THINK BEYOND LEGACY SCALE-OUT NAS

Built from the ground up as a modern, all-flash architecture, VAST Data Universal Storage brings an end to decades of storage complexity and compromise

OCTOBER 2021

INTRODUCTION: ELIMINATING TRADEOFFS TO MAKE STORAGE SIMPLE

Most scale-out network-attached storage (NAS) systems, including Dell EMC PowerScale (aka Isilon), were built on 20-year-old "shared-nothing" concepts designed to solve the problems of that time. Since then, data has exploded and the rise of game-changing artificial intelligence (AI) and analytics applications has been a period of reckoning for many enterprise customers, who are questioning old habits, traditions, and systems that no longer serve them. In the era of hyperscale flash, the idea of using a complex, tiered storage system such as Dell PowerScale is one that has stopped making sense.

Of course, every application should be powered by a single tier of flash. But customers have never been able to afford a true, all-flash data center with Dell PowerScale because the cost difference between hard disk drive (HDD) and flash storage systems is extreme. Hence, organizations continue to tier their data between HDD and flash in an attempt to lower their costs. But, this tiered approach means that organizations have to constantly devalue significant data sets by placing them in tiers that are up to a hundred times slower than the network.

VAST Data's Universal Storage redefines these economics, making flash storage affordable for all of an organization's data, from high-performance datasets to the largest data archives. Universal Storage blends game-changing storage innovations to lower the cost of flash with an exabyte-scale file and object storage architecture, breaking the painful storage tradeoffs that have persisted for decades. When an organization stores all of its data in a single tier that is fast enough for its demanding applications, large enough to manage all of its data, and affordable enough that it does not need HDDs, exabytes of data become available in real time. And everything becomes dramatically simpler.

THE RISE AND FALL OF SHARED-NOTHING SYSTEMS

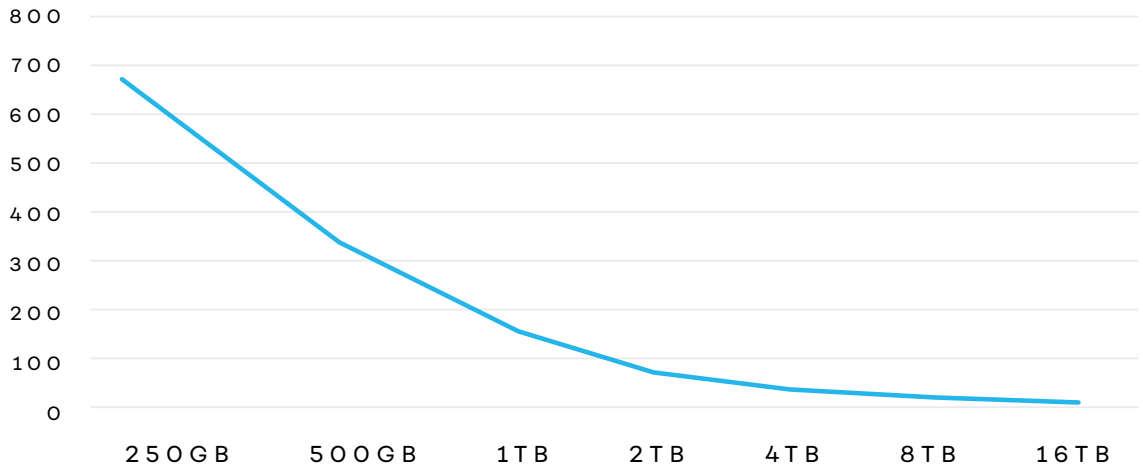
In 2003, a Google whitepaper on the company's new file system changed data storage and management forever. This new distributed data storage system scaled to hundreds of terabytes by clustering commodity server nodes over networking systems, composed of thousands of shared-nothing servers that shared neither memory nor storage. Shared-nothing architecture defined the next twenty years of scale-out infrastructure and includes storage systems such as Dell PowerScale (formerly EMC Isilon), Nutanix Acropolis, Pure Storage's FlashBlade, Qumulo, VMware vSAN, and Red Hat Ceph.

In 2003, shared-nothing systems were a highly effective solution. However, nearly 20 years later, the basic principles on which shared-nothing systems were founded are much less valid. To begin with, data is exploding. IDC estimates that organizations now create 640 times more data per year than in 2003. Today's hyperscale data-centers have requirements for multiple petabytes if not exabyte file systems—not the 100 TB for which the shared-nothing systems such as Dell PowerScale were built.

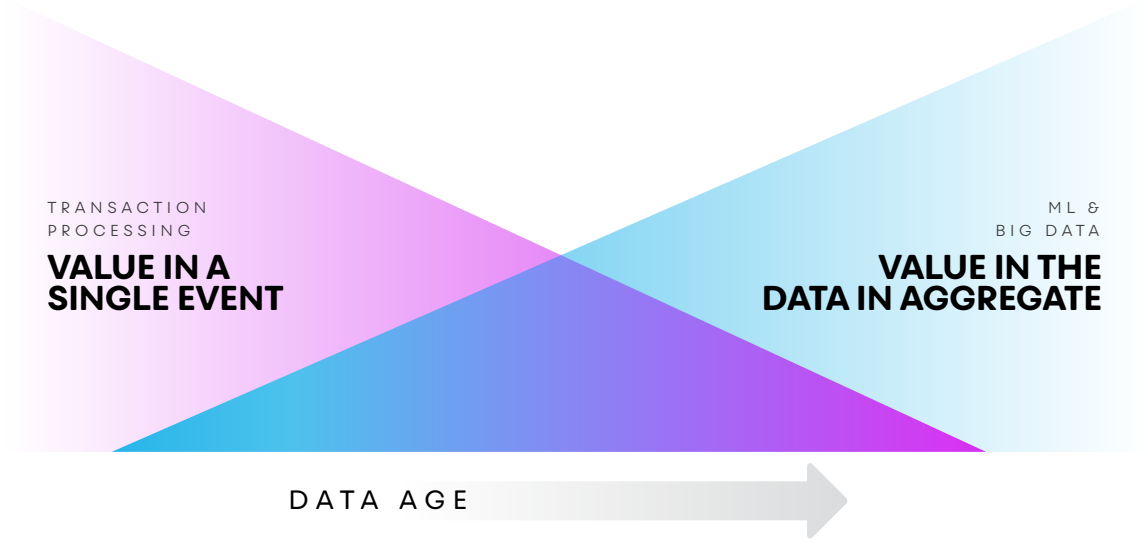
Second, networks are no longer the bottleneck. Shared-nothing systems were designed to co-locate storage with compute, under the assumption that networks were slower than direct-a attached storage (DAS). But, with the advent of NVMe over Fabrics (NVMe-oF), a new high-performance storage networking protocol, storage can be disaggregated from compute without compromising application performance, and much larger systems can be built that can directly connect thousands of servers and storage devices in a single system without any data locality concerns.

Third, shared-nothing systems were architected for HDDs, and HDDs have regressed. Although the disks may have grown in density, they haven't gotten any faster. In many cases, disks in 2003 delivered exponentially more IOPS per gigabyte compared to the denser disks of today, bringing down the cumulative performance of HDD-based clusters.

RANDOM READ IOPS/TB PER HDD HAVE DECREASED



But most important, today's applications demand more than what shared-nothing-based scale-out systems can provide. These systems were designed at a time when no one envisioned the transactional I/O requirements AI would impose on large data stores. The need to randomly read through all events at GPU speed makes affordable flash infrastructure a prerequisite.



NO MORE TIERS

Initially, tiering was introduced as a means to optimize data center costs by deprecating older data to lower-cost (and slower) tiers of storage. While customers may see some savings when applying this legacy model, tiering creates multiple challenges in the long run:

Tiering is Complicated

Having multiple tiers of storage forces storage users to create complex workflows to move and/or copy data back and forth between storage tiers. Users can spend extraordinary amounts of time balancing their applications across a myriad of storage silos, while subjecting their applications to a response time that's much slower than the network access itself. Even with automation frameworks and data management systems, all of this data movement creates integration complexity and reduces overall infrastructure efficiency by creating data copies and wasting precious I/O on data movement.

Tiering is Expensive

Originally designed for HDD capacity, flash was implemented as an afterthought in shared-nothing systems such as Dell PowerScale. Because of its HDD-centric filesystem layout,

PowerScale is unable to write to low-cost flash (such as QLC) in an endurance-optimized fashion. Tiered systems use flash as a high-performance tier for transient data, so applications overwrite this resource constantly, introducing a level of flash wear that ultimately requires the use of expensive, enterprise-grade flash to overcome the limitations of the architecture.

Tiering is Not Ready for AI

As explained previously, flash infrastructure is critical for your modern AI applications, as algorithms need to randomly read across a massive volume of files. But, that's not all. AI models need to be optimized by training on the largest datasets possible. Deep learning models become more accurate once they're run against the entire dataset—not just the 15-30% on an expensive flash tier. So, the classic storage tiering model does not work in the AI era.

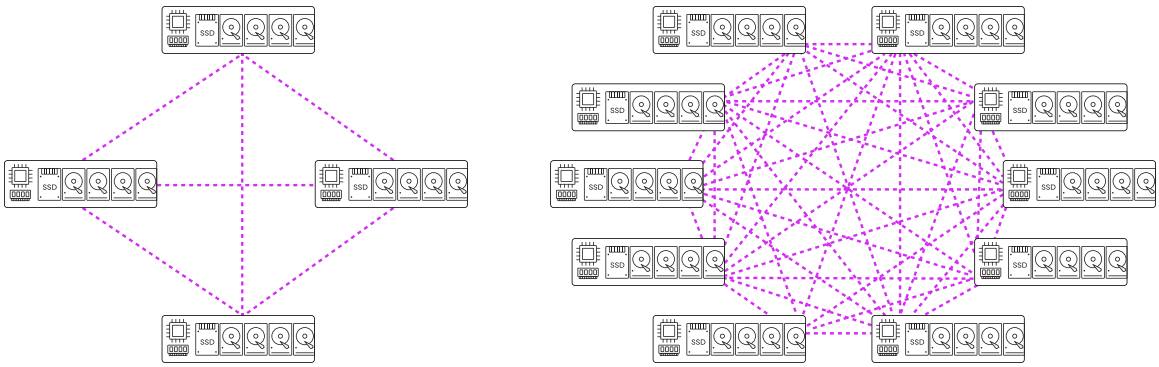
BUT THE PROBLEMS DON'T STOP THERE...

East-West Traffic Hampers Scale on Shared-Nothing Architectures

In shared-nothing systems such as Dell PowerScale, each node "owns" a portion of cluster data, so nodes MUST communicate with each other in order for client read and write operations to complete. Unfortunately, this creates a flood of east-west traffic affecting multiple servers to maintain cache coherency or during node rebuilds.

As systems grow larger, so does the cross-talk within the cluster. This limits the effective performance scale to a few dozen nodes, beyond which there are diminishing returns that limit customers from seeing linear performance gains, especially when running parallel applications on shared-nothing systems.

SERVERS NEED TO COORDINATE
EVERY SINGLE READ/WRITE REQUEST



Resilience and Cluster Stability Suffer, Especially at Scale

A shared-nothing node consists not only of disks or solid-state drives (SSDs), but also of a CPU, memory, disk controller, network controller, an operating system, and storage software. Unlike disks that are typically protected using RAID or erasure coding, these components are not fault tolerant or redundant. So, a single memory or CPU error can cause the entire node to go offline, taking all of its storage offline with it. Most scale-out systems protect against full node failures, but customers need to initiate lengthy, I/O-intensive node rebuild operations of potentially hundreds of terabytes of data, leading to significant performance degradation for applications.

Further compounding this problem is that as scale increases, so does the chance of a node failure. Node unavailability can not only result in performance degradation, but it can also lead to data unavailability when multiple nodes fail simultaneously. Because node rebuilds can be lengthy, this 'window of risk' can span days or even weeks.

Quality of Service is Almost Impossible in Multi-Tenant Environments

East-west traffic makes it almost impossible to protect clients from heavyweight jobs that can inflict significant pain in multi-tenant environments. That is why most customers create a different storage pool for their batch applications, versus more interactive home directory storage. But in shared-nothing systems, batch jobs can overwhelm all the cluster resources—even if the batch application clients are not mounted to all of the shared-nothing servers.

This inconsistent quality of service (QoS) applies to all competing styles of I/O. For this reason, many organizations will deploy different shared-nothing clusters for different users, thereby creating the very islands of storage these systems were designed to eliminate.

Inflexible Infrastructure Scaling Results in Recurring Refresh and Migration Cycles

Shared-nothing systems require clusters to be built from homogeneous pools of servers and storage, forcing customers to make hard-coded and compromised decisions on how they would like to size performance and capacity. Further, the tight coupling of storage and compute in shared-nothing systems, forces customers to scale both capacity and performance in lock-step.

As a result, the decisions customers make on Day-1 will absolutely impact their choices in the future. If their needs evolve, customers are locked in and unable to adjust, leading to wasted storage capacity and performance. These rigid infrastructure boundaries also mean that

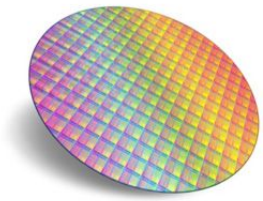
customers have to perform frequent forklift upgrades or migration cycles if they want to take advantage of newer generation of nodes.

The reality is that shared-nothing-based systems such as Dell PowerScale were a great solution to the problems seen in the 2000s and 2010s, but it has become increasingly clear that these types of systems pose significant challenges when customers try to operate them at scale for their modern, next-generation AI and analytics applications.

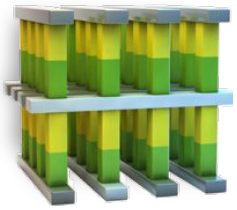
INTRODUCING DASE: STORAGE ARCHITECTURE OF THE FUTURE

VAST Data's Disaggregated and Shared-Everything (DASE) architecture eliminates many of these tradeoffs and compromises that have emerged through the global adoption of shared-nothing architectures. Starting with a blank sheet of paper, VAST Data designed a high-performance, highly scalable, and highly agile enterprise storage platform that uses three new storage technologies, NVMe-oF, low-cost hyperscale flash, and persistent storage class memory, in an innovative way to solve the historical problems of shared-nothing architectures.

NVMe-oF enables remote NVMe devices to be accessed with direct-attached performance.



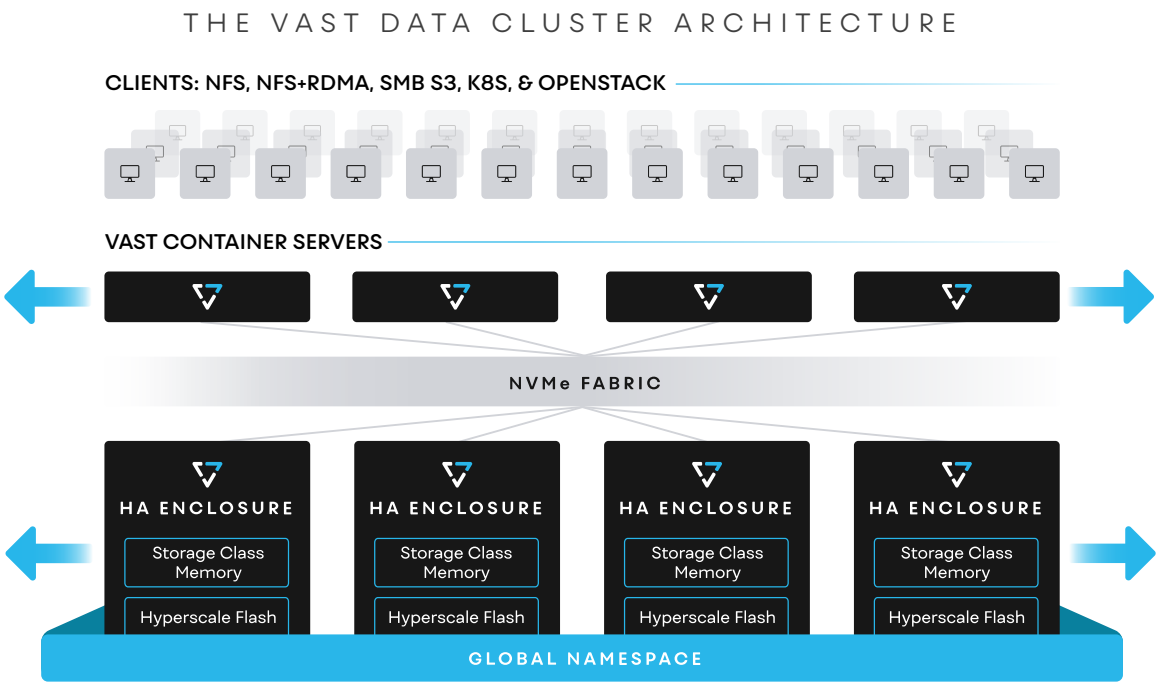
Hyperscale Flash costs dramatically less than enterprise flash while delivering enterprise levels of performance.



Persistent Storage Class Memory is used to reliably buffer writes to low-cost hyperscale flash and create large, global metadata structures.

The use of NVMe-oF makes it possible to disaggregate CPUs from the state of the cluster and for each stateless containerized storage server to simultaneously access all of the data. When servers are disaggregated from storage, everything is better:

- Servers are stateless, and failures of any server never require data reconstruction across a network.
- Servers are loosely coupled in a scalable cluster, each operating independently while all accessing a shared, global namespace, thereby eliminating cluster cross talk and enabling virtually limitless scale.
- New, global algorithms are implemented to achieve game-changing levels of efficiency and resilience.



Declared as "the storage architecture of the future" by IDC in 2020, DASE lays the foundation for game-changing levels of scale, efficiency, resiliency and QoS such that it's finally possible– and affordable– for companies to consolidate all of their data and all of their applications onto one single and universal tier of flash infrastructure.

Enable Radical Storage Economics

Legacy storage systems were not designed to work with low-endurance hyperscale flash

devices that cost significantly less than traditional enterprise flash. VAST data's DASE architecture resolves the performance and endurance challenges of low-cost flash, enabling their use over a decade of longevity. To underscore the viability of this design choice, VAST Data also provides a 10-year flash media endurance guarantee.

Global Wear Leveling and Write Amortization: Flash wear leveling is done at a global level, and the system balances the needs of transactional workloads with the needs of archive data to deliver higher endurance.

Enhanced Data Protection Efficiency: VAST's data protection codes not only support more concurrent failures in a failure domain, but also allow significantly wider stripes, which brings the system's overhead down to just 2.7% to 11% (vs. other systems that carry 30%-66% overhead).

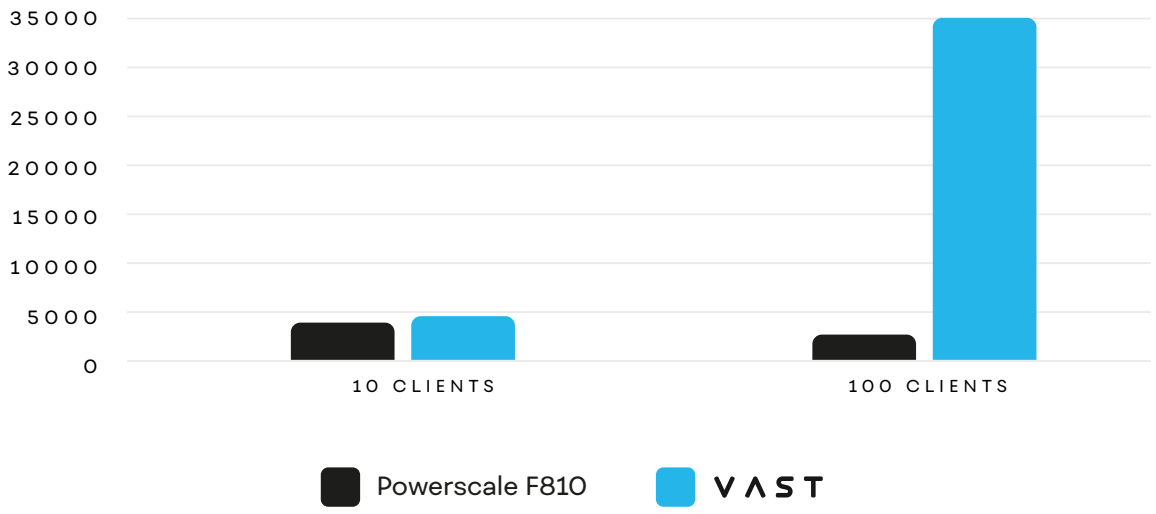
Similarity-Based Data Reduction: Similarity-based data reduction technology combines the global nature of deduplication with the fine-grained byte-granularity of pattern matching, resulting in unprecedented levels of storage efficiency—without compromising performance or endurance. When run against real-world datasets, including unstructured data, backup data and even compressed data, VAST Data's Similarity-Based Data Reduction is designed to deliver anywhere between 2:1 to 20:1 storage-level data reduction.

Enable Simplicity and Resilience at Scale

VAST Data's Universal Storage is purpose-built for the exabyte age, so that enterprises can achieve virtually limitless scale, while preserving the stability, resiliency, and performance of their cluster.

Linear, Predictable Parallel Scale: VAST's DASE architecture provides direct-attached levels of storage access from every NVMe SSD to every stateless storage container in the cluster. This new shared-everything approach ensures that no machine needs to talk to any other in the data path. No east-west cluster traffic means every node adds a proportionately linear amount of performance to the cluster, allowing customers to achieve massively-parallel scale without having to use complex HPC file systems.

DASE ENABLES PREDICTABLE SCALE AS YOU ADD NODES

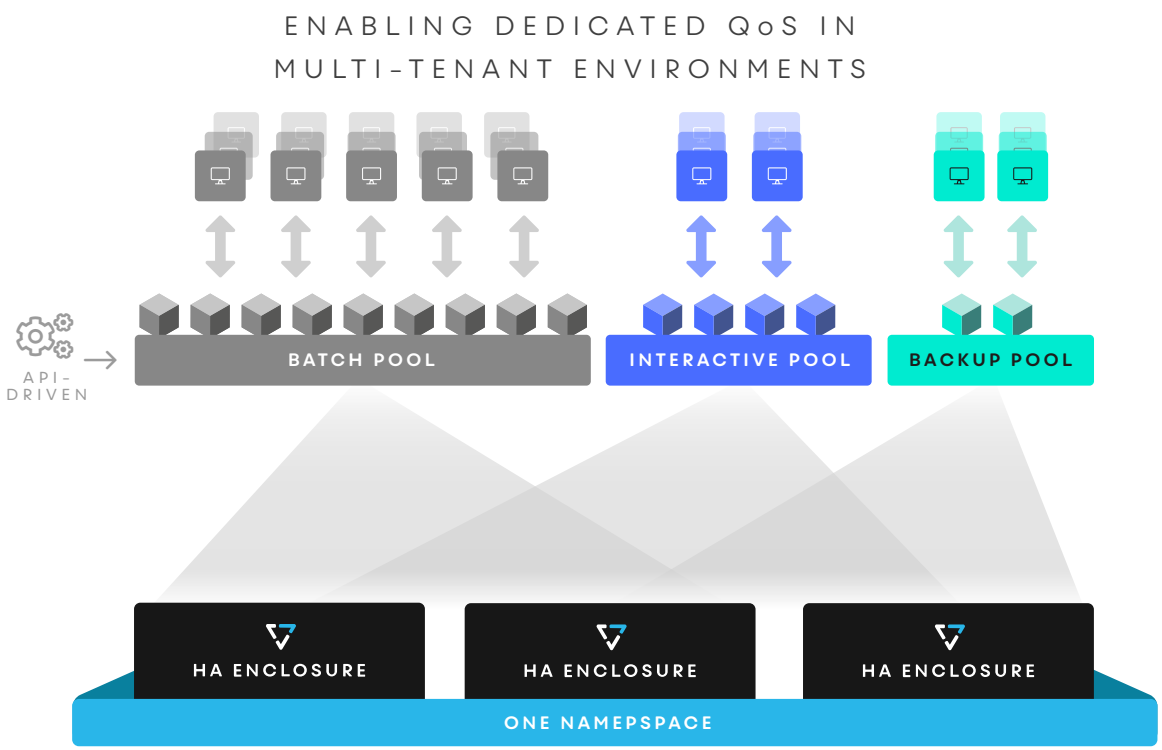


Postgres Testing Performed by a US Government Customer

Enhanced Resilience: VAST combines stateless container-based servers with a stateful protocol and metadata layer to maximize uptime for performance workloads at scale. Containers make it simple to deploy and scale Universal Storage as a software-defined microservice while laying the foundation for a much more resilient architecture, where container failures do not trigger a rebuild event and are, hence, non-disruptive to system operation.

Independent Compute And Storage Scaling: By breaking away from the classic 'shared-nothing' model of building storage, customers can now scale compute independent of capacity. This means that they no longer need to buy capacity when all they need is performance, and vice versa.

Dedicated QoS For Multi-Tenant Environments: No east-west traffic means compute servers can be composed into pools that provide dedicated QoS for demanding applications, while ensuring data isolation across multiple tenants/ applications. Pools can be composed and recomposed dynamically using APIs, thanks to their statelessness.



SUMMARY

Enterprise data centers have requirements for multiple petabyte—or exabyte—file systems, more than what outdated shared-nothing systems such as Dell PowerScale can support reliably and affordably. For such data centers, VAST Data Universal Storage makes flash affordable for ALL of their data and applications, from the highest performance databases to the largest data archives.

VAST Data's disaggregated, shared-everything approach, enabled by new solid-state storage technologies such as NVMe-oF, hyperscale flash, and persistent storage class memory overcomes the performance, availability, recovery, and cost disadvantages of legacy designs, while enabling workload consolidation at scale. The result is a modern architecture that is more powerful, more simple, more flexible, and less expensive than traditional storage.

In the the following chart below, you can see the shows the unrivaled benefits of VAST DASE architecture over shared-nothing architectures.

	Dell EMC Isilon/Powerscale	VAST Data's Universal Storage
Cluster Architecture	Shared-Nothing	Disaggregated Shared-Everything
Cluster Orientation	Symmetric	Asymmetric
Rebuild Upon Server Failure	✓	✓
NFS, SMB, S3	✓	✓
All-Flash Nodes	✓	✓
Hybrid Nodes	✓	✗
HDD Nodes	✓	✗
Supports Low-Cost QLC Flash	✗	✓
Erasure Code Overhead at Scale	20%	2.7%
Global, Cross-File Compression	✗	✓
Composable QOS Using Disaggregated CPU Pools	✗	✓
All-Flash Effective Cost/GB	$$$$$$	$
Hybrid Storage Effective Cost/GB	$$$	$
HDD Node Effective Cost/GB	$	$
Disaggregated Sourcing Model	N/A	Gemini

© 2021 VAST Data, Inc. All rights reserved. All trademarks belong to their respective owners.