PURE STORAGE® AND FLASH

Why it's Time for Flash to Replace Disk and Why Pure Storage Continues to Lead in Flash Systems

As digital data grows in scale and diversity, storage infrastructure costs put increasing pressure on IT budgets. In the last decade, users have adopted two tiers of online storage media: flash for peak performance and disk for high capacity. Observers believe that while flash cost continues to decline, it will be 2030 or later before all-flash storage systems cost less than disk-based ones.

But if the TCO of flash storage were the same as or less than disk, would there be any reason for users to buy disk-based systems?

With its co-designed hardware and software, Pure Storage has accelerated the crossover point to *today*, eliminating the need for disk-based systems in modern storage infrastructures. This acceleration is the result of three key innovations:

- Adopting each new generation of flash without sacrificing storage system reliability or economics.
- Scaling storage module capacities well beyond those of commodity *solid-state devices* (SSDs).
- Developing a system architecture whose controller software manages storage modules with directly addressable flash to overcome the shortcomings of conventional SSDs.

This brief explains why SSDs, although they are superior to disks in enterprise storage systems, do not fully exploit flash, and describes how architectural innovations in Pure Storage systems realize the full potential of flash today and will continue to do so in the future.

THE STATE OF DISK STORAGE IN THE ENTERPRISE

Today

Over the past decade, flash-based storage systems have essentially supplanted disk-based ones in performance-sensitive applications. But disks still dominate in the “nearline” storage systems that hold the majority of the world’s online digital data (much of it in public clouds).

The 7,200 RPM disks in nearline systems transfer data at 200–300MB/s but seeking and rotational latency limit them to about 100-150 random IOPS (with latency in the 5-10ms range). As disk capacity increases, streaming data transfer rates increase somewhat due to higher linear bit density, but random IOPS and latency remain about the same, so random IOPS *per TB* decrease. Systems based on these devices use erasure coding (e.g., RAID) to protect against data loss due to unrecoverable I/O errors and device failures.

To improve random I/O performance, storage systems *virtualize* capacity by striping data across several disks. For example, a 12-disk stripe of nearline disks could deliver as many as 1,500 random IOPS, although uneven distribution of requests would likely reduce that number.

While disks are among the most robust mechanical products made, their components wear out with age. Conventional wisdom holds that disk failure rates follow the “bathtub curve” shown in Figure 1.[1] The referenced publication goes on to mention that while “infant mortality” seems to have lessened in recent years, a disk in its fifth year of service has about a 4% chance of failing—a significant problem for large data centers. For example, a data center with 100PB of disk storage would have about 5,000 of today’s 20TB disks. Statistically, it would expect 200 failures in the fifth year of service. Each failure means a performance-sapping data rebuild lasting for hours or even days with data at risk from a second failure, as well as inconvenient data center visits for disk replacement. Disk failures require prompt replacement, and data durability requires elaborate erasure coding, which reduces media efficiency. In short, disk failure is the main reason that disk-based storage system service life is generally considered to be five years.

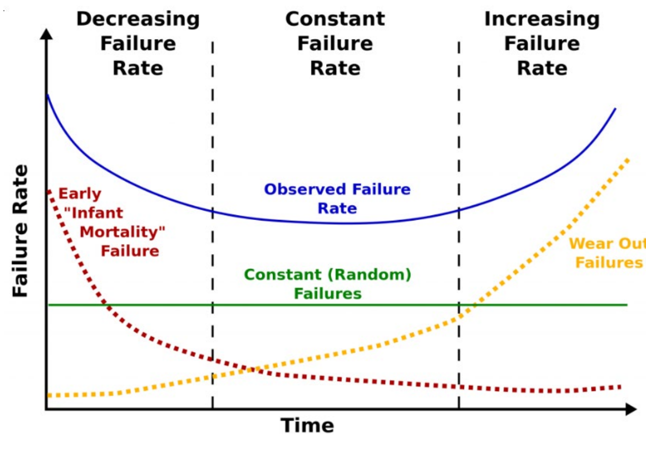

Figure 1: The Disk Failure “Bathtub Curve”

In the Future

Disk manufacturers regularly publish “roadmaps” that forecast product introductions and technology innovations. Current maps indicate expected capacity increases through the use of new

[1] Adapted from https://www.backblaze.com/blog/drive-failure-over-time-the-bathtub-curve-is-leaking/

recording techniques like *energy-assisted magnetic recording* (EAMR) and *ordered granular media* (OGM)[2], as well as mechanical innovations like thinner platters[3] for greater capacity and dual actuators for improved IOPS/TB. With these innovations, they predict 40–50TB devices by the end of the current decade. But if the disk industry's history over the past decade is a guide, these predictions may be optimistic. Even if they are accurate, it is safe to assume that 40–50TB is the upper limit for disk capacity in the current decade.

The 3.5" disk form factor sets some basic parameters for nearline storage systems. Higher capacity disks will still occupy about 400cm^3 (plus carrier) and consume 5–6 watts of power when idle[4] (10–11 watts when active). Even if 50TB disks *do* materialize in 2030, disk storage will still require 40% as much power, cooling, and rack space per TB as it consumes today.

Most importantly for large systems, disk random access time is constrained by seeking and rotational latency to about 5–10ms and is unlikely to change much. Even if dual-actuator disks were to become a reality (optimistically doubling per-device IOPS), today's solid-state devices (SSDs) would still be capable of thousands of times more random IOPS (>1,500,000[5] vs. 300).

ENTER FLASH

Flash-based solid state storage devices have a 30+ year history,[6] but it wasn't until the early 2000s that the first enterprise-ready SSDs were introduced. Developers realized from the outset that the fastest path to broad adoption would be to build devices that could directly replace disks, even if they did not deliver the full benefits of flash. Early in the following decade, all-flash storage systems using SSDs, including *FlashArray*™ systems from Pure Storage were introduced.

Flash has clear advantages over mechanical disks—near-zero latency, reliability, longevity, space efficiency, power consumption, cooling requirements, and end-of-life environmental impact—but the technology has two inherent quirks that challenged early system developers—reading was *much* faster than writing, and endurance[7] was limited. Flash-based SSD developers used a combination of DRAM buffers, overprovisioning[8], dynamic data relocation and remapping to overcome these limitations and deliver expected lifetimes[9] comparable to those of disks.

Today, SSDs are used in everything from laptop computers to enterprise-class arrays. Most enterprise SSDs are in the 2.5-inch form factor, originally developed to fit disk bays. They use a variety of interfaces, including SATA, SAS, and direct connection to PCIe buses with NVMe

---

2 For example, see https://blocksandfiles.com/2022/05/11/western-digital-disk-roadmap/
3 https://toshiba.semicon-storage.com/eu/company/news/2022/03/storage-20220308-1.html
4 In principle, idle disks can be spun down to conserve power, but accessing a spun-down disk incurs about 10 seconds of latency for spin-up. Moreover, frequent starting and stopping is a major cause of disk failure. Cf. Non-RAID drive architectures - Wikipedia
5 https://media-www.micron.com/-/media/client/global/documents/products/product-flyer/9400_nvme_ssd_product_brief.pdf
6 Cf. https://en.wikipedia.org/wiki/Solid-state_drive
7 The number of times a flash cell can be overwritten before it will no longer retain data reliably.
8 "Extra" flash not exposed outside the device. Used for "garbage collection" and to increase device endurance.
9 Manufacturers' warranties are a proxy for storage device expected lifetimes.

protocols. 30TB devices with 3–4GB/s transfer rates, over a million low-latency random IOPS, and endurance of 1–3 *device writes per day* (DWPD)[10] are available from manufacturers.[11]

The capacity, performance, space efficiency, and power and cooling advantages of flash have made it a clear choice for persistent storage in high performance systems. Today, flash-based systems dominate the performance-sensitive segment of the data storage market.

SSDs and Nearline Storage

The story is different for price-sensitive *nearline* storage, where fewer IOPS per TB are required and enormous data sets make acquisition cost the dominant system selection criterion. But as flash technology improves, adoption increases, and manufacturing economies of scale reduce its cost. Even today, vendors ship more flash capacity than disk[12]. Some predictions have nearline disk and SSD cost per TB reaching parity later in the current decade, with solid state devices rapidly outpacing disks in the years that follow.[13]

Of course, device cost is only one element of storage system *Total Cost of Ownership* (TCO). When overall acquisition (including packaging, cabling, controllers, etc.), data center space, power, cooling, management, support, and retirement/disposal costs are considered, Pure's flash-based systems are already at or below parity with disk-based ones.

The Future of Flash

Increases in flash density over the past 30 years have consistently been greater than those for disks, and manufacturers' projections continue to reflect exponential capacity increases relative to magnetic media[14]. Increases will come from smaller cells, more bits per cell, more layers per die, and stacking dies vertically on wafers. SSD manufacturers confidently predict 100TB devices by the end of the decade, with one going so far as to suggest that 1,000TB SSDs are realistic.[15]

In summary, SSD manufacturers predict that devices with twenty times the capacity and capable of delivering upwards of 10,000 times the IOPS of disks at much lower latency, will be available at lower cost per TB—more than enough capacity and random I/O performance for any application. And with no moving parts, SSDs will continue to be more reliable than disks.

---

[10] DWPD is the number of times per day a device's entire usable capacity can be written during its warranty period. For a 30TB device with a five-year warranty for example, one DWPD would mean about 350MB/s written every day for five years—far beyond any realistic usage.

[11] For example, https://americas.kioxia.com/en-us/business/ssd/enterprise-ssd.html

[12] But not more SSDs. A lot of flash is used in appliances and personal devices.

[13] For example, https://blocksandfiles.com/2021/01/25/wikibon-ssds-vs-hard-drives-wrights-law/

[14] For example, https://wccftech.com/218-layer-3d-nand-flash-from-kioxia-western-digital-delivers-huge-leap-in-performance-cost-effectiveness/
and https://www.techradar.com/news/1000tb-ssds-could-become-mainstream-by-2030-as-samsung-plans-1000-layer-nand
and https://www.forbes.com/sites/tomcoughlin/2022/08/13/2022-flash-memory-summit-announcements/

[15] For example, https://mezha.media/en/2022/10/06/1000tb-ssds-could-become-a-reality-by-2030-as-samsung-plans-1000-layer-nand/

PURE STORAGE AND SOLID-STATE DEVICES

Since the company's founding in 2009, Pure Storage has focused exclusively on flash-based storage systems for the enterprise. As a small startup, the company initially equipped its systems with commercial SSDs and developed innovative software to minimize system cost and overcome the shortcomings of that era's devices. It was quickly realized, however, that SSDs designed for a wide range of applications are not optimal for enterprise storage, so in 2013 Pure Storage innovated by adapting its Purity software to control flash directly and designing its own flash modules built with NAND chips. The company shipped its first directly addressable flash devices in 2016. Today, hot-pluggable *DirectFlash™ Modules* (DFMs) provide persistent storage managed by Purity software in FlashArray™ and FlashBlade® systems.

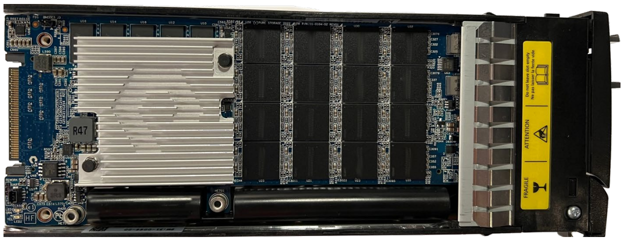

Figure 2: A 75TB DFM

DFMs vs SSDs

The integration of Purity software and DFMs eliminates the redundancy in SSD-based systems caused by SSDs' emulation of disks. DFMs differ from SSDs in two important respects:

- They expose all of their flash to Purity running in controllers. DFMs are not overprovisioned and do no internal relocation and remapping of stored data. Direct access has the beneficial side effects of minimizing latency by eliminating indirection and increasing endurance by greatly reducing *write amplification*[16].
- Unlike SSDs, which are constrained by disk interface protocols (including NVMe), DFMs are designed specifically to cooperate closely with Purity software, which optimizes I/O scheduling, data placement, fault recovery, and other operations across multiple devices.

In addition to delivering more consistent performance at the device level, DFMs are less costly to build, and their simplicity makes them more robust, than conventional SSDs.

The Purity software that drives FlashArray and FlashBlade systems schedules I/O across multiple devices to equalize wear and maximize endurance. It tracks usage and failures at the flash block level based on reports from DFMs, which enables it to optimize I/O and incoming data placement, and to retire defective blocks intelligently. From a user perspective, Purity-DFM integration makes system I/O performance more *consistent* than that of SSD-based systems, with far fewer unpredictable high-latency responses to host commands.

DFMs connect to controllers via PCIe. The NVMe protocols they use support highly parallel operations—in busy systems, a single DFM may perform over a hundred concurrent data transfers. As a result, Pure Storage systems deliver better *performance per device* than systems

[16] Writes to flash originated by the device in the course of managing stored data. A significant issue for conventional SSDs, which constantly relocate stored data during garbage collection.

with conventional SSDs. Today, a single-chassis FlashArray//XL system with 40 DFMs can execute over a quarter million 128KB random I/O commands per second (36GB/s). A single-chassis FlashBlade//S system can provide 66GB/s of throughput from QLC-based DFMs.

THE FUTURE OF DFMS

As has historically been the case with semiconductor technologies (e.g., DRAM), flash density typically doubles with each generation, compared to the much more modest increases common between disk generations. Pure Storage has close relationships with major flash manufacturers and is usually among the first to receive production units of their new products.[17] In 2023, the company will deliver 75TB DFMs. As manufacturers introduce new generations of flash in the 2–4 years that follow, Pure Storage will deliver systems with 150TB, 300TB, and even larger DFMs.

THE SUSTAINABLE ADVANTAGE

SSD manufacturers produce higher-capacity devices with each new flash generation, and other storage system vendors incorporate them in their systems. But to achieve manufacturing scale, each new SSD generation must be designed to address the broadest possible market. Thus, SSDs will continue to use disk-like interfaces that hide the internal complexity of write buffering, wear leveling, "garbage collection," and caching of the maps that relate external data addresses to intra-device flash locations. Moreover, it's safe to assume that each SSD generation's unique firmware will require "hardening" before being production ready.

SSDs use DRAM both to buffer writes and to cache data location maps, the latter growing in proportion to their flash capacity. As long as SSDs emulate disks, the typical DRAM:flash ratio of about 1:1000 is unlikely to change; higher capacity devices will need more DRAM.

SSDs also have supercapacitors to provide power for persisting DRAM contents in the event of external power failure. The power required is proportional to the amount of DRAM.

DRAM density increases have slowed in recent years, causing pricing to remain roughly constant (Figure 3). Larger amounts of DRAM and heftier supercapacitors will increase the cost, power consumption, and space requirements of future SSDs. Developers could reduce SSD DRAM:flash ratios (limiting performance), adopt larger device

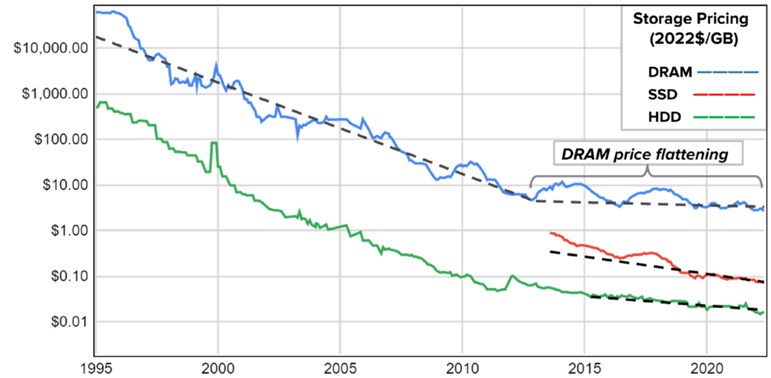

Figure 3: Storage Pricing History

[17] For example, Pure Storage has delivered two generations of QLC-based DFMs earlier than their manufacturers delivered SSDs utilizing the same flash technologies.

form factors, or create a new SSD interface paradigm. Any of these alternatives would imply major changes in storage systems that would limit the market for high-capacity SSDs and would therefore tend to keep prices elevated.

DFMs are not overprovisioned and do not relocate data internally. They do use DRAM as NVRAM cache, but they require only about a 1:1 million DRAM:flash ratio—1/1000th of the typical SSD requirement—making their inherent hardware cost lower. They reserve a small part of their flash to backstop NVRAM cache, so they require on-board supercapacitors, but with only 1/1000th as much DRAM, their supercapacitor needs are much less than those of SSDs.

Firmware is a leading cause of solid-state device failures. Using data from customer systems collected by its Pure1® service, the company analyzes its deployed products throughout their lifetimes, including DFM behavior. Regular Purity software updates include any applicable DFM enhancements, so reliability of installed DFMs actually *improves* over time. In contrast, it is rare for SSD manufacturers to monitor deployed devices or issue firmware updates for them.

In summary, Pure Storage DFMs have sustainable advantages over SSDs in capacity, performance consistency, internal simplicity, and reliability that will increase along with device capacity. Moreover, they provide additional benefits at the system level by making centralized decisions across many devices possible. To compete effectively with DFM-based systems, a developer would have to create both controller software for direct flash management and devices optimized for centralized management. Having begun development of its integrated architecture in 2013, Pure Storage has a decade lead in this space.

STORAGE TCO: THE *REAL* DIFFERENTIATOR

Vendors of disk-based storage systems argue that because disks currently remain less expensive than flash devices on a per-TB basis, they will continue to have a role in enterprise storage. This argument, however, fails to consider the *total cost of ownership* that includes acquisition, operation, maintenance, management, and disposal at end-of-life:

Power Consumption

As of 2021, data centers accounted for almost 1% of total global electricity use.[18] In the past two years, data centers have seen huge increases in power cost—in many cases doubling pr more. Some grids cannot supply data centers' power needs at any price. With ever-increasing amounts of data to store this is becoming untenable.

If the disk industry's predictions of 30TB devices in 2026 are realized, device power per TB will decrease by about a third. Today, however, Pure Storage DFMs consume only about 40% as much power per TB as disks, and the higher-capacity DFMs of the future will consume little if any additional power. Moreover, systems will contain far fewer of them for a given capacity. For example, in 2026, a 100PB data center would require 3,333 30TB disks, but only 333 300TB DFMs—10% of the disk power requirement.

[18] https://www.iea.org/reports/data-centres-and-data-transmission-networks

Cooling

DFMs generate about a third as much heat per TB as nearline disks. Depending on other data center requirements, forced air cooling might be adequate with Pure Storage arrays, whereas disk-based systems of equivalent capacity are more likely to require expensive, power-consuming liquid cooling.

Size and Weight

Today, a FlashBlade system packs 1.92PB into five RU (8.75 vertical inches of rack space). As higher-capacity DFMs are delivered, density will increase accordingly. In 2026, Pure Storage will be delivering over 8.4PB in a three RU chassis of normal rack depth. Even if disks reach 30TB in that time frame, Pure Storage systems will pack three times the storage into less than half the rack space. And that's with packaging density similar to that of today's FlashArray and FlashBlade systems. Should the company develop higher density enclosures, future systems could triple this to about eight PB/RU, and still provide full DFM benefits.

Reliability and Maintenance

Disk failures mean lengthy data rebuilds that significantly impact system I/O performance and require prompt replacement. Not only do disk-based systems require more devices per TB, but disks fail more often, especially as they age. Based on the report cited in footnote 1 and Pure Storage internal ARR data, DFMs are only about 1/16th as likely to fail as disks in the fifth year of deployment. In 2026, a 100PB data center with 333 Pure Storage 300TB DFMs would expect only 1/160th as many device failures as one with 3,333 30TB disks.

Not only do DFMs fail less often, but their high I/O speed means that rebuilding affects system performance much less than in disk-based systems. Fewer failures and faster rebuilds make error recovery a much smaller risk for flash-based systems.

Most disk failures abruptly incapacitate the entire device. In contrast, DFM failures are electronic; many are wholly or partially recoverable. For example, a 2Tb (256GB) flash die failure in a 150TB DFM disables 0.16% of the device's capacity. Purity software stops writing to the device and generates an alert that automatically initiates a customer support ticket, but 99.84% of the device's capacity is still readable. Purity rebuilds any data stored in the failed area in alternate locations—a much faster process than rebuilding an entire device. Moreover, failed DFMs can often be repaired and restocked—never the case with disks.

And reliability isn't just about disks—more devices require more chassis, more storage network ports and cables, more fans, and more power cords; in other words, more ancillary components, many of which rotate or wiggle…*and are therefore more likely to fail*. The only moving parts in a Pure Storage system are cooling fans.

End-of-Life and Disposal Considerations

At end-of-life, storage devices may contain confidential data. DFMs encrypt all stored data all the time, using keys that must be reassembled from *shards* stored on multiple devices. There's no need to physically destroy Pure Storage DFMs to protect data's confidentiality.

Increasingly, enterprises are incurring recycling and disposal costs as they retire IT equipment. A DFM weighs about a third as much as a 3.5" disk. With ten times the capacity and a third of the weight, DFMs will produce thirty times less e-waste per TB than disks. And their 50% longer life increases that to forty-five times less device-related e-waste over time.

Pure Storage's *Evergreen™* program provides additional advantages to its customers:

- Regular hardware upgrades replace controllers as technology evolves. Upgrades do not disrupt operations, nor do they require data migration or parallel operation of retiring systems and their replacements.
- Customers can add capacity, in many cases by consolidating their existing data onto higher-capacity DFMs. During consolidation, DFMs can be replaced one-by-one, or in systems with unused device bays, in groups called *data packs*. Transferring data to new DFMs does not disrupt system operation. Moreover, when capacity is consolidated onto larger devices, the company issues credits for the capacity of returned DFMs.
- Replacement of failed DFMs is covered by active and qualified Evergreen subscriptions.

At The System Level....

Occupying a tenth of the space and with less than 1/100th as many device failures per PB, flash storage systems are much less complex for data center teams to manage than disk-based ones. Fewer devices for a given system capacity means fewer connections–simpler storage networks that cause fewer problems. Long DFM lifetimes and multi-generational non-disruptive Evergreen™ system upgrades mean fewer system installations, and virtually eliminate data migration and parallel operation of old and new equipment. The consistent performance of Pure Storage systems and the extensive reporting and planning capabilities of Pure1 greatly reduce the effort IT teams devote to performance management and troubleshooting. A decade of customer experience suggests that these benefits allow IT teams to manage 10 times as much capacity per person with Pure Storage all-flash systems as with disk systems.

When all factors are considered, flash-based systems are already more cost-effective than disk-based ones. Moreover, *data reduction*,[19] possible due to flash random access speed, increases the relative efficiency of flash-based systems. For example, with a 2:1 reduction ratio (common for many types of file data), the 10:1 ratio of devices per PB of disk compared to DFMs becomes 20:1. Disk I/O latency makes data reduction difficult if not impossible for many disk-based systems. This, and the other advantages of flash storage will become even more pronounced as higher capacity devices are delivered in the future.

[19] Removing redundancy from stored data by compressing and deduplicating it

SUMMING UP

This brief makes two points:

- Today, there is no reason ***not*** to replace disk storage throughout the enterprise storage realm, including in nearline systems. With higher performance, greater reliability, lower space, power, and cooling requirements, and simpler management, flash-based system TCO is already lower than that of disks, and the gap will only widen in the future.
- With Purity software directly managing the flash in DFMs, Pure Storage has a long-standing and sustainable advantage over flash storage systems that use conventional SSDs. This advantage stems in large part from its many-year (and on-going) investment in understanding flash technology thoroughly and employing it effectively in systems.

It is interesting to compare today's enterprise storage situation to transportation at the beginning of the 20th century when horses were the dominant means of personal locomotion. An alternative—automobiles—existed, but had several disadvantages: they were inconvenient, unreliable, had limited range, and there was almost no supporting environment. Automobiles, although an interesting technology, were generally considered less desirable than the incumbent equine "technology."

A short 15 years later, however, automobiles were being manufactured in volume and prices had decreased. A system of roads, service centers, and "rules of the road" had developed. In short, technology had advanced to overcome the disadvantages. Seemingly overnight, horses had become an anachronism with support disappearing. Even proponents of equine transportation converted to what had become an obviously superior technology.

Something similar is happening in enterprise storage systems today—as flash-based systems increase in capacity and drop in price, as more options appear in the market, and as flash becomes the accepted norm for nearline storage, the popularity of disk-based systems is fading, and with it, the motivation of disk and system developers to innovate. It is increasingly obvious that the future of all enterprise storage is flash.

As an early proponent of flash, Pure Storage gained an invaluable understanding of the then-available devices and their limitations for enterprise use. That understanding led the company to develop Purity software and DFMs specifically optimized for large-scale storage systems, along with expertise in flash technology which it continues to enhance with each new generation.

Purity software and DFMs are tightly integrated; together they provide synergies that are impossible in storage systems that utilize conventional SSDs. In principle, a developer *could* follow the path Pure Storage has taken, but that path is neither short nor simple. Pure Storage has a long-standing advantage in flash-based enterprise storage systems that will persist across multiple generations of flash technology.

Flash Is Poised to Replace Disk

Pure Storage Continues to Lead in Flash Storage Systems