VAST

INTRODUCING

EXASCALE NAS

VAST Data Defies Convention To Make Simple And
Embarrassingly Parallel NFS Storage a Reality For Scalable HPC



White Paper | June 2020

INTRODUCTION

Somewhere in a parallel universe, Network File System (NFS) made the graceful jump from Unix workstations to massively parallel supercomputers, and no one ever felt the need for a purpose-built, complicated HPC file system.

Back in our universe, around the turn of the century, the HPC community broke away from standard storage protocols in favor of parallel file systems that provided support for parallel streaming operations, plus scalability that is far beyond what the NAS file servers of the day could deliver. With custom SW clients, these systems could leverage modern networks to stream data over RDMA-enabled interconnects such as InfiniBand to achieve far more bandwidth than TCP based connections could deliver. Since then, HPC centers have suffered from the challenges created by parallel file systems, including wrestling with client-side file system SW and trying to overcome centralized and coordinated metadata activities. As bandwidth challenges have been solved, a growing problem around handling distributed metadata has been left in the wake. As IOPS-intensive analytics workloads become more prevalent, the bandwidth-centric nature of many of these parallel file systems is becoming an analytics inhibitor as much as they were previously a simulation enabler.

Furthermore, the move from NAS appliances to integrated file systems has encouraged HPC centers to cobble together disparate file and storage systems that together receive nowhere near the polish nor the investment that the broader NAS community enjoys from much more portable and featureful storage appliances. Scalability has come at the expense of uptime and operational simplicity.

Simplicity and scale have never been words that HPC storage architects could combine in a single sentence—until now. VAST Data has created a new system architecture that solves the scalability challenges of NFS, adds RDMA support to maximize the use of high-speed networks, and combines an embarrassingly parallel system architecture with a revolutionary approach to flash efficiency to make scale-out flash affordable for all HPC applications.

SOME HPC STORAGE BACKGROUND

To many, the idea that a scale-out all-flash NAS platform can be used to replace everything from cloud storage systems, to parallel file systems, to burst buffers may seem too good—or

foreign—to be true. Today's big cloud builders who engineer their own platform designs,making storage software, processors, memory, storage, and networking for their very precise application needs, have moved beyond general-purpose protocols such as NFS because they own all of their applications and can engineer around a slimmed-down set of application requirements. The HPC community, which endorsed NFS wholeheartedly from the early days of democratized simulation and modeling on Unix systems in the 1980s through the 1990s, has, by and large, adopted parallel file systems such as Lustre and IBM's GPFS because NFS was never designed to scale to meet the needs of parallel computing environments.

That said, NFS is simply a client-access protocol. Despite its name, it is not itself a file system. While most NAS systems were not designed to scale, it's possible for file systems that underlie NFS to be built with massive namespace and infrastructure scalability, all while exposing said namespace either via standard NAS protocols or with custom clients that run custom client side protocols. Evidence of this is that even Lustre and IBM Spectrum Scale (aka: GPFS) today also support accessing their systems via both NFS and SMB. With these products and others, the basic reasons why HPC centers abandoned NFS don't always apply.

With the right underlying storage and namespace architecture, NFS file services can scale across a distributed network attached storage array and work just fine for supporting HPC simulation and modeling workloads. With new approaches, this filesystem access protocol can scale to exascale proportions, can meet the needs of new demanding AI applications and (with the right appliance packaging) can dramatically reduce the deployment and administration burden that many supercomputing centers take on.

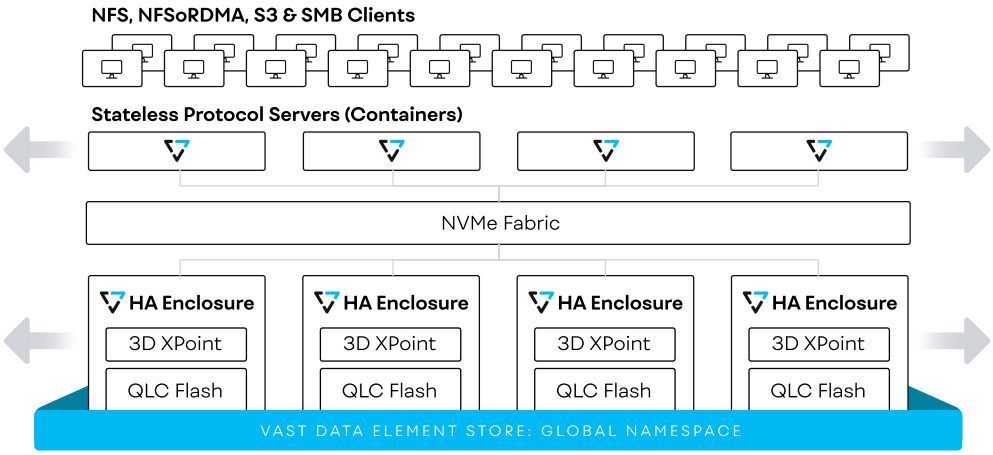
QUICK DOWNLOAD: VAST DATA INTRODUCTION

VAST Data's Universal Storage redefines the economics of flash storage, for the first time making flash affordable for all applications, from the highest performance datasets to the largest data archives. The Universal Storage concept blends game-changing storage innovations to lower the cost of flash with an exabyte-scale file and object storage architecture breaking decades of storage tradeoffs.

With the advantage of enabling technologies that weren't available before 2018, this new Universal Storage concept can achieve a previously-impossible architecture design point. The system combines low-cost QLC flash drives and 3D XPoint memory (such as Intel® Optane™

Technology) with stateless, containerized storage services that all connect over new low-latency NVMe over Fabrics networks to create VAST's Disaggregated Shared Everything (DASE) scale-out architecture. The software underlying Universal Storage applies next-generation global algorithms to this DASE architecture to deliver new levels of storage efficiency, resilience, and scale.

While the architecture concepts are sophisticated, the intent and vision of Universal Storage are simple: to bring an end to the data center HDD era and end the complexity of storage tiering that is a byproduct of the decades of compromises caused by mechanical media.



VAST Universal Storage is based on a new scale-out architecture consisting of two building blocks that are scaled across a common NVMe fabric. First, the state and storage capacity of the system is built from resilient, high-density NVMe-oF storage enclosures. Second, the logic of the system is implemented by stateless Docker containers that each has the ability to connect to and manage all of the media in the enclosures. Since the compute elements are disaggregated from the media across a data center scale fabric, each can scale independently—thereby decoupling capacity and performance.

By breaking the long-standing price/performance tradeoff that has held back HPC applications, VAST has quickly won the attention and support of many of the world's largest supercomputing centers. VAST Data systems are now deployed in many top 500 data centers where they are used for storage applications ranging from simulation scratch, to deep learning, to project and home directory stores.

By eliminating the need for intra-cluster coordination and by scaling across high-throughput commodity networks, Universal Storage makes it possible to deliver at exascale proportions:

Max # of VAST Servers Supported

10,000 Storage Servers
scale to over 100TB/s, over 500M IOPS

Max # of VAST Enclosures Supported

1,000 NVMe Enclosures
675PB (raw), 1.5EB (at 2.5:1 data reduction)

NFS FOR EXASCALE: UNDERSTANDING THE APPLICATION EXPERIENCE

For HPC customers, the four most important aspects of a file system are application compatibility, client performance, scalability, and price. VAST improves the experience for applications that have more recently become accustomed to parallel file systems for cluster I/O, so let's look at these one by one.

Application Compatibility

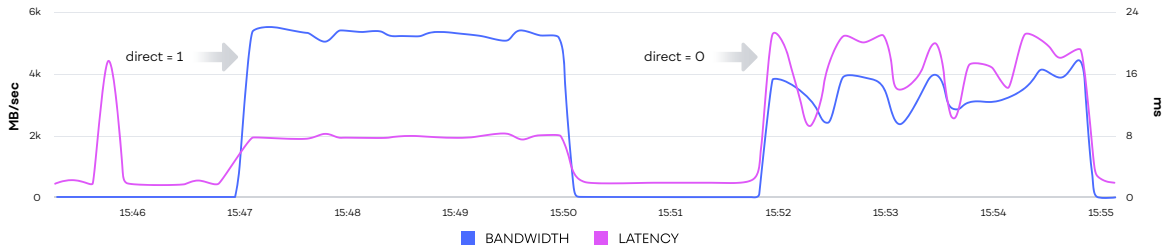
Cache Coherence

Not all exascale applications require strict global caching semantics. In fact, many customers run well-designed applications that create a file per process that don't require locking or cache coordination to ensure global namespace consistency. On the other hand, not all HPC IT organizations have control over user behaviors and many do not know about or care about how to achieve consistency when using parallel I/O.

One of the HPC community's principal concerns is the close-to-open problem that stems from the fact that NFSv3 clients support client-side caching, and at the same time does not coordinate cache evictions in the case of a parallel file or directory operation that extends across multiple NFS clients. NFS was never designed for atomic parallel I/O: If two clients on a network are writing to a cached copy of the same file and issue a close command in close proximity, the protocol cannot ensure which bits will become the authoritative version of data, leading to the possibility of data inconsistency.

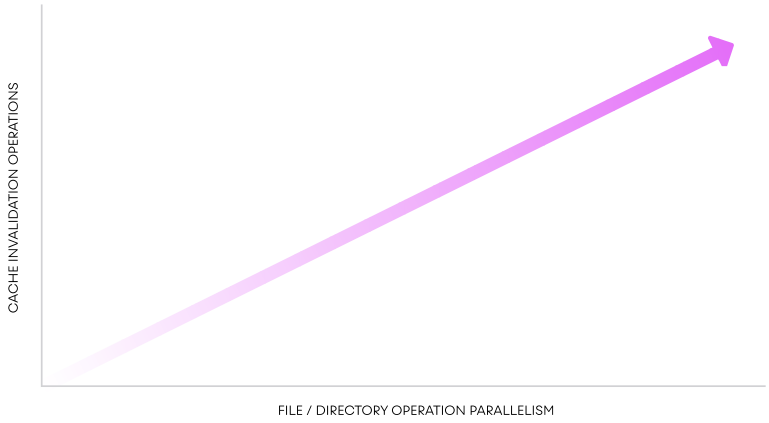
To understand the problem, it's important to understand the underlying premise: Client caching is an age-old method of making sure that networks are well utilized and that HDD storage receives large, sequential well-formed I/Os from the client and is one of the best ways to ensure good performance utilization of mechanical media. Write caching has helped everything from NFS to parallel file systems accelerate performance. However, in an age of solid-state infrastructure, the same requirements for well-formed client side I/O don't necessarily apply.

It's counterintuitive, but the pagecache flush by a cached NFS client will put backpressure on the CPUs of a VAST Server. WIthin a VAST cluster, all data is first written to 3D XPoint media, therefore any attempts by NFS clients to sequentialize the data don't provide the same benefit that caching once did for HDD storage. Everything, in a VAST system,is optimized for random I/O, and alignment no longer matters. With VAST Data's Universal Storage, the system would prefer to ingest a synchronous dribble of atomic writes so it can use the CPUs more consistently and evenly over time. When disabling client caching by using the **sync** or **o_direct** options, writes are never cached and become atomic operations to the VAST storage cluster. Without caching, this approach solves the client consistency issue often found with legacy NFS systems while also providing the benefit of delivering superior sustained write performance.

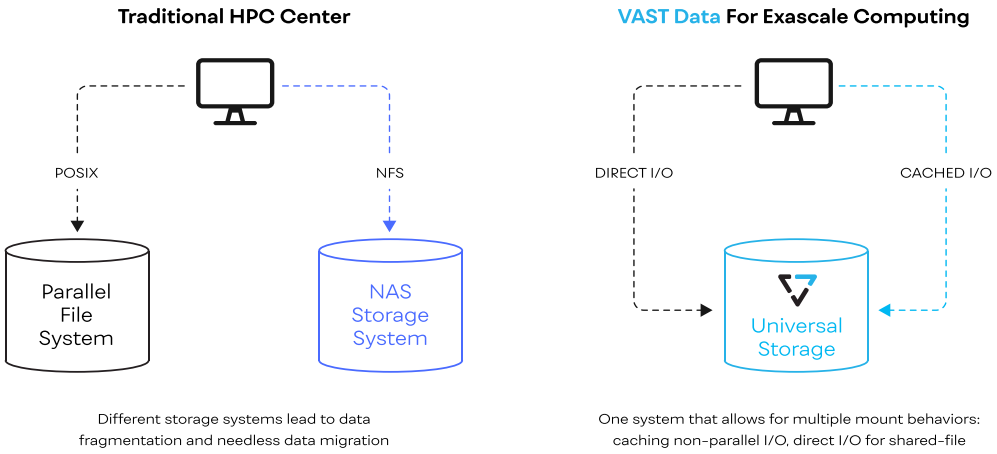


Caching Optionality

Now, many applications also benefit from having caches=on, particularly when it comes to frequent metadata operations (eg. stat calls). As we've studied this requirement, we've come to understand that there are applications where NFS caching can be utilized and there are other instances where caching should be disabled. If an organization knows this in advance, which they often do, it's easy to make distinct choices to cache or write synchronously to the same cluster. In many ways, even parallel file system caches deliver less benefit as parallelism increases on a single piece of data. The observation is simple:



NFS caching is used for everything from home directory I/O to being an accelerant to applications that need more transactional performance than what parallel file systems have historically been capable of delivering. In the past, customers would deploy independent parallel file systems and NFS systems to satisfy the needs of both types of jobs... with VAST, this can be consolidated onto one system:



Global File Locking

VAST NFS also supports the NLM byte-range locking protocol. NLM, which originally stood for Network Lock Manager, defines a standard mechanism for NFS clients to request and release locks on NFS files and byte ranges within those files. NLM locks are advisory, so clients must test for and honor locks from other clients. NLM provides the support for shared and exclusive locks to applications and is designed for parallel applications where many byte-ranges can be locked concurrently within a single file. VAST's approach to NLM locking is inherently scalable because locking and lock management is fully distributed across the VAST Cluster.

Unlike some parallel file systems that use central lock managers, VAST clusters leverage the DASE architecture to eliminate the need for centralized lock management. Instead, lock information is stored as extended file system metadata for each file in the VAST V-Tree, distributed globally across the system's 3D XPoint. Since all of the system metadata is available to all the VAST Servers in the cluster, each VAST Server can create, release, or query the lock state of each file it's accessing without the central lock manager server that can so often become a bottleneck on other systems.

Client Performance

RDMA Support

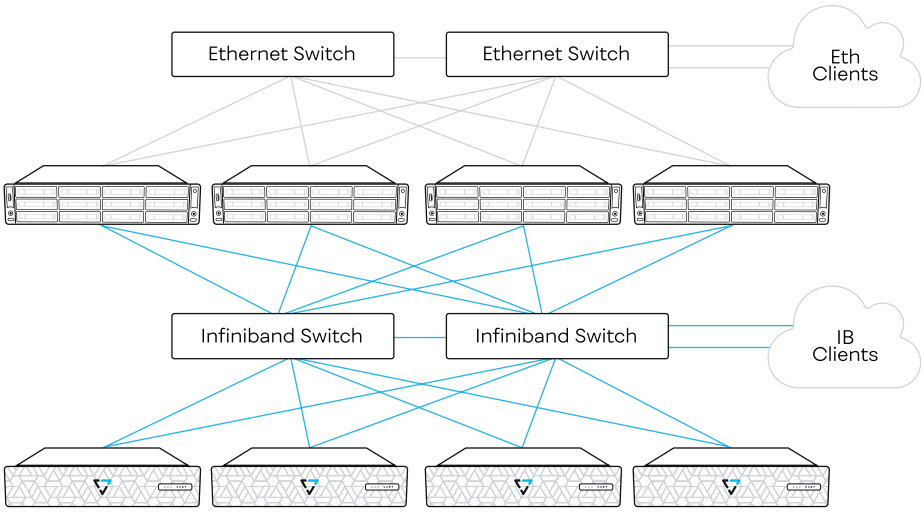
The market's most successful NFS appliances have only ever been built for TCP networks, which creates two problems for HPC centers:

1. These systems are difficult to connect into clusters that only have RDMA interconnects, such as InfiniBand, without some additional I/O network or network gateways… both of which introduce cost and complexity.

2. Client performance has been limited to what a single TCP connection can provide. This limits a client's ability to saturate a network connection.

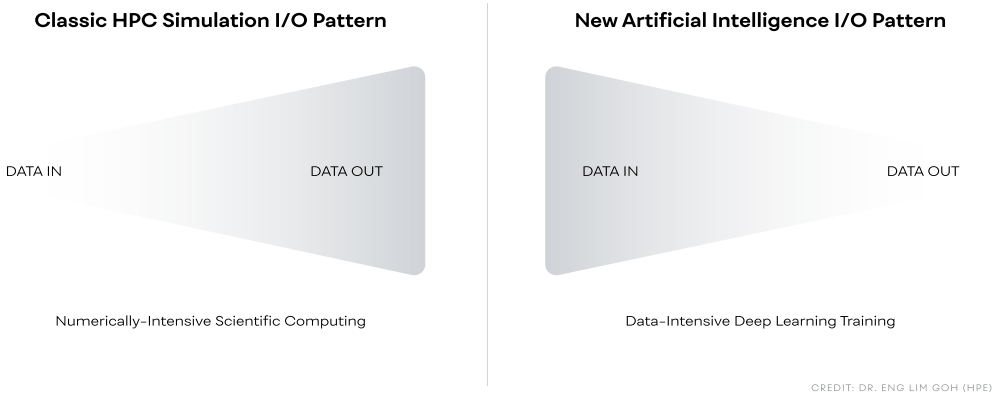
Here, VAST has a simple and clean solution. Many years ago, Oracle added RDMA extensions to the NFS kernel as a means to improve NFS performance beyond the limits of TCP. VAST Servers can be mounted over either NFS or NFSoRDMA. The result is that a single NFS client can nearly saturate a 100Gb connection (11GB/s over EDR has been observed).

VAST Clusters can also be connected over Ethernet or InfiniBand networks, or both. Without the need for network gateways or storage protocol "routers" VAST servers can be homed to one or many networks simultaneously. With the VAST Server Pooling concept, the cluster can be mounted to multiple subnets without needing each subnet to have access to the entire Server pool… this is a key benefit of disaggregation: each Server has access to the global state of the cluster over NVMe-over-Fabrics. The architecture lends the system to a good number of topology options when deploying a cluster across multiple networks. A simple example is provided here:



Client I/O that's ready for GPU computing

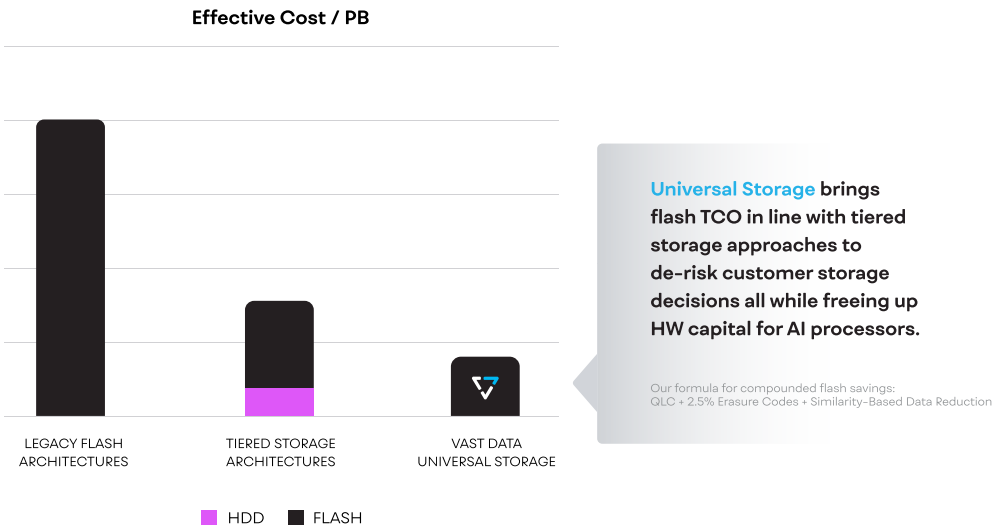
Artificial intelligence applications are turning long-standing HPC storage requirements on their heads. Not only do these applications require random access to read large stores of training data, but they also need to concentrate this performance into high-density GPU/AI computing servers. Dr. Eng Lin Goh summed up the differences between HPC I/O and AI I/O best:



CREDIT: DR. ENG LIM GOH (HPE)

For years, HPC centers have been conditioned to focus on write throughput as the principal method of sizing HPC storage. This is the market that gave birth to the burst buffer. On the other hand, as AI methods encroach into CPU and GPU environments, organizations are faced with not only a growing requirement for read performance (95%+ of deep learning traffic is read-oriented), but also a realization that these reads are small, random, and cannot be deterministically pre-fetched into some buffer or cache layer in a storage environment. Cache misses that go down to HDD storage can cost organizations as much as a 98% loss in performance and considerably impact AI computing efficiency.

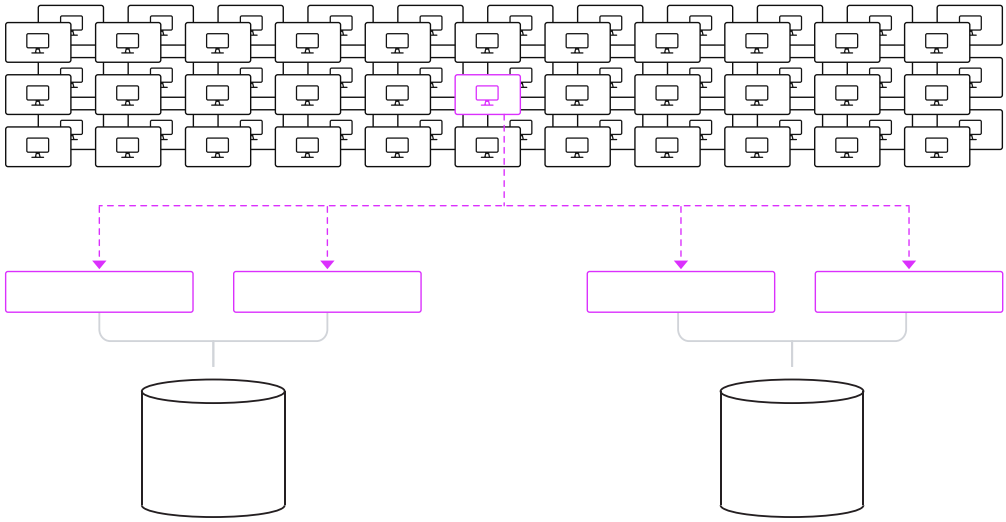
VAST's marrying all-flash infrastructure with HDD economics enables organizations to service AI clients with high-density throughput as well as limitless IOPS.



NFS Doesn't Broadcast The Pain

As VAST has gotten more traction within computing centers that previously used parallel file systems, we've also gained a greater appreciation of the performance isolation capabilities of NAS and how this can be used to resolve much of the noisy-neighbor problem.

- Parallel file system clients, because they mount all of the file servers simultaneously, can expose the entirety of the storage environment to the pathological behaviors of one bad user or application. As such, their inherent parallelism makes them—in many ways—ill-designed for multi-user and multi-disciplinary HPC environments. If one user goes crazy with a wild metadata operation or a big bandwidth storage operation, the impact can be felt upon all of the underlying parallel file system storage cluster and all its controllers.

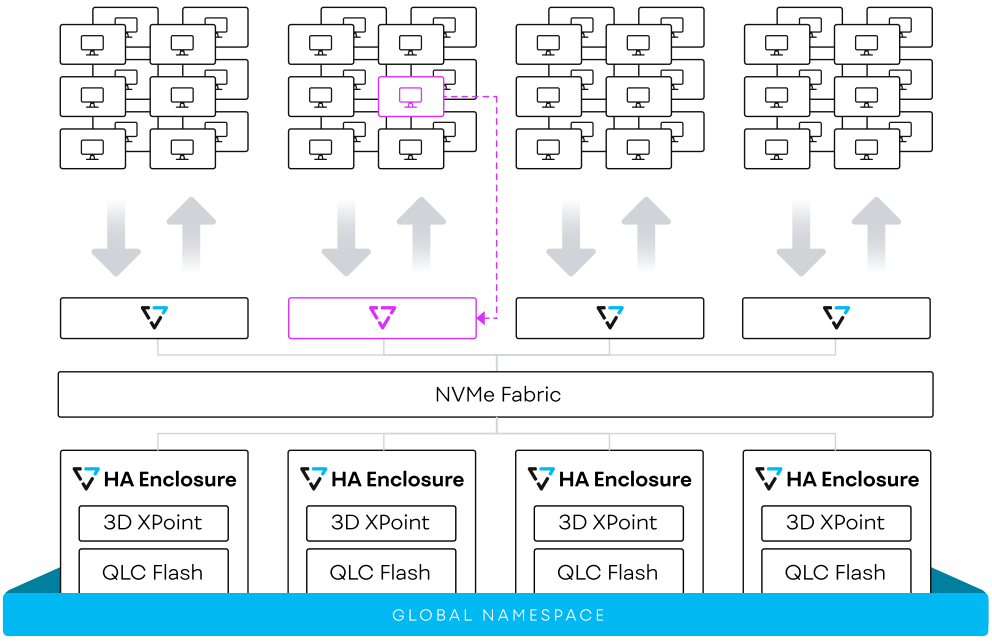


Parallel file systems can 'broadcast the pain' of one bad client across the whole storage cluster

- NFS clients, on the other hand, create a single connection with one of many file servers in a scale-out NAS system. Legacy NAS clusters, which ascribe to the shared-nothing design, receive the requests from one overwhelming client and broadcast them across the rest of the machines; because these activities need to be coordinated. VAST, on the other hand, has eliminated the need for any east-west traffic in the cluster thanks to the Universal Storage DASE architecture.

  Disaggregation combines with NFS to minimize the impact of any one bad actor to just the file server that this client is mounted to. While the clients mounted to this same VAST Server

will feel the pain, all of the other clients mounted to all of the other servers will be completely isolated from this activity. It's cleaner than legacy scale-out NAS and much cleaner than how parallel file system clients broadcast their pain broadly.



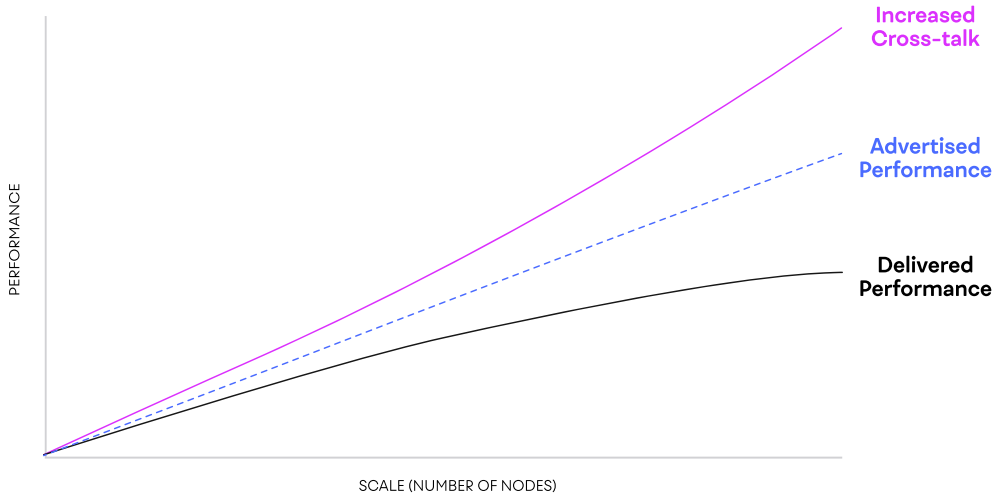
NFS + Disaggregation limit the impact of any bad user/application

Namespace Scalability

Embarrassingly Parallel

Many shared storage systems can experience laws of diminishing returns as organizations scale them out in an attempt to meet the needs of supercomputing applications. The root cause is often the underlying shared-nothing approaches that storage vendors take to build scalable systems. There are a variety of implementations of shared-nothing that can be found in everything from conventional scale-out NAS systems to today's leading parallel file systems.

Sharing comes at a price, however. Because each cluster node has to coordinate data and metadata activities with other cluster nodes, the cross-talk relating to cluster coherency and storage rebuilds limits the effective performance of shared-nothing systems. As systems grow larger, so does the chatter within the cluster. Most commercial scale-out storage appliances don't scale beyond a few dozen nodes before the law of diminishing performance returns limits organizations from seeing linear performance gains.



VAST's new DIsaggregated and Shared-Everything (DASE) is built from stateless containers which each mount all of the system's flash and XPoint devices over a low-latency, data center scale NVMe fabric. DASE systems eliminate east-west cluster traffic to deliver linear scale units of performance scalability as the cluster grows in both capacity and CPU.

There are many examples where scalable storage architectures have shared or centralized processes.

- Metadata-server based architectures, such as Lustre, centralize metadata activity and distributed lock management around a small collection of metadata servers
- SAN-based and shared nothing architectures, such as IBM Spectrum Scale, require communication across nodes in a cluster for locking, metadata coordination and other internal functions

With Universal Storage, each VAST Server provides both the data and the metadata of the entire cluster, thus eliminating the need for independent and centralized metadata servers. In addition, no two stateless VAST Servers talk to each other AT ALL in the synchronous write or read path, none of them have cache, and there is zero requirement to enforce global coherence since the underlying globally-accessible data structures are transactional and atomic across the whole storage cluster. As a result, VAST is the MOST scalable distributed NAS system ever developed, and even scales more elegantly than legacy parallel file systems.

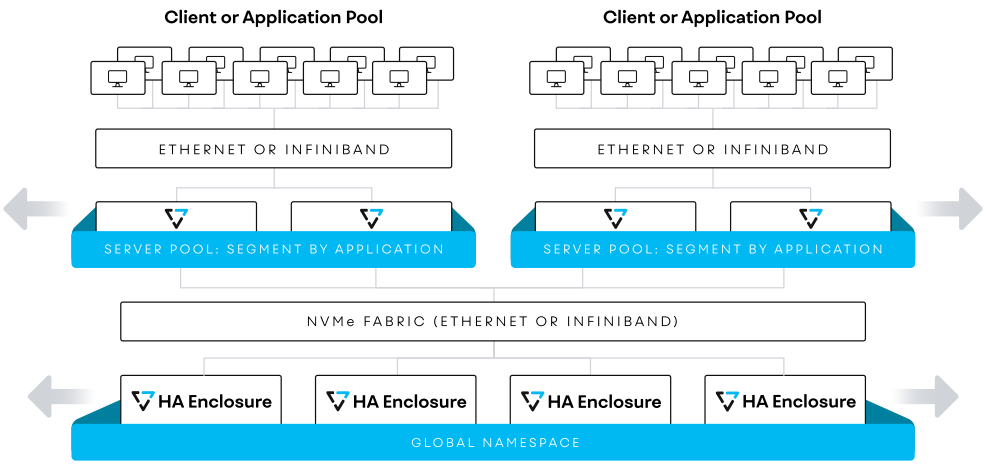
Many of the world's largest computing centers tell us that the bandwidth challenges of yesterday

have all been solved. As modern environments scale to millions of cores, these legacy parallel file systems have created a new bottleneck, where legacy file systems can't service the distributed metadata transaction requirements of these many-core systems. VAST Data's disaggregated, shared-everything architecture solves this challenge, where scalable systems can now deliver 100Ms of data OPS and IOPS to millions of compute cores.

Server Pools For Consolidation and QoS

Scalability is relative. While many parallel file systems deliver large levels of scale for scratch for large petascale to exascale sized machines, many of these storage systems become single purpose clusters designed to service the needs of only a small part of an overall computing center. In many environments, you'll find dedicated storage systems assigned to each large cluster and then other NAS, viz, archive and analytics storage systems elsewhere in the same data center. The reason for this typically traces back to the same problem we address above with respect to performance isolation: sharing.

The VAST Servers in a cluster can be subdivided into Server Pools that create isolated failover domains, making it possible to provision the performance of a arbitrarily-sized pool of servers to a set of users or applications in order to isolate application traffic and ensure a quality of ingress and egress performance that's not possible in shared-nothing or shared-disk architectures.



Server pooling also provides the advantage of being able to support multiple networks simultaneously. Users can build their backend fabric from a common Ethernet or InfiniBand storage network, while also provisioning additional Server front-end ports to talk across multiple, heterogenous, Infiniband, or Ethernet subnets. Pooling makes it easy to provide file services to hosts on different networks that all access a global namespace.

With Universal Storage, the promise of data center consolidation becomes a reality. Pools are now in use by many of VAST's largest customers, and there are many use cases:

- A scratch pool can service the batch needs of a supercomputer, while another pool can be scaled to meet the interactive I/O requirements of users that access the VAST Cluster either from their login nodes or via their desktops. This approach isolates the interactive users from the punishing workloads of the supercomputer, and ensures that the batch jobs enjoy clean I/O that is uninterrupted by the users' metadata-intensive activities (eg: SW builds).

- Pools can be built to service each of the different subnets that support different HPC clusters within a supercomputing center.

- VAST's native support for SMB also helps make it also very easy to build pools that natively connect the namespace to Windows and Mac users, all at sub-millisecond latency. Now, organizations no longer need separate enterprise NAS systems to service these user communities, simplifying data management and eliminating data migration jobs just by eliminating silos of data.

Thanks to the architecture's statelessness, VAST Servers can be easily provisioned and dynamically scaled, even programmatically by API, to adapt to the needs of an evolving application stack.

Price

In 2016, VAST looked out over the time horizon and realized there would be an opportunity to combine a collection of new technologies that ultimately did not become available until 2018 (specifically: NVMe-over-Fabrics, 3D XPoint, and QLC NAND) into a new storage architecture intended to bring the total cost of an all-flash distributed storage cluster architecture in line with what users had previously spent on hard drive and hybrid HDD+flash storage systems. VAST DASE architecture lays the foundation to implement a new collection of global algorithms that were previously impossible.

This new approach manages flash to get more efficiency out of silicon storage than what was ever possible. New innovations include:

- A new approach to managing low-cost, low-endurance flash that makes it possible to deploy commodity media in transactional environments, where other vendors can't use these low-cost devices because of their specific write pattern requirements. Moreover, since flash is

declining in cost at a rate much faster than HDD, this media will continue to drop in cost at a rapid rate as it approaches the raw cost parity of mechanical media.

- 10 years of flash longevity to help customers amortize storage for longer than ever.
- New data protection algorithms (Locally-Decodable Erasure Codes) which bring the 'tax' of protecting a system from drive failures from the roughly 30% average of other vendors to only 2.5%, all while increasing system resilience as compared to legacy approaches to data protection
- And finally, a new approach to data reduction that is neither compression nor deduplication—but rather a combination of the best of both—into a new capability called Similarity-Based Data Reduction (aka: Similarity). Similarity is designed to work globally across all of your files (as with block-based deduplication), but the pattern matching is byte-granular in ways that customers only ever enjoyed with local compression. Because the pattern matching works at 1/4000th the granularity of traditional deduplication algorithms, it is consequently much less sensitive to noise (i.e., differences) in data. This innovation makes it possible to dramatically shrink the footprint and expense of modern file and object data, where other storage systems show no gains, and has proven to even further reduce data that has already been compressed by applications.

  2.5:1 is becoming a commonly observed data reduction ratio in large-scale HPC centers, and this last approach to gaining global storage efficiency is signaling an extinction-level event for the hard drive in the exascale data center.

> The detail behind these inventions is too long to share in this whitepaper. To learn more, read the **Universal Storage** architecture whitepaper at vastdata.com/whitepaper

SUMMARY

While it was definitely possible for VAST Data to design and sell a custom file system client, the pains associated with complex HPC storage systems administration compel us to make the product as simple as possible.

By inventing this new DASE architecture concept, VAST has solved many of the problems that

have limited the utility of HPC in modern computing centers. By reinventing storage on new infrastructure technologies that weren't available until 2018, it's now possible to:

- Build scale-out NAS systems that deliver TB/s and millions of IOPS of performance
- Power demanding CPU and GPU client applications with RDMA throughput
- Achieve atomic consistency across a pool of parallel writers, with high performance
- Eliminate the cache miss penalty that tiered architectures inflict on AI algorithms
- Minimize the impact of noisy neighbors
- Achieve embarrassingly parallel scale, to exascale levels
- Pool infrastructure to build site-wide storage systems that ensure cluster QOS
- Break the price/performance tradeoff to make flash affordable for all of your data

VAST NFS is already powering some of the world's largest government labs, animation studios, seismic processing centers, bioinformatics pipelines, hedge fund research grids, AI efforts, and more. Customers are reporting more performance from VAST than what they've seen from their legacy parallel file systems and any other NAS options. VAST is simple, affordable, and applications thrive once they've made the move to flash.

To learn more about how **Universal Storage** can help simplify your exascale initiatives, reach out to us at hello@vastdata.com.

© 2020 VAST Data, Inc. All rights reserved. All trademarks belong to their respective owners.