



White paper

Scaling through flash scarcity with purpose-built architecture

Table of Contents

Executive Summary	3
Introduction	3
The capacity efficiency solution space	4
Eliminating efficiency tradeoffs through architecture	5
Durability	6
Performance	6
Scalability	6
VAST's unique data reduction techniques	7
Deduplication	7
Compression	7
Similarity	8
Storage efficiency in practice	8
Benefits of locally decodable erasure codes	8
Benefits of data reduction	10
Conclusions	12

Executive Summary

Data generation rates are outpacing the global manufacturing capacity of storage media at a time when AI is becoming a strategic asset, driving a shift towards long-term retention and always-on accessibility. Sustaining innovation in this climate requires maximizing the efficiency by which data is stored in flash. However, traditional shared-nothing architectures were developed in a time when capacity challenges could be addressed by deploying more hardware, driving a tradeoff where high availability and performance require spending 20–66% of the raw flash capacity on data protection, and data reduction algorithms only work within small partitions of the entire system.

VAST's Disaggregated Shared Everything (DASE) architecture breaks these tradeoffs, reducing data protection overheads to under 3% and allowing global data reduction across exabytes of data without performance penalties. In this paper, we describe how the DASE architecture enables these capabilities and allows significantly more data to be stored on a fixed physical footprint of flash. We then quantify this architectural advantage by analyzing VAST's fleet-wide telemetry covering exabytes of production data. While shared-nothing systems demonstrate overheads between 21 and 28%, the overwhelming majority of data stored in VAST systems operates at less than 12% overhead. In addition, global similarity-based data reduction delivers a capacity-weighted median data reduction ratio of 1.87:1, with some domain-specific datasets exceeding 8.0:1.

These findings demonstrate that VAST's DASE-based architecture effectively decouples data growth from the constraints of the global supply chain. By maximizing the utility of every physical drive, organizations can sustain data growth rates that would otherwise be unattainable through hardware procurement alone.

Introduction

Demand for storage media has grossly outpaced manufacturing capacity in 2026, and the industry has arrived at a point where even the world's largest technology companies cannot buy their way out of this shortfall. Due to the confluence of several economic and technological drivers, there simply isn't enough manufacturing capacity in the world for storage media—whether it be hard drives, flash, or memory—to meet the global demand.

Given that data generation rates show no signs of slowing, the only way to keep up is to focus on efficiency and extracting more usable capacity from the raw capacity that can be deployed. The companies that thrive in the coming years of chronic shortages¹ will be those who can radically innovate in their approach to managing data at scale. Whereas realizing a few percent improvement in storage efficiency may have been a success in the past, the years ahead will require fundamentally new approaches to realize efficiency improvements measured in multiples, not percents.

¹ Pua Khein-Seng, CEO of the world's largest independent provider of SSD controllers, has stated, "[NAND will face severe shortages next year. I think supply will be tight for the next ten years.](#)"

The capacity efficiency solution space

There are only two broad ways in which data growth can be managed in a capacity-constrained world:

1

Keep less data

Enforce tighter data retention policies, apply more aggressive tiering strategies. Force data producers to be accountable for the value of the data they retain.

2

Store data more efficiently

Scrutinize the inefficiencies being incurred in the name of data availability, durability, and performance. Store less redundant data for datasets that do not require it.

Keeping less data (Option #1) is possible, but if data is the lifeblood of an organization, this approach incurs a steep opportunity cost. It risks stifling innovation by forcing engineers and researchers to prescribe value to data prematurely, and it risks destroying data that may have unrealized, long-term value. This ultimately addresses a near-term problem (media supply shortages) with an irreversible solution (deleting data) and is acutely dangerous as emerging AI models are able to extract new insights from large corpora of historic data.

Thus, storing data more efficiently (Option #2) is the only viable option to continue realizing the most value from the increasing volumes of data being generated. If we look at today's state of the art in efficiency at scale, we can see that there are a few major dimensions in which storage efficiency has been traded off:

1. **Durability:** Data is replicated or erasure coded to ensure that failures of individual drives, servers, racks, or datacenters do not result in data loss. Erasure coding typically reduces usable capacity by 15-25%, while replication reduces it by 50-66%.
2. **Performance:** Data is stored predominantly on TLC flash² to ensure high write performance and sufficient endurance. This reduces usable capacity by 25-50% relative to QLC with zero overprovisioning. For IOPS-intensive data, data is often replicated to avoid the latency overheads of read-modify-write inherent in block-based parity schemes.
3. **Scalability:** Data is distributed across independent servers that rely on sophisticated, decentralized control planes to enable bandwidth and capacity to scale together. However, this decentralization also makes global data reduction techniques prohibitively complex, as these methods rely on strongly consistent hash tables. The usable capacity lost is data-dependent but has no upper bound.

These tradeoffs resulted from hard design choices made when hyperscale storage systems were first designed decades ago, but the architectural constraints that motivated them no longer apply. The severe shortage of flash is now a strong incentive for organizations to revisit these tradeoffs and consider alternatives that make the most of the latest in storage, network, and computing advancements to eliminate many of these steep efficiency penalties:

- **Durability:** Standard Reed-Solomon codes used in most systems require minimal computation (to minimize write latency) and minimal metadata (to offer good performance on sequential media). They trade space efficiency for durability and latency, which was sensible when data was stored exclusively on HDDs.

2 TLC holds the majority of the market share of flash as of 2024. QLC remains a relatively small footprint by comparison. See [NAND Flash Memory Market Size & Share Analysis - Growth Trends And Forecast \(2025 - 2030\)](#).

- **Performance:** TLC can endure a wide range of workload patterns without sacrificing performance or endurance. It trades space efficiency for performance across many workloads and allows simplistic storage systems to ignore write shaping and treat SSDs as if they were HDDs.
- **Scalability:** Networks had high latency, so storage was designed to have clients talk directly to storage servers with minimal network hops to maximize throughput. These shared-nothing architectures trade efficiency for scalability in a time before low-latency interconnects.

The result is that many large-scale storage systems that directly serve AI and analytics workloads have high storage overheads and limited ability to reduce data. This makes them a costly choice to deploy when facing capacity constraints.

Eliminating efficiency tradeoffs through architecture

VAST Data's Disaggregated Shared Everything (DASE) architecture benefits from being designed without the aforementioned constraints. As a result, storing data on flash managed by VAST offers significantly lower overheads and higher data reduction capabilities than shared-nothing architectures.

Technology	Architecture	Typical protection scheme at scale	Protection Overhead	Data Reduction
Proprietary S3-compatible object store	Shared nothing	12+4 ³	25%	None
		8+8 ⁴	50%	
Open-source S3-compatible object storage	Shared nothing	8+3 ⁵	27%	Compression-only (64 KiB block size)
		3x replication	66%	
Distributed analytics file system	Shared nothing	6+3 ⁶	33%	None
		3x replication	66%	
VAST	Shared everything	146+4	3%	Compression (global) Deduplication (global) Similarity (global)

Table I. Comparison of data protection and reduction techniques

³ Recommended in *"From 1PB to 100PB: Architectural and Performance Considerations for Massive SmartStores"* (2020).

⁴ From *"Data and Drive parity on AIStor"* (2024), "N/2 each (N being the total number of drives in use by the MinIO server) ... still remains our recommended configuration."

⁵ For example, the Echo Ceph cluster at RAL is described as having 57 PiB usable of 78 PiB raw using 8+3 erasure coding in *"Deploying and Running Ceph Clusters for Analysis Facilities"* (2023).

⁶ For example, Uber has disclosed their use of 6+3 coding with HDFS in *"Infra Meetup | Uber's Storage Evolution"* (2023).

As summarized in Table I, the DASE architecture enables an order of magnitude lower overhead on data protection and data reduction approaches that identify similar or duplicate data across exabyte namespaces to conserve as much physical flash as possible. The following sections describe how this level of data efficiency is made possible by the DASE architecture.

Durability

VAST uses advanced [Locally Decodable Erasure Codes](#) to enable significantly better efficiency than Reed-Solomon codes—with large systems realizing only 3% capacity lost to parity compared to 15–25% for Reed-Solomon and 50–66% for replication as shown in Table I. While VAST's codes are more computationally and memory-intensive than Reed Solomon codes, the DASE architecture allows thousands of CPU cores and petabytes of storage-class memory (SCM) to collectively build these stripes asynchronously.

This results in the same extreme durability as cloud object stores⁷ but with less than 5% of raw flash capacity being lost to data protection. In principle, this is equivalent to a billion-object dataset experiencing a data loss event once per hundred years. In practice, VAST's Locally Decodable scheme has not resulted in data loss in over a trillion object-years of production service to date.⁸

Performance

VAST uses SCM to persist incoming writes instead of relying on high-capacity flash to provide both high random write performance and capacity. By disaggregating the media used to receive new data from the media on which data is persisted en masse, VAST can deliver 33% more capacity per SSD compared to TLC without compromising the latency of writes or endurance of the drives.

In addition, VAST's use of SCM obviates the need for overprovisioning of capacity drives to provide consistently high write performance. While SCM typically represents less than 3% of the total capacity of a VAST system, overprovisioned TLC typically costs 7–28% of total capacity.⁹ VAST's ability to effectively buffer incoming writes through SCM without compromising steady-state performance is enabled by its DASE architecture; every client's data can write to every SCM drive across the datacenter, removing the architectural propensity for localized hotspots, uneven wear, and unpredictable latencies intrinsic to shared-nothing systems.

Scalability

VAST's DASE architecture is built upon a scalable, low-latency RDMA interconnect between SSDs (in DNodes) and CPUs (in CNodes) which allows every CPU to synchronously serve and process all data requests. While this provides independently scalable performance and capacity, it also allows thousands of CPUs to leverage petabytes of SCM collectively to apply sophisticated data reduction algorithms that exceed the capabilities of the individual servers in shared-nothing architectures. For example, DASE enables global, partition-free deduplication because its deduplication table is distributed across all of the shared SCM regardless of cluster size.

The write shaping performed in SCM that enables use of QLC also allows these reduction algorithms to happen asynchronously. Thus, while global deduplication and aggressive compression algorithms typically incur high latencies, VAST hides this latency from applications by applying data reduction as data is being asynchronously written to QLC.

⁷ Eleven nines (99.999999%) of data durability is the standard offered by [AWS S3](#), [Azure Blob](#), and [Google Cloud Storage](#) and matched by [VAST DataStore](#).

⁸ Object-years calculated by multiplying the average number of logical objects in use by the years in service of all VAST clusters with phone-home capability enabled. The average number of logical objects per VAST system is assumed to be half of its instantaneous value as observed in December 2025.

⁹ See Maneas et al, "[Operational Characteristics of SSDs in Enterprise Storage Systems: A Large-Scale Field Study](#)." FAST'22.

The amount of capacity that can be saved through these global-scale data reduction techniques is dataset-dependent. However, across production deployments, VAST's global data reduction delivers over 1.5x effective capacity; a typical exabyte of physical storage managed by VAST supports an additional 500 PB of data.¹⁰

VAST's unique data reduction techniques

This level of capacity efficiency through data reduction is enabled by the flexibility of DASE, which allows VAST to employ layered approaches to finding opportunities to store data with fewer bits. An in-depth description of these techniques is detailed in a separate document, [All Data Reduction Is Not Created Equal](#).

Deduplication

VAST leverages two unique aspects of DASE to deliver significantly greater deduplication efficacy: unpartitioned, global deduplication and adaptive chunking:

Global deduplication allows duplicate data anywhere on the system—regardless of its location in the namespace, location on specific SSDs, or its representation (file, object, table, or block device)—to be identified and deduplicated. This provides significant benefits in multi-user environments where datasets, application containers, and other artifacts are commonly shared and managed by people. Deduplication domains are only limited by the underlying encryption domain in multitenant environments.

Adaptive chunking identifies duplicate data despite shifting byte alignment by using content-defined boundaries. It is effective for datasets with high semantic similarity but unstable layout, such as time-series sensor data (common in scientific and financial workflows), container images and persistent volumes (common in cloud platforms), and high-volume unstructured event logs built from repeated templates with variable fields (common in high-traffic application backends).

The result of this is highly effective, granular, global deduplication. For example, one VAST system in a multi-user academic research environment saw 22% of bytes written being deduplicated within a randomly selected 24-hour period. This duplicate data was identified across the existing 20 PB of user data on the cluster, representing datasets ranging from astronomy and cosmology to biomolecular simulation and genomics.

Compression

As with deduplication, VAST applies compression asynchronously to writes during the time when data is migrated from SCM to SSDs. As a result, it performs byte-granular, data-aware compression that would have prohibitively high latency if performed inline with writes. There are two optimizations that it applies:

- **Floating-point optimization** is a technique where sequences of 32-bit numeric values are transposed so that all values' four bytes are grouped by significance. This allows ranges of data with similar values (such as collections of timestamps) to compress very well, since their high-order bytes are all identical.
- **Delta encoding** is a technique where the difference between successive values is saved instead of saving multiple, nearly identical values. Delta encoding is highly effective in many formats that contain tiles or sequences of data with low entropy between them such as high-resolution image formats.

¹⁰ Measured as the median of data reduction ratios when weighted by physical capacity in use across all VAST clusters. Only clusters that report phone-home telemetry and have similarity enabled are represented. Measurement taken in January 2026.

Because VAST has no a priori knowledge of the content of the data being written to it, these techniques are tested on small samples of written data asynchronously. If one technique is found to greatly improve the compressibility of a piece of data, it is then applied to the rest of the data as part of the process of migrating it from SCM to SSD.

Similarity

Similarity-based data reduction is the final unique approach taken by VAST to reduce the amount of physical capacity required to store data. Like deduplication, this approach leverages the large pool of globally shared storage-class memory within a VAST cluster to fingerprint all of the data written to flash. Unlike deduplication, though, similarity uses locality-sensitive hashes¹¹ to identify old data that is similar, but not identical, to newly written data.

If similar data has been previously written, similarity-based data reduction follows a process similar to delta encoding whereby a diff is calculated between the new data and the similar, older data. This diff is saved to SSD in lieu of the full write, and VAST automatically reconstructs the original data when it is read back. However, unlike with the other data reduction techniques, similarity introduces a small amount of added latency to reads because of this reconstruction. Fortunately, this latency is hidden behind the high concurrency present in the large VAST systems where similarity offers the greatest benefits.

Storage efficiency in practice

As described above, VAST's extreme capacity efficiency comes from the DASE architecture which enables highly efficient locally decodable erasure codes and sophisticated data reduction techniques. These advantages are not simply theoretical though; to illustrate how these two capabilities benefit real environments, we analyzed telemetry from the VAST fleet to understand how real datasets are reduced in production deployments.

In the following analyses, we demonstrate the fraction of data stored on VAST systems worldwide that realize the benefits described above. We do this by weighting data reduction ratio (DRR) and storage overhead of each cluster by the total user data stored on it. This naturally diminishes the effects of smaller or empty clusters which often have spuriously high reduction ratios or low capacity overheads.

Benefits of locally decodable erasure codes

While erasure code rates are a major factor in determining the overheads of data protection on a storage system, all storage systems have additional formatting overheads which increase the gap between physical capacity and usable capacity. To best reflect the baseline user experience, we combine all of these overheads into a single "protection and layout overhead" metric that reflects the percent of raw capacity (number of SSDs multiplied by per-SSD unformatted capacity) that cannot store user data. This overhead assumes zero data reduction; it simply reflects the overheads of protecting data and formatting the storage system.

¹¹ See [U.S. Patent No. 12,007,948 B1](#), "Similarity based compression," assigned to VAST Data Ltd., granted June 11, 2024.

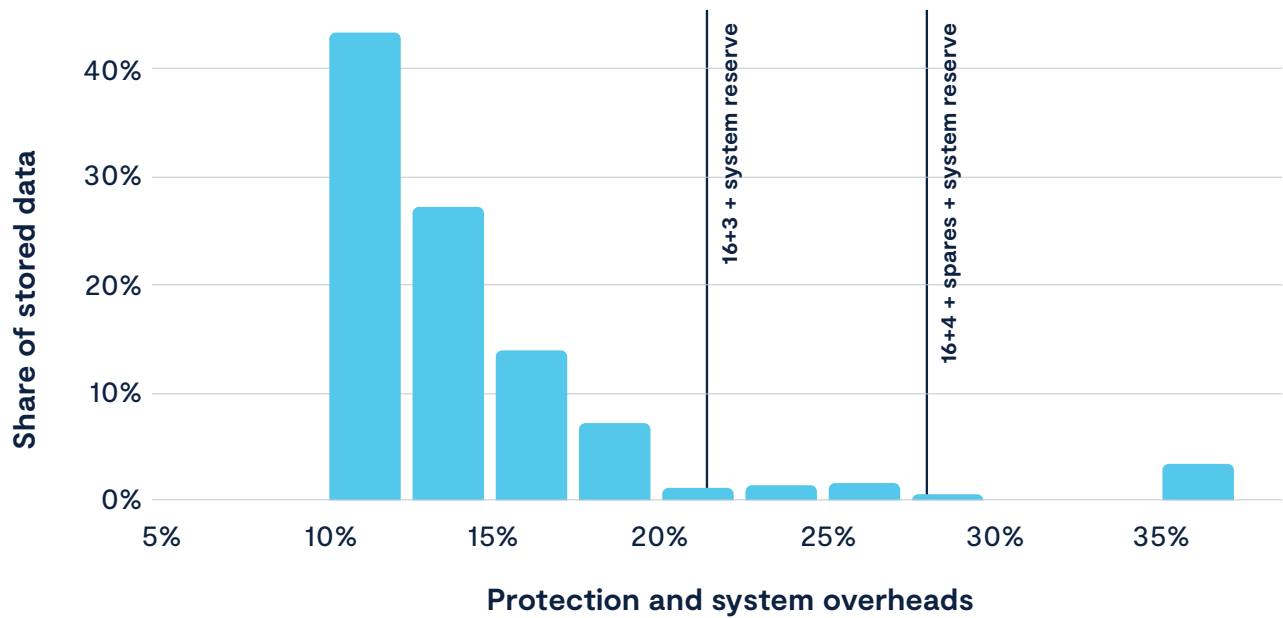


Figure 1. Distribution of measured protection and system overheads across data stored on VAST clusters. Most deployed capacity operates at ~12% overhead, well below the 20–25% overhead associated with conventional erasure-coding approaches that lack locally decodable codes, demonstrating the efficiency benefits of the DASE architecture in production.

We see that the majority of data stored on VAST enjoys only ~12% overhead, which is the minimum given the 3% overhead from the 146+4 data protection and 9% reserved for data management within the system. For comparison, we include two other all-flash, scale-out storage systems that rely on shared-nothing architectures:

- The first technology, which uses 16+3 coding (16% overhead), also reserves an additional 5–8% for sparing and other system functions, resulting in a minimum 21% total overhead.
- The second technology, which uses 16+4 coding (20% overhead), also reserves an additional 8% for sparing and other system functions, resulting in a minimum 28% total overhead.

The distribution of data stored across the VAST fleet indicates that 93% of data stored on VAST realizes less than 21% overhead, and 96% of data stored on VAST has less than 28% overhead before any data reduction is applied.

The remaining data stored on VAST that has relatively high protection/system overheads is concentrated on VAST clusters with exceptionally high resilience requirements. Specifically, the data with >35% system overheads is stored on relatively small clusters that employ [Rack Level Resilience](#). Similarly, the data stored with >20% overhead are concentrated on clusters that employ [enclosure-level resilience](#). As these clusters are expanded to more racks and enclosures, their efficiency will increase as more fault domains allow for wider stripes.

Benefits of data reduction

VAST's sophisticated data reduction techniques are the other method by which VAST allows more data to be stored on physically fewer SSDs. To understand how effective these techniques are in practice, we also examined all VAST clusters with all data reduction capabilities enabled. We plot the cumulative distribution function of these data to illustrate what fraction of data stored on VAST systems achieve different levels of data reduction.

As with the previous data, this distribution is weighted by the amount of data deployed on each cluster to accurately capture what the majority of data—not clusters or customers—experiences. This avoids over-representing small or empty VAST clusters with spuriously high DRR values.

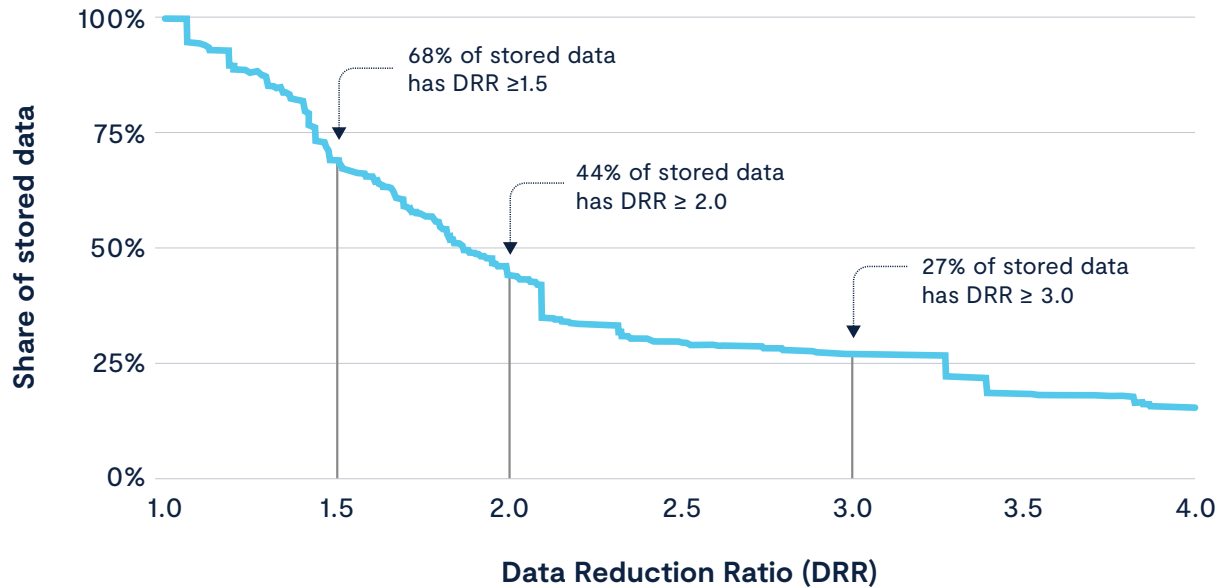


Figure 2. Capacity-weighted complementary cumulative distribution of measured data reduction ratio (DRR) across data stored on VAST clusters. This distribution shows the share of stored data that realizes a DRR of at least x ; the median point ($x = 1.87$, $y = 50\%$) indicates that 50% of data managed by VAST realizes a reduction of at least 1.87:1.

We find that the capacity-weighted median DRR is 1.87, indicating that half of the data stored on VAST systems occupies 53% or less of its original size on physical SSDs. A significant amount of data (44%) has a DRR of at least 2.0, and an appreciable amount (27%) is 3.0 or greater.

In practice, we find that much of this long tail of stored data with high DRRs comes from two sources:

1. **Financial datasets** which are physically massive but have relatively low entropy. These datasets benefit from data-aware compression techniques that specifically optimize financial market data that is logged over time. The sharp cliffs at DRR = 2.09, 3.27, and 3.39 fall into this category.
2. **Data warehouses** that naively store data in triplicate on the assumption that they are running atop unreliable storage. VAST immediately realizes DRR > 3.0 for these datasets, and they account for the long tail of data with DRR > 4.0 on the above distribution.

There is a minority of data with poor data reduction; 32% of data experiences DRR < 1.5. In practice, this data is typically dominated by precompressed formats. For example, one large telecommunications company achieves a DRR of 1.3 because it stores its data in precompressed ORC files. However, their choice to use VAST followed their experience storing these data in triplicate on an HDFS system.

Realistic DRR values are ultimately dataset-dependent, but the above data suggests that outstanding data reduction is achievable across the broad range of application areas where VAST is currently deployed. Anecdotally, we observe the following DRR values:

Typical DRR	Data type / modality
1.1	Precompressed video
1.1 – 1.5	Precompressed financial market data
1.5 – 3.0	Scientific datasets (climate, seismic, and bioinformatics); animation
3.0 – 8.0	Raw LLM training datasets (multimodal)
> 3.0	Log stores, enterprise backups

These ranges of values broadly align with the distribution calculated in Figure 2.

Conclusions

The chronic shortage of NAND flash and other storage media is a foundational constraint of the modern data economy that will persist through 2026. As data generation rates continue to accelerate across multimodal AI and large scale research, organizations can no longer rely on simply adding more hardware to keep pace. Traditional storage architectures are fundamentally ill-equipped for this new reality because they often trade away significant portions of their raw capacity to maintain basic performance and availability.

VAST Data's Disaggregated Shared Everything architecture represents a necessary shift in how data is managed at scale. By decoupling the constraints of durability and performance, VAST enables organizations to extract maximum value from every bit of physical flash. The implementation of key VAST innovations like locally decodable erasure codes and similarity-based global data reduction provides a significant uplift in effective capacity. These efficiencies are proven across the VAST fleet where users achieve high levels of data reduction and efficiency.

In this environment of scarcity, the ability to store data more efficiently is a strategic imperative that determines which organizations can continue to innovate. VAST provides a critical advantage by allowing organizations to continue growing and innovating on their data even when hardware is difficult to obtain. By choosing an architecture that avoids the efficiency tradeoffs required during a time when drives were mechanical and networks were slow, organizations can decouple their velocity from the physical constraints of the storage market. This efficiency is the key to thriving through the current shortage and keeping pace with the ever-accelerating pace of AI-enabled breakthrough.