Western Digital

White Paper

The Long Term Case for HDD Storage

An analysis of the economic forces driving the future use of SSDs and HDDs in the data center.

October 2024

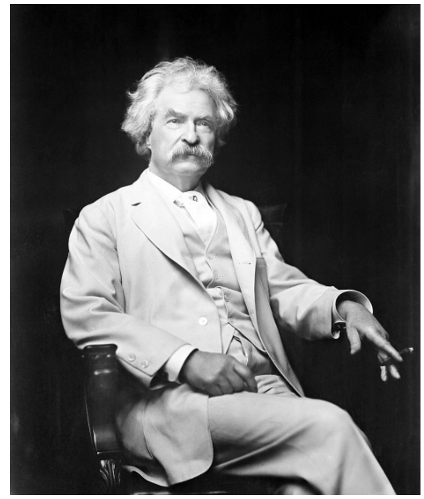
The last decade has been a tumultuous time in the data storage industry. Solid state drives (SSDs) have transformed from being a luxury item that only gamers and power users could justify purchasing, to become the mainstay of personal computing. During this time, hard disk drive (HDD) sales unit volumes have shrunk as consumers opt for SSDs, due to high performance, low power consumption, shock tolerance, and compact form factors. In lower capacity applications, SSDs may even often be available for a lower unit price cost than HDD.

A decade ago, a typical home could have contained several HDDs. Each personal computer typically had an HDD as its primary storage. Some users would have a network attached storage (NAS) in their home containing HDDs to back up those PCs, and others would have USB direct attached storage (DAS) containing HDD for backup. Most cable/satellite receivers would have an HDD enabling its DVR functionality, often having multiple receivers to support multiple TVs, each containing its own HDD. Some homeowners even had home security/ surveillance systems, which stored the recordings from each camera on an HDD.

This has all changed. Much of "computing" has been replaced by tablets and phones, or chromebooks, which are all flash-based. Nearly all new PCs are flash based. Many consumers now resort to cloud backup solutions outside their home instead of a home NAS or DAS solution. TV for many has moved to streaming devices with DVR functionality being "cloud DVR" outside the home. Security and surveillance systems are still mostly HDD-based, but fewer consumers are choosing this option as the cloud-based ecosystems underpinning consumer-focused smart cameras make adoption relatively frictionless.

Much of this has been driven by flash becoming less expensive over time, a trend that is expected to continue. As a result, typical consumers are now living in a world where they rarely see, much less think about, HDDs. It causes many to ask, "Why do we still need HDDs?" And when most learn that HDDs still make up the bulk of all datacenter storage, might be tempted to ask "But won't flash take over there as well, just later?"

Yet the storage industry, analysts, and major customers all expect HDDs to remain viable in the future. And not just to "remain viable", they expect enterprise HDD unit volumes and capacity shipments to continue grow year over year, for the foreseeable future. Those who predict the demise of the HDD must be missing something, but what is it?



"*The report of my death was an exaggeration*."

-Mark Twain

The Explosion of Data

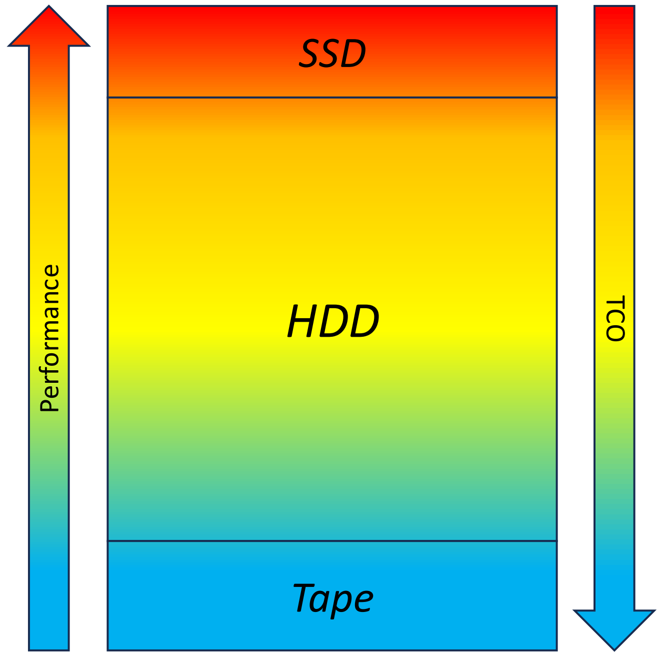
Data is being created at an exponentially growing rate. Research and market intelligence firm IDC projects[1] that global data generation will increase from 132.4 zettabytes (ZB) in 2023 to 393.9 ZB in 2028, more than doubling. They highlight that only a small portion of this data is stored long-term, but they likewise project[2] a growth in installed data storage capacity from 9.3 ZB to 19.2 ZB over that same span, also more than doubling. The bulk of that 9.8 ZB increase in installed data storage capacity will come in the enterprise or data center segment. The data center market is rapidly expanding, with robust new data center construction and vacancy rates at or near historic lows[3]. The industry is building the capacity rapidly to both store and utilize data.

Data—and the ability to transform that data into actionable information—is the lifeblood of the modern economy. Data is the light that gives us insight into what is happening in markets, in health research, in climate change modeling, and in nearly every other discipline in our world. As a result, the need for data storage was already projected to continue increasing, even before the arrival of generative AI. Generative AI, which is still very much a new and emerging phenomenon, is expected to accelerate global data creation and the demand for data storage platforms to handle that data. While much of the current use of generative AI is based upon large language models generating text which is compact to store, the future of generative AI will increasingly include image and video generation. Images and video will require orders of magnitude more storage capacity than storing text. It is still unclear how much or how quickly this acceleration will materialize, but it is coming.

As a result, the first lesson about the SSD and HDD markets, particularly in the datacenter rather than consumer devices, is that the two technologies are not fighting over a fixed pie. The pie is growing rapidly, as data storage is a secular growth market. Both technologies are going to be fighting not only to capture a share of that pie, but merely trying to keep up with rapidly growing demand.

1. Source: IDC Global Datasphere Forecast, 2024-2028, May 2024, US52076424
2. Source: IDC Worldwide Global StorageSphere Forecast, 2024-2028, June 2024, US52312824
3. https://www.cbre.com/insights/reports/global-data-center-trends-2024

Workloads and TCO



In any datacenter application, "data" is not a homogenous pool, where every piece of data is handled the same way. Each application and each piece of data has requirements on how often it must be accessed, how quickly it must be accessed, and how much it's worth paying for that speed of access.

It may be surprising to some, but magnetic tape still comprises about 10% of all installed cloud data center storage capacity[4], and is projected to remain at roughly that percentage over the next 5 years. Tape, which has access latencies of minutes to hours, is still the most cost-effective storage option for data that is primarily archival, backed up for disaster recovery purposes, or other applications where the data is considered truly "cold."

For the bulk of the data center, however, we're largely dealing with active "warm" data. But there can be a big difference between warm and hot. For example, consider the data pipeline, latency, throughput, and uptime demands of a popular internet video streaming site. They may host a music video from a globally famous recording artist that will receive billions of views, on the same streaming platform hosting a video from a struggling influencer trying to amass views in the thousands, or even a video of an amateur golfer who filmed her swing to share with her golf coach, and will only receive a few views. The storage demands for these videos are significantly different. The music video likely needs to be maintained on SSD for primary storage or even held in memory, due to its high frequency of access, while the latter two video types can be stored on HDD and still *easily* meet the performance needs of the application to serve the video to a viewer.

As a result, data center operators do not choose their storage based purely on seeking the highest performance option, because not every application requires near-zero latency across multiple regions requesting simultaneous access to the same data. They choose storage that performs adequately for the application's requirements while optimizing their storage investments to minimize the total cost of ownership (TCO) of that storage over time.

Similarly, a global company needing to ship products around the world will use container ships and rail transport for some shipments that are cost-sensitive but not time-sensitive, while using more costly air shipping and sometimes even same-day human couriers for shipments that had to be there *yesterday*. Choosing the right logistics approach—or the right data storage device—is much more complex than simply choosing the highest performance storage technology.

4. IDC Revelations in the Global Storagesphere Aug 2023

TCO Comparison Between HDD and SSD

When a data center operator is evaluating a storage solution for TCO, there are many factors that must be considered. One of the key factors, of course, is acquisition cost of the storage device itself, usually measured in $/TB rather than $/unit terms. However, there are many more factors involved. How much power will the storage device consume, whether idle or active, and if they are different, what portion of time will the devices be active vs idle? With power, of course, comes heat, and the requisite power draw to cool the devices. What is the storage density of the device, as datacenter floor/rack space is a constraint on total storage capacity that can be installed? What is the usable life cycle of the product, and at what point will it need to be replaced either for reliability purposes or to take advantage of newer platforms with higher storage capacity or performance characteristics that can yield better TCO? Today, data center operators are increasingly including environmental, social, governance (ESG) considerations into their acquisition strategy, contemplating the embodied emissions of their equipment and the emissions necessary to power their data center.

Indeed, there are several key areas upon which TCO is typically evaluated:

- **Acquisition Cost:** This is a strong advantage for HDD. When looking at the $/TB premium for enterprise grade SSDs to capacity enterprise HDD, the products used in data centers, IDC shows that the premium has been between 7.4x and 9.9x between 2020 and 2024, except for a drop to 5.1x in 2023 when the flash market was in a historic oversupply condition[5]. IDC projects that gap to remain above 5x through 2028, the reasons for which will be discussed later in this document.
- **Power Consumption:** This is close enough to parity that it does not become a major factor. While in the consumer product category, SSDs can go into efficient idle modes that consume very little power, enterprise grade SSDs typically remain in a state of higher readiness than consumer devices, which places their idle power in the same range as capacity enterprise HDDs. In active mode, enterprise grade SSDs can use significantly more power than capacity enterprise HDDs. With this added power comes impressive performance, but that is still power that will be consumed and heat which will be generated and require cooling. Due to the acquisition cost gap, very few data centers buy SSDs to keep them idle, so the active power plays an outsized role in the total power consumption of SSDs relative to HDDs.
- **Storage Density:** This is an area where SSDs ***can*** have an advantage over HDDs. Capacity enterprise HDDs in a 3.5” form factor are currently limited to approximately 32TB . SSDs in the same form factor are already commercially available at 100TB. Enterprise grade SSDs in a U.2 form factor (similar to 2.5” HDD) or E1.L form factor (aka “ruler”) in ~64TB or ~128 TB capacities will be available soon. However, this storage density advantage is somewhat mitigated by the limitation of the SSD interface—it is often better for high-performance applications to use multiple smaller capacity drives, to achieve higher aggregate system throughput. And of course, the $/TB premium still exists for these higher capacity drives. This is evidenced by very low commercial uptake of these ultra high capacity SSDs given their prohibitive pricing.

With the power envelope being similar between the two, and storage density not being a completely clear advantage, the most compelling driver of TCO is acquisition cost. Capacity enterprise HDD is a far more compelling TCO solution for any application at scale where the performance is adequate. TCO metrics where enterprise grade SSDs have a perceived advantage do not come nearly close enough to bringing the TCO gap created by disparate acquisition cost into a range where substitution makes sense.

5. Source IDC Worldwide 2024–2028 Enterprise HDD: Market Overview, IDC Doc# US52411324, July 2024

As a result, modeling the future of the storage industry suggests that the key factor that we should be looking at going forward is acquisition cost comparisons between the two capacities. And that story starts with a discussion of economics.

The Economics of Cost and Price

One of the most basic lessons of economics is simple: price is based on the balance point between supply and demand. The price of a good is not determined by production costs. The price of a good is not determined by profitability. In fact, contrary to popular belief, the price of a good is not "set" by the producer of that good, or producers would never lose money. It is merely the point at which the market demand for a product and its supply hit a point of equilibrium to sell all the product on offer—*the market-clearing price*. It is a snapshot in time of that equilibrium point.

This is not to say that cost or profitability are not important. In fact, they're critically important, because they drive what happens beyond a snapshot in time. For every for-profit company, they make decisions based on costs, competition, and perceived market demand, to try to align the supply of a good or service with demand at profitable levels. This longer-term decision making is important, because it reveals another basic lesson of economics:

*Prices can never decrease, **long-term**, faster than the production costs.*

In the short term, prices certainly can move wildly. They can move to high profitability if supply is constrained and product shortages emerge. They can move to low—or negative—profitability if supply exceeds demand by a large amount.

But in the long term, prices and costs must trend down at similar levels. When prices decline faster than costs, it often results in reduced investment in production, because profitability declines and/or companies lose money. In industry, the word "oversupply" is used when this occurs, but the truth isn't that supply is objectively too high. Instead, it means *supply is too high to support financially sustainable pricing*.

This was demonstrated only very recently from roughly mid-2022 through the end of 2023 in the data storage market. The effects of the COVID pandemic wildly distorted the market and supply chains in the data storage industry as well as the wider information technology sector. This led to a boom-bust cycle where very large customers over-bought inventory of data storage devices, buying them faster than they could be deployed. When the inventory situations at these customers became untenable, they drastically cut or eliminated their purchases of new data storage devices. The effects were seen by both HDD and NAND flash / SSD vendors during this time. With poor demand, the market was in an "oversupply" condition. This led to price and sales volume declines, and many producers were unprofitable during this time.

This occurred despite long term growing demand for data storage. It occurred despite the predicted growth in both data creation and data storage needs. The fundamental knowledge that the data storage market is a growing pie did not change, but for a short period, the pie shrank. It was a temporal disruption to the supply/demand balance.

Across the industry, a predictable response occurred: production was cut to reduce supply and align to that reduced demand, allowing the over-inventory situation to be digested in the market and to hold off new production until demand showed signs of recovery. NAND flash makers reduced wafer starts to cut production. HDD makers reduced factory utilization to cut production. The prices, which were not financially sustainable, started to recover, at which point more normal HDD and SSD production levels could resume.

So if we want to predict the future of the market and the future value of HDDs relative to SSDs in the datacenter, we must understand that pricing in the long term ***is*** determined by costs. And thus we must understand what prices have done in the past, and why, and then look at what we predict for the future.

Historical Acquisition Cost

The general narrative about SSD pricing is that it has been declining to become equivalent to HDD pricing. However, this is only true in very low-capacity applications. The nature of HDD is that there is a floor capacity at which it is not feasible to produce a lower-price product at a lower capacity. An HDD with one disk and one or two heads is a minimum configuration. With modern storage technology, that is typically a 1TB 2.5" HDD, or a 2TB 3.5" HDD.

This capacity is now higher than is needed in most traditional desktop / notebook computing, where it might be common for a user to only need 128GB or 256GB of local storage. In other products, such as Chromebooks, large local storage is eschewed in favor of microSD™ card slots and cloud storage, so these devices come equipped with even lower local storage capacity. Thus, SSD capacities in this range reached price parity with larger-capacity HDDs years ago, which has driven the dominance of SSD in personal compute storage.

The unit price comparison between lower-capacity SSDs and higher-capacity HDDs has led to a belief that a crossover between HDD and SSD pricing is inevitable, and that it is only a matter of time before it arrives. However, when traditional pricing is compared not in unit costs for low-capacity options, but is compared to $/TB cost, the picture looks much different.

Western Digital, based upon industry research and company data, has modeled the $/TB cost of SSD and HDD products over many years. When looking at historical $/TB data from the third quarter of 2015 to the third quarter of 2023, an 8-year period, the price declines between the two on a $/TB basis were nearly identical:

- SSDs across this time period reduced in price on a $/TB basis at about 13.76% annually
- HDDs across this time period reduced in price on a $/TB basis at about 12.83% annually

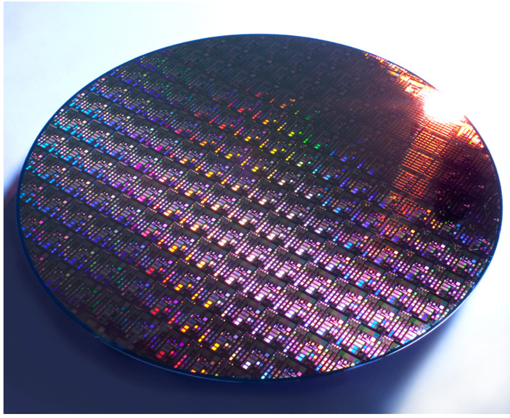
These rates are close enough to identical to be the same. Currently, basing on IDC data showing a 7.4x cost difference between enterprise grade SSDs and capacity enterprise HDDs, this would mean that it would take over 15 years at -13.76% CAGR for pricing to cross over—***if HDD pricing did not decline at all***.

But of course HDD innovation continues apace, and HDD prices are expected to continue to decline on a $/TB level. To model this, and understand what to expect in the future, requires that we delve into the economic drivers behind these declines for both SSD and HDD technologies. Only then can we accurately predict what is to come.

The Economic Unit

As stated above, prices cannot drop in a financially sustainable way faster than production costs. And since we're talking about prices, and production costs, on a $/TB basis, we need to understand how the economic and technological drivers for both the SSD and HDD industries operate. Note: when we discuss SSDs, ultimately we must discuss NAND flash. NAND flash comprises the majority of the bill of materials (BOM) cost of an SSD, and the price in $/TB terms for SSDs closely tracks the price of NAND flash. So we are comparing NAND flash with HDD.

To do this, we look at the economic units of production. The economic unit of production for NAND flash is the wafer. Each NAND fabrication plant (fab) has a fixed capacity of how many wafers they can process per year. The way to increase wafer production capacity is typically to build a new fab. When a new fab is built, it is an investment requiring billions of dollars of capital expenditure (CapEx), and that CapEx is only invested if the belief of the industry is that storage demand, **at profitable sales prices**, will justify the investment.



However, building new NAND fabs is not the only way to increase bit supply, measured as the total amount of bits that the industry can ship to consumers. Another way is to increase the quantity of bits per wafer. This is a technological improvement based upon innovation. The NAND industry has many examples of doing this. The original 2-dimensional or “planar” NAND pushed to smaller and smaller process lithography, shrinking the memory cells in size, which allowed for more bits per wafer. The industry then went from single-level-cell (SLC) NAND to multi-level-cell (MLC), which uses the same physical cell to store additional data states. The initial push from SLC to MLC resulted in a doubling of data bits per wafer. However, smaller lithography had a downside: smaller physical cells could not retain charge as well or as long as larger cells, so endurance and data retention suffered. The industry needed a solution, and found it in 3D NAND. With 3D NAND, the cells can be physically larger, allowing not only for MLC, but now for triple-level-cell (TLC), quad-level-cell (QLC), and at some point in the future, possibly even five-level-cell (PLC). This expands the number of bits per cell, while the vertical stacking of layers allows for these larger cells to no longer be a limiter to wafer capacity.

Each of these technological improvements is part of what has driven the price declines in NAND flash and thus in SSDs, not increased production capacity. Thus, the driving factor in SSD price is not production capacity, but it is the number of bits per wafer. The industry measures this using the GB/wafer metric.

For HDDs, we must ask a question if the platter is the economic unit, or in modern data center usage, whether then 3.5” capacity enterprise HDD is the economic unit. Over the last decade, the introduction of sealed and helium-filled HDDs has allowed for the HDD to be built with higher platter counts, increasing the possible storage per HDD, even more quickly than the bit storage per platter increased. So one might argue that the HDD itself is the economic unit. However, the industry is reaching volumetric limits where it is unlikely that platter counts per HDD will continue to increase. So the *future* of capacity increase will depend on the ability to expand the number of bits per platter, or areal density, as measured by terabits per square inch (Tb/in^2).

As in the NAND industry, increasing the areal density capability (ADC) of an HDD platter has been a multi-decade story of innovation pushing recording technology ever further. Looking back only through relatively recent history, a major development less than

20 years ago was to move from longitudinal magnetic recording (LMR) to perpendicular magnetic recording (PMR). This allowed the HDD vendors to significantly shrink the size of bits, increasing ADC. Introduction of two-dimensional magnetic recording, the triple-stage actuator, energy-assisted PMR (ePMR), OptiNAND®, and UltraSMR have all been ways to better write ever-smaller bits to the media and read them back accurately. Going forward, the introduction of heat-assisted magnetic recording (HAMR) will provide the next major inflection point in attaining even higher ADC.

So when we think about the production of NAND flash or HDDs, we think about the number of bits we can produce on each economic unit, and how that will change over time. This GB/wafer or TB/in^2 can be used as a *proxy* for cost. It will never be a perfect relationship, because often these technical innovations that improve bit density carry additional costs to implement. If nothing else, higher bit density and larger storage devices take longer to test than smaller, and even test time is an increased production cost. That said, the proxy can still give a rough idea of a relative change over time between the two technologies. Ultimately, this will tell us whether we can predict a crossover point or even a significant narrowing between them.

Future Bit Density Projections

If bit density is the driving factor in production cost, we should look at what the industry is saying about future bit density.

Tech Insights, a research organization tracking the NAND industry, publishes a quarterly report on the NAND market. The Q3 2024 report contains a 3D NAND GB/Wafer projection showing that from 2023-29, the bit density (GB/wafer) will increase at a CAGR of 18.2%[6]. That's impressive, and it reflects the growth in number of layers of 3D NAND and the transition towards more QLC adoption.

On the other side, the Advanced Storage Research Consortium (ASRC), a forum for joint research initiatives among and between storage industry participants, published in 2022 at the IEEE 33rd Magnetic Recording Conference (TMRC) their projection of the future ADC roadmap for HDD technology. Their projection shows the adoption of HAMR leading to a CAGR of 20% ADC growth to 2030, and other technologies (HAMR+/HDMR) extending that 20% CAGR out through 2035[7].

NAND GB/wafer and HDD ADC are the fundamental technology metrics that will drive future production cost, and they are expected to increase at nearly identical rates. ***Most importantly, this suggests that the gap between the two technologies in the market as measured by $/TB acquisition cost will continue to persist.***

Industry and Customer Behavior

Through the discussion to this point, several key pieces have been established:

1. There is a significant gap in acquisition cost between HDD and SSD, with a corresponding significant TCO advantage to HDD.
2. Industry players will only increase bit supply if the product can be sold profitably.
3. The driver of production cost is bit density whether at the NAND wafer or HDD platter level.
4. The production cost curves for each technology appear to be trending down at similar rates.
5. Customers will purchase storage based on minimizing TCO.

6. TechInsights NAND Market Report Q3 2024
7. A new Advanced Storage Research Consortium HDD Technology Roadmap

NAND vendors *can* increase bit supply to meet current total storage demand. While NAND production is a heavily capital-intensive business, building new fabs is quite possible. It is an economic calculation, not a technological limit on supply.

However, it is clear that increased NAND bit supply will only be sought if it can be achieved profitably. Since customers in lower-performance workloads can use HDD or SSD interchangeably, and we know that the customers will seek to minimize TCO, customers will choose HDD over SSD if it meets their performance demands and has lower TCO.

Even in the severe industry downturn from mid-2022 until the end of 2023, when SSD pricing fell significantly—relative to HDD—the TCO advantage of HDD was still too significant to shift data center market share to SSD. In this downturn both NAND vendors and HDD vendors were losing money, and took measures to restrict supply. The NAND industry cannot increase bit supply without reducing production cost and expect to remain profitable.

In order for NAND and SSD acquisition cost to meaningfully close the gap, the NAND bit density as measured by GB/wafer will need to increase at a faster rate than the HDD industry can increase ADC. This is not projected to happen. This suggests that there is no way for the SSD industry to bridge the TCO gap. The takeaway is simple: due to inherent technology projections that drive production costs, there is no scenario where SSD simply takes over the market from HDD. HDD is here to stay, at least for the foreseeable future.

Counterpoints

We have established that trajectory of the market does not support the thesis that HDDs will be replaced by SSD in the data center. It is important, however, to look at what could change that trajectory, and what might falsify this conclusion:

- Major technological advance in NAND (or other non-volatile media) technology: The current model of wafer bit density growth is 19.7% CAGR for NAND flash. This projection includes growing 3D NAND layers and the continued adoption of QLC and perhaps even PLC flash. But if a revolutionary technology is invented that significantly reduces the cost of production, it would allow the gap to close to HDD. If the gap closes enough, SSD may become close enough in TCO to meaningfully displace HDD. There is no current known or hypothesized invention capable of doing this.
- HDD areal density hits a wall: The projections of HDD areal density capability rely on the ability of vendors to be successful with HAMR and other technologies to keep growing. As with all new technologies, there is uncertainty until they are field-proven at scale. If HDDs were to be unable to keep growing ADC, due to any insurmountable technological hurdle, that would allow SSD to close the gap and potentially displace HDD. Today, the industry strongly believes in the future ADC technology roadmap, and continue to execute to it each year.
- Storage demand declines precipitously: There have been suggestions that technologies like compression or deduplication could cause a reduction in overall storage demand. If this were to occur, the effects on the industry are unknown, but could result in there not being enough overall demand to support both technologies, and with total storage budgets likely decreasing, may lead to SSD displacing HDD. Industry projections of future data creation, storage demand, and global data center construction tell the opposite story; continued exponential increase of data and storage needs, likely accelerating due to new technologies like generative AI.

The future is never crystal clear, but there is nothing currently on the horizon that meaningfully will cause SSDs to displace HDDs *en masse*. If it happens, it will likely be due to something that nobody today sees coming.

Conclusion

Data generation and data storage are secular growth markets. What this means for storage vendors is that the future is very bright for SSD. But the economics of these two technologies mean that the future is also very bright for HDD.

The two technologies have different strengths and weaknesses. SSDs offer stunning performance, but it comes at a cost. HDDs cannot match that performance, but have a tremendous advantage in TCO. The result is that the storage solution that is chosen for any particular application or workload is going to be dependent on what is needed. *Customers will use HDDs where they can, but SSDs where they must*.

The applications where HDDs are more than adequate are large, and growing. With the TCO advantage that HDDs offer today, and the expectation that this TCO advantage will persist in the future, it seems the rumors of their demise were greatly exaggerated.

5601 Great Oaks Parkway
San Jose, CA 95119, USA
www.westerndigital.com