The Need for and Emergence of Hyperscale NAS

February 2024

Written by: Marc Staimer, President Dragon Slayer Consulting, Contributor for the Cube Research

Premise

There has been a massive ground swell of actionable information that has caused a significant paradigm shift for IT organizations across all industries. The realization that information from analytical databases, a wide range of sensors dispersed everywhere, machine learning (ML), predictive and process AI, deep machine learning (DML), generative AI large language models (LLM) and many other sources can now be leveraged to derive additional value has taught IT professionals to keep all of their data and files alive and accessible. They have quickly learned to never throw away or delete any of it because IT organizations have discovered their data/files potentially hold substantial informational value. They simply need to unlock the value that is hidden within their digital assets.

But there are problems with keeping all data/files. Every file, object, or block has metadata associated with it. Both the data and the metadata must be made available with high-performance local to compute clusters for processing and stored, managed, migrated to where it's needed, value-tiered to the correct cost/performance storage, made accessible to users, analytics, AI engines, and application workloads, whether on-premises or in the cloud. None of that is cheap. More data and metadata mean more cost.

IDC analytics shows that ≥ 80% of all data is unstructured – primarily files. Unstructured data is also growing 3x faster than structured data, and this is the data that Generative AI models are hungry to get access to in order to derive competitive advantage. The storage of choice for most unstructured data/files is network attached storage, or NAS. Unfortunately, all current NAS systems and software have severe issues that cause them to come up short in this new high-performance 'analytical everywhere' paradigm.

Like every new technological paradigm, new supporting systems must innovate and advance to meet its needs. If all current NAS models are coming up short, then a new NAS model must be developed. That new NAS model is Hyperscale NAS.

Introduction

The reason NAS has become ubiquitous in most IT organizations is its extreme simplicity to implement, operate, manage, and maintain while also being intuitive to users. Yet, it has several idiosyncrasies and problems that range from annoying to downright frustrating. There have been various iterations and developments by storage vendors along the way that attempted to solve both the idiosyncrasies and the problems. Some have been somewhat successful, albeit not completely, and others not so much.

Offshoot file storage technologies such as object storage, Hadoop file system (HDFS) storage, and file sharing solutions were also developed along the way to try to address the traditional NAS limitations. Although they too introduced their own difficult problems.

NAS costs are always increasing. Storage is the only data center technology that's both used AND consumed. This means there is always a need for more storage, especially when data can't or won't be easily discarded, deleted, or destroyed. Current NAS technologies have not been keeping up with the performance demands of GPU compute farms or other HPC-like workloads. Nor are they keeping up with the mixed I/O workloads of different analytics applications, highly geographically distributed workloads, file sharing between parallel file systems and non-parallel file systems, or even maintaining high-performance while also managing/storing the massive file counts that continue to grow. NAS performance at scale is a primary problem.

These NAS shortcomings have generated an urgent need for new NAS architecture called Hyperscale NAS. What is Hyperscale NAS? It's NAS that can start small and scale massively in both performance and capacity, independently and linearly to the performance and scale requirements needed for today's data environments. Hyperscale NAS must overcome the limitations of all current NAS, object storage, and Hadoop storage models. From a performance perspective, it must be able to keep up with the largest GPU compute farms with the HPC-levels of performance needed to handle AI machine learning (ML), and generative AI deep machine learning (DML) workloads. Significantly, Hyperscale NAS is also based on open standards and affordability. Most importantly, Hyperscale NAS is what legacy NAS must become in order to solve these very difficult modern day unstructured data problems.

To get a clear appreciation of why there's a need for Hyperscale NAS, the problems it solves, and how it solves them, a short recap of the evolution of NAS will help set the context. We'll see how each iteration of NAS architecture solved some problems but not others, and how new problems were often created along the way. A brief NAS history provides that clarification.

NAS Evolution

NAS is a well-known file storage technology with a relatively long and storied history. It is a storage type specifically architected for files and unstructured data. NAS is commonly connected to multiple heterogeneous users and application workloads over a local area network.

In the Beginning

NAS was initially just a server with some drives associated with it. Originally called a file server, it morphed over time into being called NAS. To grow performance and/or capacity, the NAS server could be upgraded to a more powerful one and/or additional drives could also be added. But NAS was originally just a single server – a.k.a. storage controller or node. Performance and capacity scalability is restricted by CPU resources, amount of memory, number of drives, and drive capacities exploitable by a single NAS controller.

High Availability

As NAS grew ubiquitous and became increasingly important within IT organizations, high availability (HA) quickly emerged as a requirement. It was introduced with dual active-passive or active-active storage controllers. Each controller had its own storage and would take over the storage of the other controller if it should fail. However, capacity and performance scalability were still quite constrained. Typical NAS HA best practices put an upper limit of approximately 50% capacity for each NAS controller to ensure there was sufficient space available for failover.

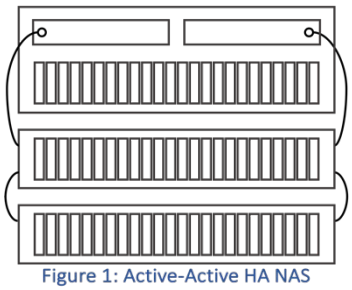


Figure 1: Active-Active HA NAS

The major downside to dual controller HA NAS is that when either more performance or capacity were required, another NAS system needed to be deployed. This resulted in storage sprawl, i.e., creating lots of NAS islands or silos. These NAS silos are highly inefficient NAS resource usage. As more NAS systems are added, the administration, management, operations, and costs caused by this architecture increase geometrically, not linearly. It also meant problematic tech refresh was always happening with one NAS or another, which further impacted utilization and increased costs. This led to the next step in NAS evolution – clustered or scale-out NAS.

Shared-Nothing Scale-out NAS

Scale-out NAS has mostly been a shared-nothing architecture that enables multiple NAS controllers to be clustered. Every controller is active and each of them can access all files in the NAS cluster. Adding performance is as simple as adding another node in the cluster.

The shared-nothing NAS architecture enables the distributed storage controller nodes to satisfy update requests by a single node in the shared-nothing scale-out NAS cluster. This architecture is supposed to eliminate contention among NAS storage nodes. Each node has its own dedicated memory and storage media.

© 2024 The Cube Research

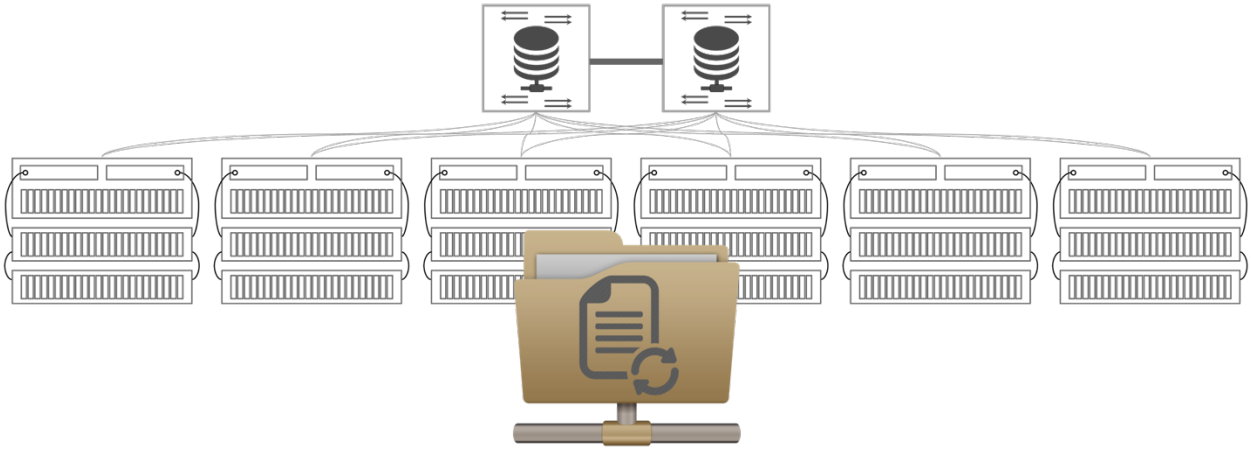


Figure 2: Shared-Nothing Scale-Out NAS

Adding capacity is accomplished with additional drives, larger drives, or more nodes in the cluster. Tech refresh is also much simpler – add an upgrade node, migrate the data, retire a node, rinse, and repeat. Its aim is to eliminate the local NAS silos so prevalent with dual-controller architectures.

Shared-nothing scale-out NAS changed the NAS market by enabling customers to reach unprecedented levels of performance, scale, and resilience. It introduced the ability to scale-out the file system across multiple storage nodes, what the storage vendors call a "Global Namespace[1]." This helped usher in the era of big data in the early aughts. Many of the well-known – NetApp, Dell PowerStore, Qumulo, Pure Storage Flash Blades – and not so well-known big data NAS and cloud NAS offerings are shared-nothing scale-out NAS. This architecture has also had a big influence on the development of object storage, cloud object storage, and even Hadoop storage.

Yet, despite all of that, shared-nothing clustered NAS has turned out to have some serious problems that storage vendors have not been able to overcome. Hard limits in performance and capacity scalability are the main ones. Each new NAS node in a cluster adds less incremental performance than the last one. As a result, at some point the next new clustered NAS node will actually decrease the overall clustered NAS performance. Just as important is the lack of scalability or parallel performance across the cluster as systems grow. Application workloads cannot take advantage of additional nodes in parallel, as they would in a typical parallel file system in HPC environments. Scale-out NAS solutions are limited to the performance of a single NIC. There is no multi-NIC bonding.

A major reason for these performance bottlenecks is related to inefficiencies in how NAS storage software sorts and indexes data/files. It's a metadata issue, i.e. the data about the data. Metadata provides essential information about when the data/file was created, where exactly it's stored, who or what created it, when it was last changed, and so much more.

Inefficient metadata processing shows up in the form of random read patterns, longer and inconsistent query performance, slowed writing and reading, longer application response times, I/O hangs, write amplifications, space amplification, and write stalls. These issues become increasingly problematic and frustrating as the NAS systems scale.

Additional decreases in scale-out NAS node performance are directly tied to data and metadata coordination between nodes. The cross talk between NAS cluster nodes severely inhibits performance scalability. Typical limitations where performance breaks down in scale-out NAS solutions come when cluster sizes get to several dozens of nodes. Internode cross talk grows exponentially with each new node that is added. Economics calls this the law of diminishing marginal returns. Storage administrators have a much more profane name for this issue. Crosstalk is exacerbated when there are mixed workloads or multi-tenancy. Every NAS node must be updated when any NAS node is updated. It's also why many customers isolate their mission-critical workloads with a separate scale-out NAS system, which once again recreates the problems of storage sprawl and silo management.

[1] What storage vendors mean by a "Global Namespace" is they can spread their file system across multiple storage nodes within **their** system. It does not mean they can spread their file system across different storage systems from the same or different vendors or spread it across separate geographic locations.

This internodal communications traffic becomes the fundamental performance bottleneck in all scale-out NAS systems. It can cause large shared-nothing scale-out clusters to massively underperform when processing large CPU-intensive workloads such as rendering. What's worse, this underperformance will generally not reveal itself until the job is run. Scale-out NAS vendors try to overcome this with high-performance internode networks such as InfiniBand or others buried within their systems. But such techniques add complexity while only masking the problem to a point, and don't fundamentally solve this performance limitation.

Storage tiering is yet another shared-nothing clustered NAS issue. Storage tiers or pools in shared-nothing cluster NAS systems tend not to be universally scalable because of the tight coupling between a storage controller and its storage media. Metadata, data management operations, backup/DR, and user access to file directories, have trouble going cross-tiers or pools. Data may become isolated within a storage tier or pool recreating the data silo problem clustered NAS platforms were supposed to eliminate. The shared-nothing scale-out NAS cluster puts the onus on the customer when it comes to sizing tiers in performance and capacity. In addition, this problem has spawned a growth industry of point solutions for supplemental data management, hierarchical storage management, and other techniques to overcome the fundamental limitations of the architecture and the resulting silos.

One other problem is data protection when drives or nodes fail. Most shared-nothing scale-out NAS platforms protect data in those situations via triple-copy or multi-copy mirroring between nodes. The problem with this approach is the cost of the added capacity it requires; each file copy consumes 100% more storage capacity. Protecting against two concurrent failures requires 3x the total capacity for the data. Protecting against three concurrent failures requires 4x the total capacity. This is an unsustainable cost.

That's why a few shared-nothing scale-out NAS systems have implemented erasure coding, which rebuilds the data itself and not the drives or the nodes. Erasure coding chunks the data and includes additional metadata on each chunk so files can be rapidly rebuilt if some of the chunks are inaccessible. The entire datagram remains accessible during the data rebuild, as opposed to the downtime required to do RAID rebuilds.

The use of erasure coding in scale-out NAS architectures has a noticeable downside; it puts an intense strain on the NAS controller processors, memory, and internode communications on reads, writes, and especially rebuilds. The result is a noticeable reduction in overall system performance. Application response times can increase to unacceptable levels. This is why many users elect the higher cost of multi-copy mirroring instead for application workloads that need faster performance.

Hadoop Distributed File System (HDFS) Storage

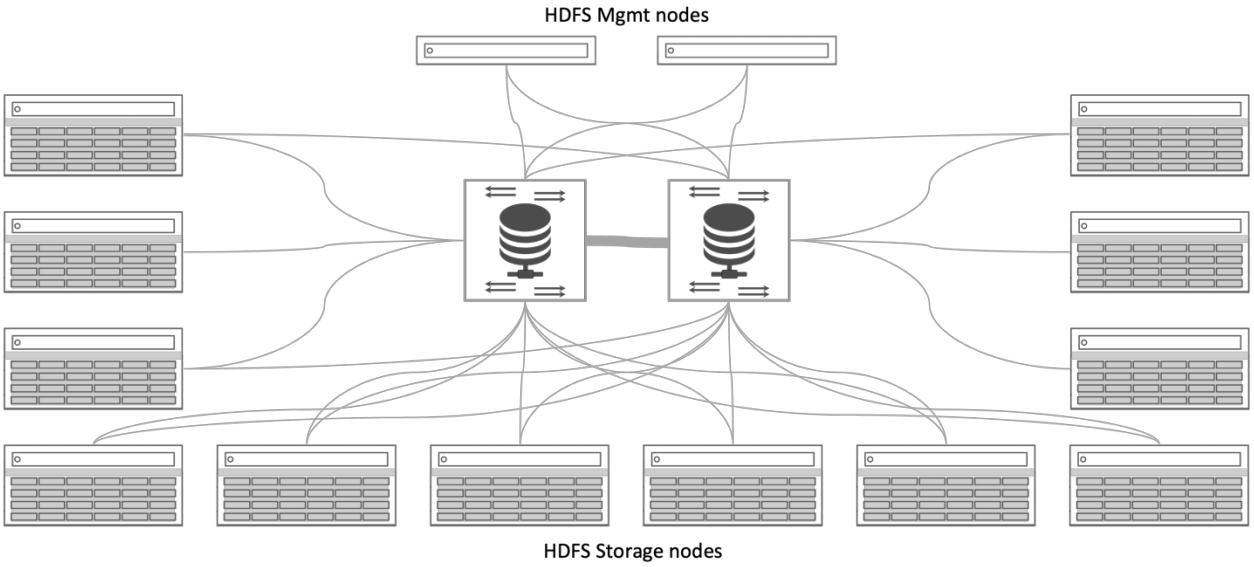


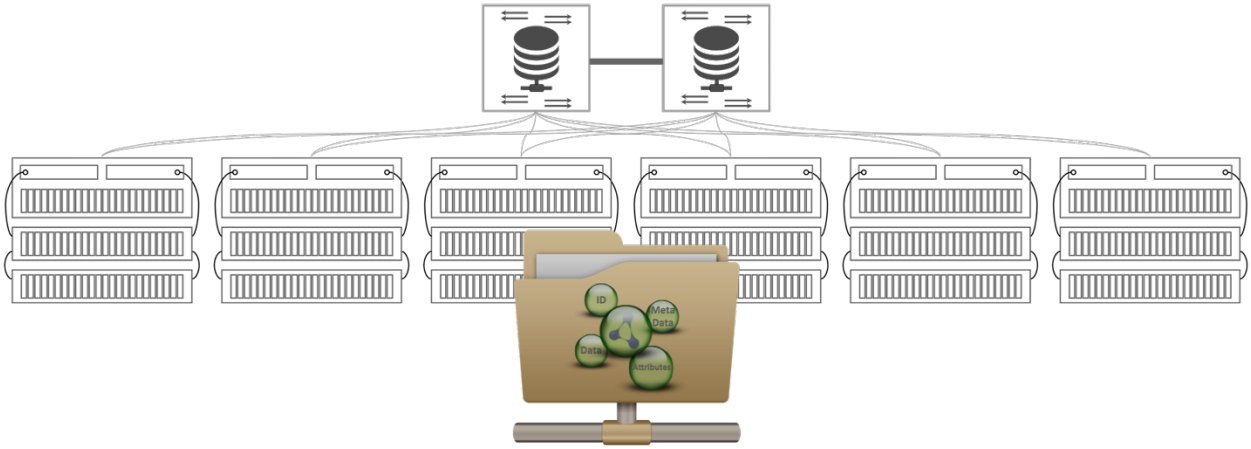
Figure 3: Hadoop Distributed File System

HDFS was developed with the goal of reducing the cost of storage for the large amounts of data so necessary for data lakes, analytics, AI, ML, DML, GenAI, and LLM. It too is a shared-nothing scale-out architecture. The concept is

to leverage cheap white box servers and high-density hard disk drives to provide the lowest possible cost performance for large volumes of data.

Sounded good in theory. In-practice is another story. Performance scalability in HDFS systems is subject to the same issues as shared-nothing scale-out NAS platforms. Protecting the data from multiple concurrent drive or node failures is back to multi-copy mirroring and the resulting high consumptive cost. HDFS lacks other basic data services and protection capabilities such as snapshots and geographically dispersed replications. As a result, HDFS has proven to be quite cumbersome and has been falling out of favor in recent years.

Object Storage



But Figure 4: Shared-Nothing Scale-Out Object

Object storage was developed to overcome the shared nothing scale-out NAS issues with significantly increased metadata performance, much reduced nodal crosstalk, plus increased resilience with standard erasure coding to protect against multiple concurrent drive or node failures. Most importantly, Object storage solutions offer the advantage of near infinite capacity scalability.

Unfortunately, the cost of these benefits is that object storage adds unacceptable latencies that make reads and writes significantly slower. The result is substantially increased application response times, lower productivity, a decline in work quality, morale, and turnover. These very high latencies are attributable to the additional metadata processing and erasure coding, which limits the use cases appropriate for such solutions.

Another object storage issue is the use the S3 protocol instead of NFS or SMB. Applications that have not been written specifically with S3 RESTful API will have to be refactored to access the object storage. The NFS/SMB gateway alternative comes with additional software cost, plus the considerable negative impact on performance, and the loss of common NAS services such as snapshots.

Object storage performance is mediocre at best even when used with solid state drives (SSD). Object storage is generally not suitable for the vast majority of mission-critical or high-performance applications. It is generally relegated to secondary or tertiary application workloads that do not require high IOPS or high throughput. Workloads such as archiving or backup.

Shared-Everything Scale-out NAS

An alternative architecture to shared-nothing scale-out is shared-everything scale-out. Share-everything scale-out satisfies requests with arbitrary combinations of nodes. However, this is also known to introduce contention at times when multiple nodes seek to update the same data at the same time.

Shared-everything scale-out NAS solutions attempt to limit this contention by separating the performance nodes from the capacity nodes. It essentially enables every NAS performance controller access to all capacity nodes within the system. It eliminates the NAS controller crosstalk at the performance layer, multi-workload crosstalk, multi-tenancy noisy neighbor crosstalk, and many of the shared-nothing scale-out NAS tiering issues. However, it has its own performance issues. Issues such as the need to add multiple network hops to data no longer in the performance nodes – data has to go from the capacity nodes to the performance nodes through a layer-2 RDMA fabric before it reaches the application or user. This adds a considerable amount of latency. These issues become clearly evident as

© 2024 The Cube Research

these systems scale, making shared-everything scale-out NAS inadequate to address the performance requirements at the scale required for AI and other HPC-class workloads. There are other restraints as well.

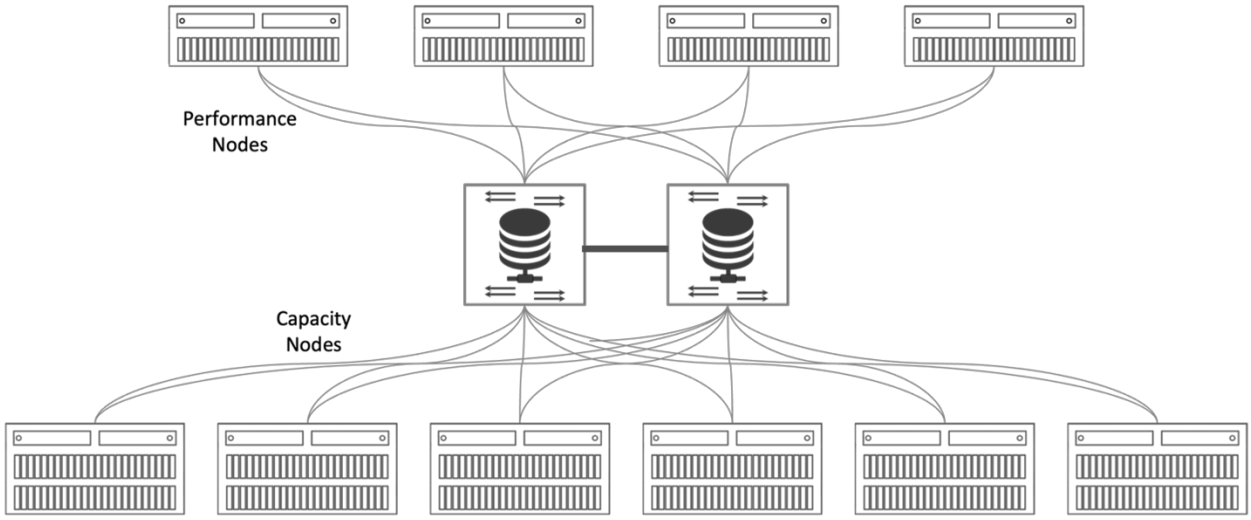


Figure 5: Shared-Everything Scale-out NAS

Restraints such as not being able to scale the global namespace to multiple geographic locations – shared-nothing scale-out NAS has the same problem. File access and storage in such systems are centralized. In order to make files accessible from multiple geographically disparate locations requires one of two methods. The user or application can either access it remotely, with the resulting latencies, or copy the files and replicate them over the WAN or MAN if appropriate. The latter then requires some form – typically manually – of synchronization.

Supplemental file sharing applications are also utilized to address this issue. As are distributed file systems. However, in both cases files must be duplicated/replicated and transported, resulting in the need to consume 100% more storage for each copy. This is an okay process for individual files. But it is extremely costly and time consuming for any significant quantity of unstructured file data, such as is common in media/entertainment, healthcare, pharmaceuticals, energy, CAD/CAM, life sciences, physics, data lake analytics, and AI. This also ignores the fact that when files are created at other locations, they are not accessible for users or application workloads at the central location unless copied and moved or made accessible to the centralized users and applications. Once again calling for duplicate capacities.

Shared-everything performance scaling has its limits too. Although it scales much higher than the shared-nothing architectures, each capacity NAS node still adds metadata processing to every performance NAS node. At some point the performance of the performance nodes declines as it becomes overloaded with the substantial metadata it has to process. This performance limitation becomes a serious obstacle to high performance workloads, such as the deep machine learning (DML) training so necessary for Gen-AI LLMs. Time is, after all, money.

Parallel File Systems

What about parallel file systems such as Lustre, Spectrum Scale, BeeGFS, Panasas, and others? These were developed for the extreme performance requirements of high-performance computing (HPC) environments. Each parallel file system node in the cluster has access to all the data in every node in the cluster. To get that performance, the parallel file system stripes the files into chunks or segments across multiple nodes. File reads and writes are performed in parallel across many HPC client nodes to deliver that aggregated increase in performance and metadata updates are handled out of band.

It's analogous to a shared-nothing scale-out clustered NAS except that it delivers extremely high throughput and IOPS performance at any scale that clustered NAS systems simply cannot achieve, due to the bottlenecks noted

© 2024 The Cube Research

previously. Parallel file systems are architected to support several billions of files. However, there are still highly problematic issues with these parallel file systems.

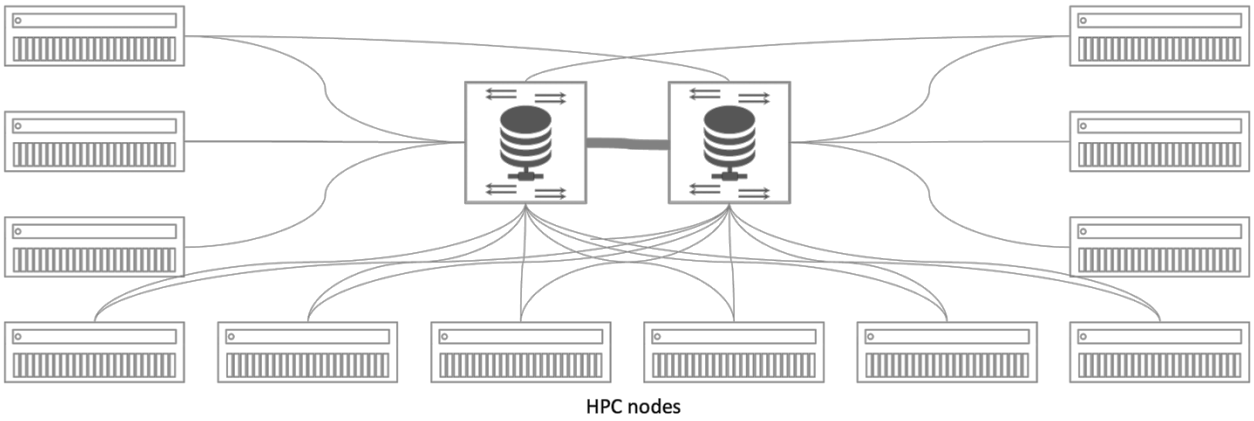


Figure 6: HPC Parallel File System

HPC parallel file systems are not immune to the laws of diminishing marginal returns. The performance addition with the 5,000th node is not going to be anywhere near as prominent as the addition with the 50th node. That helps explain why the workaround is to have separate clusters.

A bigger issue is that Parallel file systems are not standards-based. They are quite challenging to implement, operate, manage, and troubleshoot. For the longest time they have been the only option for traditional HPC use cases requiring specialized staff to maintain the proprietary client software, tune the system, and manage the specialized environment and networking. Open-source parallel Network File System (pNFSv4.2) with Flex Files is in the process of changing that paradigm. More on that a bit later.

In traditional HPC deployments, parallel file systems also require every client node to install proprietary software and/or hardware in order to access the other client nodes and its shared storage. Access is done over a high-speed layer-2 interconnect, commonly being InfiniBand, OmniPath, or Aries. IT administrators are generally very reluctant to open up their servers for proprietary anything. Traditional HPC or technical computing environments had no choice but to install proprietary clients to achieve the performance they needed with parallel file systems. But proprietary clients are generally not acceptable in Enterprise environments, where HPC-level performance is increasingly needed to satisfy AI/ML and similar workloads.

But perhaps the biggest limitation for the traditional parallel file systems is their inability to natively share any of the parallel file system files and data with applications or users that are not in the HPC cluster. Parallel file systems are simply not directly accessible via standard NAS protocols such as NFS, SMB, or even the object storage S3 standard, unless I/O funneled through some form of gateway. Gateways in general, are analogous to breathing through a straw. It's doable, but painfully slow. That means they throttle the parallel file system high-performance. Simply put, it's difficult to share data with non-HPC application workloads, commonly leading to the creation of more bottlenecks and data silos.

Cloud NAS and Cloud Object Storage

Some believe cloud NAS or cloud object storage solves these issues. Those who do are misinformed. Both have substantial limitations.

Cloud NAS is currently the most expensive storage available today, and it’s not even close. The price is currently averaging more than $300 per TB per month. It’s either a variant of HA NAS or shared-nothing scale-out NAS with their same issues and problems.

Whereas cloud object storage is more than an order of magnitude cheaper than cloud NAS, and delivers near infinite scalability, it uses a different interface to the application (S3 vs NFS), introduces high latencies and much slower I/O performance and throughput while concurrently pushing data management into the application. Cloud object storage is identical to on-premises object storage in that it’s a poor choice for mission-critical or any high-

© 2024 The Cube Research

performance application. Its mediocre performance, regardless of scale or use of SSDs, means it's relegated to tier-2 and tier-3 application workloads.

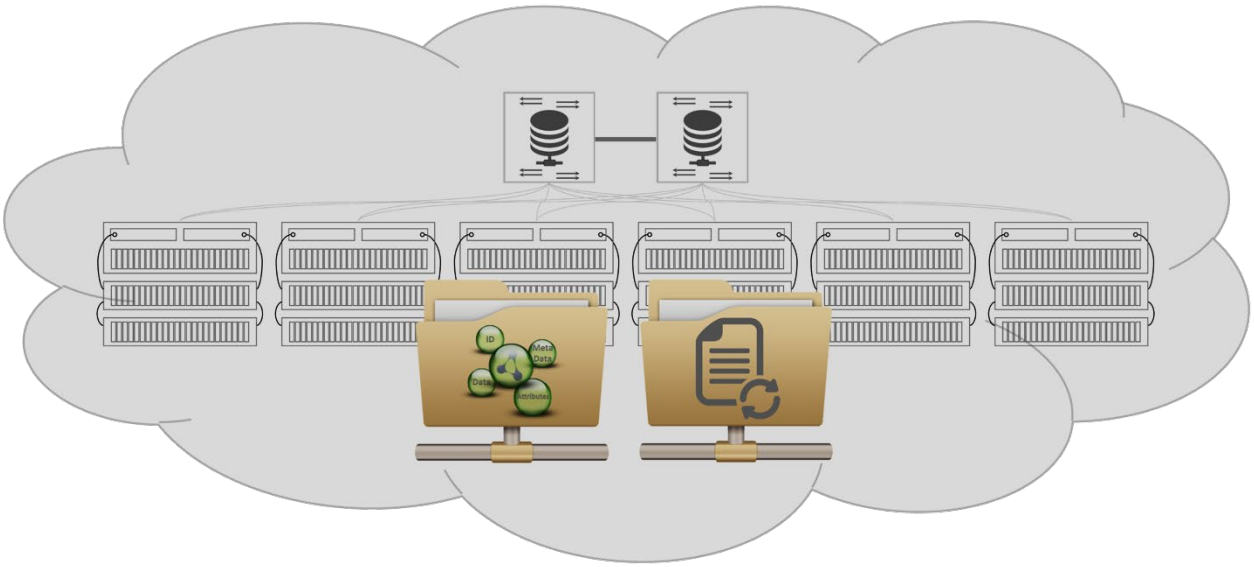


Figure 7: Cloud NAS or Object Storage

The actual value of cloud NAS and cloud object storage is that it's elastic on-demand. Customers only pay for what they use when they use it. However, since storage uniquely is both a used and consumed infrastructure resource, capacity requirements are always expanding and rarely decreasing. Neither cloud NAS nor cloud object storage solves the current data/file scale-out performance and management issues requiring a new model of NAS.

Overarching Problems with All Current NAS Types

All current NAS types and variations have positive attributes. They also introduce problems IT organizations must deal with and work around. Some of these issues are so problematic they're hindering the adoption of process or predictive AI, generative AI LLMs, autonomous analytics, and more.

Disruptive Data Assimilation of Current Data into a Big Data NAS Model



Shared-nothing scale-out NAS, shared-everything scale-out NAS, parallel file systems, cloud NAS, object storage, and cloud object storage vendors all suffer from a similar hubris. They all assume their storage systems will exclusively fit the customer's entire data storage needs. In other words, a customer's existing NAS and object systems must be replaced in order to take advantage of the advertised benefits of the new platform. It's the definitive forklift upgrade requiring every single byte of the customer's current data and metadata be moved or migrated to this new scalable storage, with the promise that it will be the last data migration they'll ever have to suffer through. Unless, of course, they decide somewhere down the road to change storage system vendors.

This is a non-trivial problem. And it's highly demoralizing to customers, who are reluctant to go all in with a single storage vendor ecosystem.

Data assimilation to bring those systems into a new environment typically means wholesale data/file migration. That's highly human labor-intensive, time-consuming, error-prone, and often disruptive to application and user workflows. Disruptions and downtime during such migrations must be scheduled, which slows the process down to a crawl and adds operational costs. And until they move their data to the new platform, customers can't take advantage of the global namespace or other storage services the new platform may offer. More importantly, when the migration is in process, or if they don't move all of their data/files, everything becomes more nutso (technical

© 2024 The Cube Research

term) and complicated. The users and application workloads will then have to navigate multiple NAS silos to access, store, and read their data.

Such data/file migration efforts ultimately become a massive, non-trivial, error-prone, multi-month or even years-long effort that are frequently disruptive with scheduled outages. The end result is ultimately a very costly effort with high risk. And rarely do all of the data/files make it into the new one-size-fits-all system, which is the origin of the term data gravity.

One of the great promises that all types of scale-out NAS, object storage, and Hadoop file system solutions touted was the ability to consolidate storage silos. Except each consolidation comes right back to this data/file migration problem. The cure is frequently worse than the silo disease.

For too many large organizations that's simply a no-op. Even the IT organizations that bite the bullet and do these extensive data migrations find it takes much more effort, far more time, with more problems, and greater cost than anyone anticipated. And when they're finally done – a big if – they're now locked into that vendor's NAS or object storage ecosystem, with no way out except to go through the same painful experience all over again.

Scaling Total NAS Performance Linearly

As previously noted, every NAS type has a problem with performance as they scale. And the threshold where such problems arise is usually much smaller in size than the vendors assert. It's analogous to EPA mileage statistics for internal combustion engines or the range for fully charged battery electric vehicles. The real-world numbers are always noticeably less than what is printed on the marketing brochure.

This performance issue becomes increasingly impossible to ignore as data that was previously primarily just stored but now is being consumed by large scale compute environments and the performance issues of the storage are keeping expensive processors idle. Object storage does not have the same degradation issue, but its performance is so latency heavy that application response times becomes unacceptable for mission-critical or even business-critical applications.

IBM research into application response times[2] showed a marked decrease in user productivity as the response times began to exceed 400ms. Productivity dropped even more rapidly when those response times exceeded 1s, and completely fell off a cliff when they exceeded 2s. Object storage latencies equate into application response times that commonly measure in seconds.

Geographically Distributed Global File Access

Providing geographically distributed global file access with any of the current NAS models or object storage requires substantial gymnastics and duplicated storage to accommodate the inevitable copy sprawl. Gymnastics such as replicating the files, metadata, and permissions while keeping the files in sync add complexity to IT staff and friction to users and application workloads. Each of those file copies has its own metadata, which effectively means they are forked copies. Changes in one copy are independent of others, requiring those changes to be synchronized. Synchronization can be automated with file sharing applications or other point solutions but are generally manually reconciled by IT staff or users. These gymnastics grow exponentially as the number of sites, files, and users increase, becoming unsustainable.

Excessive Total Cost of Ownership (TCO)

As discussed, the data/file migration needed to add another NAS into a customer environment is by itself excessively costly. Both the old systems and new must be paid for and used concurrently. This equates to twice the cost just for the systems. And that does not include the loss of productivity or other operational costs of the migration.

The workaround to overcome the limited performance scalability of these systems is often to implement additional separate NAS platforms. More systems are more costly in licensing, cabling, switches, allocated data center overhead – or in the cloud more application license instances, database sharding, etc. – operations, management, troubleshooting, personnel, training, human errors.

The gymnastics of geographically distributed global file access also adds to increasingly pointless costs. Costs that include substantial higher capacity utilization from copy sprawl – that grows another 100% with each location,

[2] The Economic Value of Rapid Response Times

© 2024 The Cube Research

increasing network utilization, and the need for more manual labor-intensive human intervention, resulting in both disruption and people costs.

All of these issues make it clear that a new innovative NAS model is required. The name of that NAS model is called Hyperscale NAS.

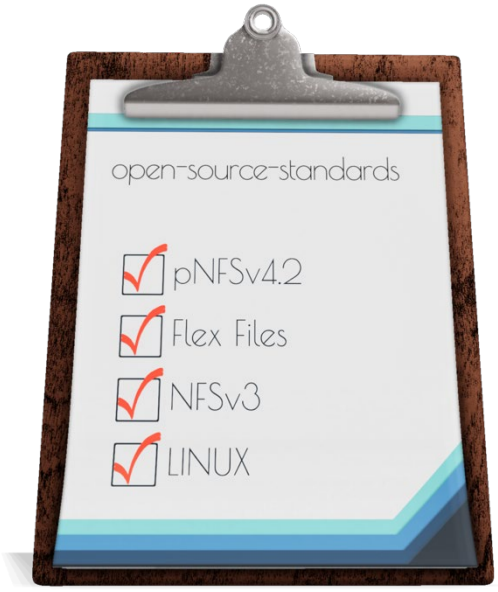
Hyperscale NAS – The New High-Performance NAS Standard

Hyperscale NAS is not a product, it's a new NAS architecture model based on open standards available in all standard Linux distributions used in the industry today. Within the Linux kernel – on servers and hypervisors in practically every data center – are fundamental elements to enable standards-based HPC-level parallel file system performance on commodity hardware and even existing legacy storage types from any data center or cloud vendor. Specifically parallel NFS v4.2, with the pNFSv4.2 client[3], Flex Files[4], and NFSv3.

Hyperscale NAS is specifically aimed at solving all the aforementioned problems for high-performance workloads when using scale-out NAS, object storage, and Hadoop. In addition, to ease adoption Hyperscale NAS needs to make the data assimilation problem non-disruptive as a background operation, so it is transparent to user and application workloads. The performance must scale linearly to enable HPC-level high-performance workloads regardless of geographic locality; and do all of this while significantly reducing the total cost of ownership (TCO). Ultimately, all this needs to be delivered in vendor neutral solutions, compatible with any storage type from any vendor, to eliminate the silo problem entirely, even with legacy infrastructure.

In short, Hyperscale NAS needs to deliver HPC-levels of parallel file system performance, but with the ease of use, standards-based access, and RAS (reliability, availability, and serviceability) features that are non-negotiable entry criteria for storage systems in Enterprise IT environments.

This may sound like a pipe dream to many IT storage professionals. It's not. Hammerspace is the first vendor to deliver on the Hyperscale NAS promise.



[3] pNFSv4.2 – a.k.a. NFSv4.2 – All standard distributions of Linux from RedHat RHEL6.2, Suse SLES 11.3, Oracle Linux 6, and Ubuntu 14.10 and later all include the pNFSv4.2 client.

[4] Flex Files – a.k.a. Flexible Files – is the layout for pNFSv4.2 enabling aggregation of standalone NFSv3 and pNFSv4.2 servers into a scale-out name space. The Flex Files feature is part of the NFSv4.2 standard as described in the RFC 7862 specification.

© 2024 The Cube Research

Hammerspace Hyperscale NAS

Hammerspace Hyperscale NAS is the first solution to meet or exceed the new Hyperscale NAS standard. It each of the problems Hyperscale NAS is supposed to solve.

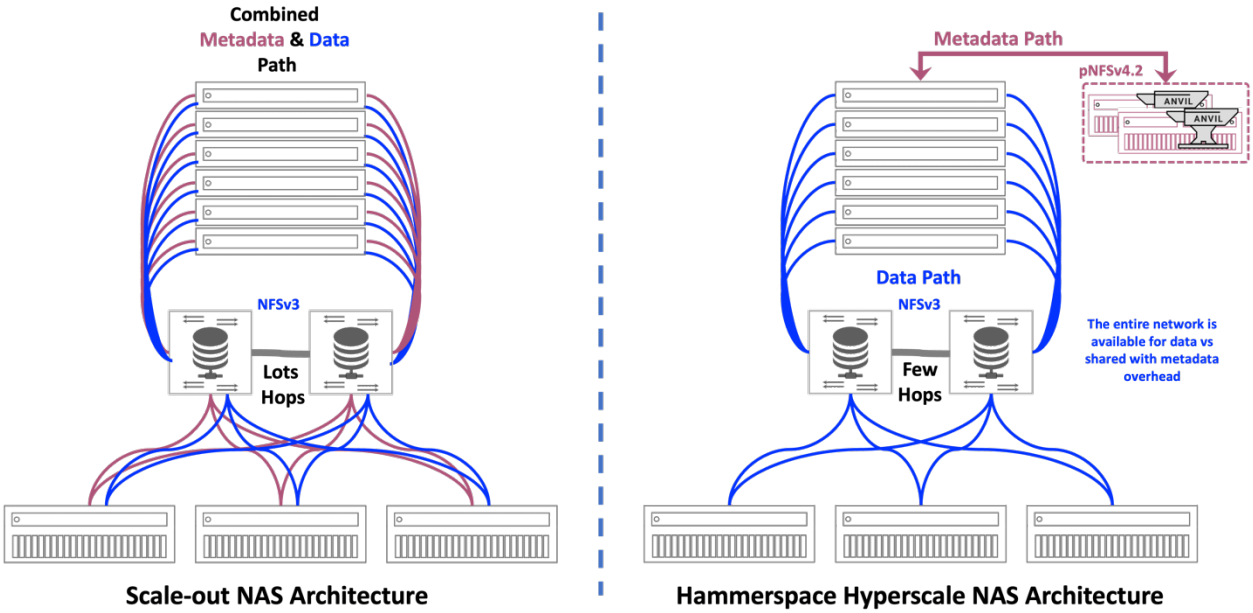


Figure 8: Scale-out NAS Architecture vs. Hammerspace Hyperscale NAS Architectures

Standards-Based

Just as the Hyperscale NAS requires open-source standards, so is Hammerspace Hyperscale NAS. It relies on those same standards of Linux, pNFSv4.2 with Flex Files, NFSv3 and other components within the Linux kernel. No custom/proprietary software or hardware is required for clients, application servers, hypervisors, or on the target storage. Hammerspace Hyperscale NAS is able to achieve exceptionally high performance primarily because it enables direct connection between application workloads and the storage, bypassing the storage controllers/nodes, multiple network hops, and the extensive latency of those hops.

Transparent Data Assimilation

Hammerspace Hyperscale NAS makes the highly exasperating and difficult data assimilation problem simply disappear. It does so by assimilating existing file system metadata while leaving data in place on existing storage. Even in very large environments this metadata assimilation is extremely fast, being performed mostly in the background, transparent to users and applications.

This is managed by the Hammerspace Metadata Services Nodes, called “Anvils”. By separating the metadata control plane from the data path, Hammerspace presents a high-performance Parallel Global File System to users and applications that not only can span multiple data silos, but also multiple sites and clouds. It means all metadata collection, harvesting, data orchestration, and management now can become automated background operations, transparent to users. This separation enables Hammerspace to capture all of the metadata from every current NAS, file server, and object store without operational disruption. This deliberate separation changes everything.

The separation of metadata from the data path provides enormous flexibility on where the data/files are physically located and automates data placement operations in the background. This architecture IT organizations can change the physical storage platform at will without disruption or outages because users are all accessing the same file system that bridges all underlying storage resources. It provides both users and application workloads access – reads, writes, modifications, deletions – to their data/files regardless of where their physical storage is located. Hyperscale

NAS can leverage Hammerspace's automated data orchestration capabilities to enable IT teams to move or change the physical storage whenever they need to transparently, without disruptions.

In other words, Hammerspace Hyperscale NAS makes data migration a non-event or a bad memory of the past. Customers can choose to move their data/files to newer, faster, or less expensive storage behind the Hammerspace Parallel Global File System. Or they can keep some, most, or all of it on current physical storage and still get the benefits of Hammerspace Hyperscale NAS to even improve the performance of their existing scale-out NAS platforms, as we'll highlight below.

Incomparable Linear Scalable High-Performance

The ability to achieve HPC-class performance with Hyperscale NAS is rooted in the integration within Hammerspace of parallel NFS v4.2 with Flex Files. This separates the metadata control plane from the data path and brings parallel file system performance to conventional enterprise storage environments. Hammerspace's Hyperscale NAS enables the extreme parallelization of I/O between clients and storage nodes similar to what is possible with Lustre and other parallel file systems. The key difference being Hammerspace Hyperscale NAS is based upon open standard hardware, networking, interfaces to applications, and eliminates the need to migrate data or install proprietary client software or hardware on customer systems.

An excellent extreme scale example is a very large AI training and inferencing pipeline in production at one of the world's largest AI LLM training environments. The customer had over 1,000 storage servers with 42PB of NVMe storage. They required the performance and resilience to feed high-performance workloads at over 12TB/s into a cluster of 4,000 servers with 32,000 GPUs. Not a trivial task.

All of the various NAS types noted above were rejected because of the limitations with those systems noted above. HPC parallel file systems were also considered but rejected. Too complex to administer, lacking the RAS requirements needed by the customer, but most importantly they required proprietary client software to be installed as well as need specialized networking.

Hammerspace Hyperscale NAS proved to be the answer they needed.

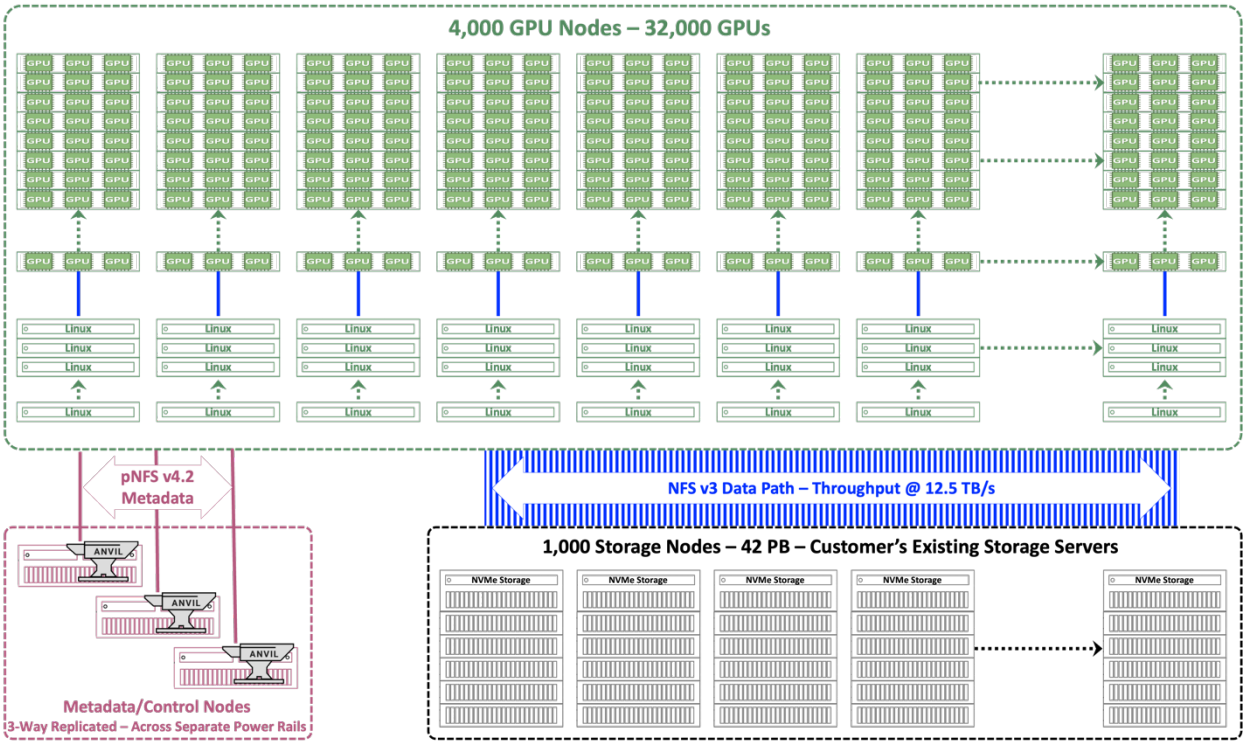


Figure 9: A Hyperscale NAS implementation for massive-scale AI workloads. Leveraging pNFSv4.2 with Flex Files within the Linux kernel on their existing systems, Hammerspace implemented the solution without data migration, and without the need to install proprietary client software.

It's unique metadata collection, orchestration, and management separation from the data path has a profound impact on performance. Metadata processing that is a performance bottleneck in all previous NAS models no longer

© 2024 The Cube Research

negatively reduces NAS scalability in IOPS, latency, throughput, or application response times with Hyperscale NAS. In fact, there is a considerable increase in performance that's comparable to adding nitrous oxide[5] to a racecar engine.

Another example shows how Hammerspace implemented a Hyperscale NAS architecture to supercharge a customer's existing scale-out NAS platform. Consider the experience of a large media and entertainment CGI visual effects studio providing high quality, visually stunning imagery for films and television. Before they implemented Hammerspace Hyperscale NAS, they were unable to get anywhere near adequate performance from their 32-node shared-nothing scale-out NAS cluster when attempting to send VFX projects to a 300-node render farm. Jobs would routinely time-out and stall, as the shared-nothing scale-out NAS cluster struggled to keep up with the data flow. Render jobs would consistently fail because the shared-nothing scale-out NAS cluster simply could not handle the I/O requirements fast enough. This resulted in severe underutilization of the render farm, but more importantly it limited their production output.

The company had to reduce the number of render nodes they could allocate to a job. Productivity plummeted making it incredibly difficult for them to keep up with their demanding production schedule.

Hammerspace Hyperscale NAS immediately flipped the script and eliminated the shared-nothing scale-out NAS performance bottleneck, but without altering the existing storage or needing to migrate data.

Data writes were now fast enough for the 32-node shared-nothing scale-out NAS cluster to completely accommodate the full output of the entire 300-node render farm without timeouts. The 32-node shared-nothing scale-out NAS cluster performance increased so much that the company is now doubling the size of their render farm to 600 nodes without changing or adding to their shared-nothing scale-out NAS cluster. And they are still nowhere near the improved performance limits achievable with their 32-node shared-nothing scale-out NAS cluster.

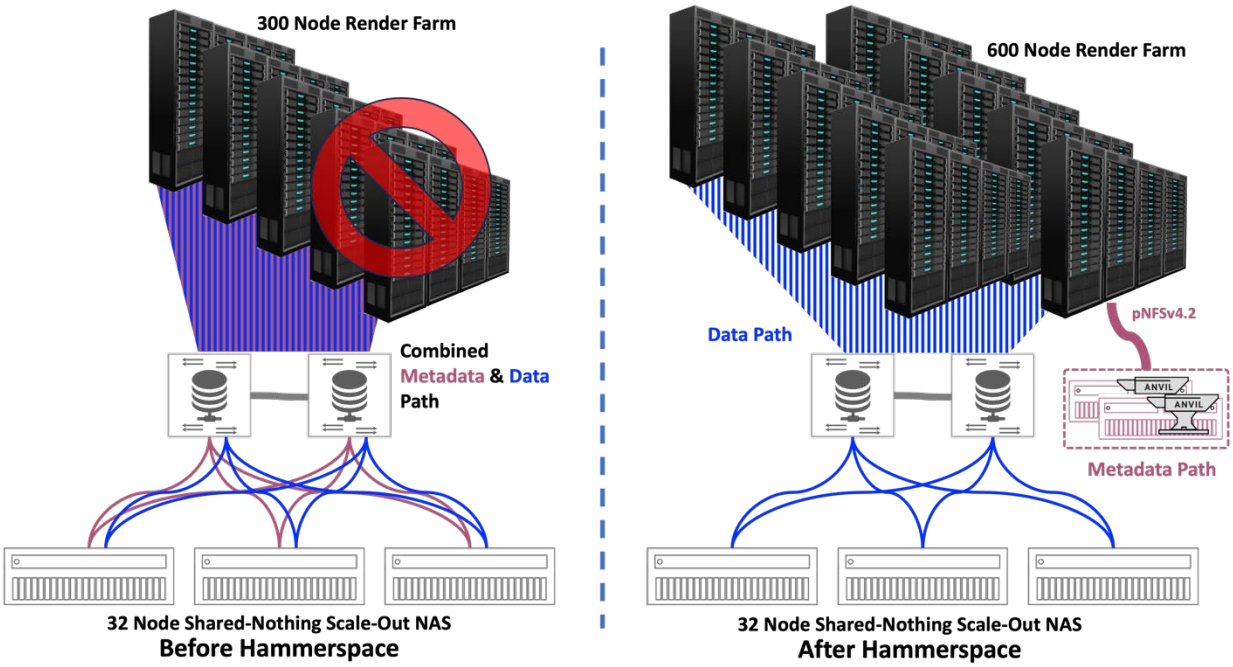


Figure 10: Before Hammerspace and After Hammerspace Configurations with VFX Organization

Since Hammerspace Hyperscale NAS leverages pNFSv4.2 with Flex Files that is already in the Linux kernel, the customer was able to implement it on their existing servers and storage, without modification to either. Based upon

[5] Per Wikipedia, nitrous oxide (often called "nitrous") allows the racecar's engine to burn more fuel by providing more oxygen during combustion. The increase in oxygen allows an increase in the injection of fuel, allowing the engine to produce substantially more engine power.

© 2024 The Cube Research

the metadata that Hammerspace assimilated from data in place on the scale-out NAS, it could now parallelize I/O dynamically between the clients and the storage nodes. This enabled direct I/O between client and storage node, eliminating the additional network hops and internode bottleneck of the scale-out NAS.

The render nodes now write directly from the pNFSv4.2 client to the exact file locations in parallel across the 32-node shared-nothing scale-out NAS cluster nodes. On the storage side, the data path uses standard NFSv3 to connect to the storage nodes in parallel, bypassing storage caching. This parallelization effectively created a virtual cache 'n' times – 'n' is the number of storage nodes in the cluster – increasing write speeds so substantially, they had to be seen to be believed.

Note that these performance improvements didn't involve altering or increasing the 32-node shared-nothing scale-out NAS cluster. Nor did it involve data migration of data from the 32-node shared-nothing scale-out NAS cluster, or any other lengthy integration process. And since pNFSv4.2 with Flex Files layouts is already in the Linux kernel on their existing servers, they didn't need to install any client software on the server side, or agents on the storage. It just worked.

They can always decide in the future to eliminate their current 32-node shared-nothing scale-out NAS cluster and just utilize the Hammerspace Hyperscale NAS with other backend storage and get even better performance. Keep in mind that achieving those dramatic performance improvements required only a small Hammerspace deployment. And in the process, it also enabled the customer to take advantage of the Data Orchestration and multi-site, multi-cloud capabilities of the overall Hammerspace Global Data Environment software stack.

Obviously, the parallelism of the data path and separation of metadata from the data path for every client is a major contributor to Hammerspace Hyperscale NAS incredible high-performance. However, it's not the only one. So are the massive improvements in pNFSv4.2 protocol efficiency.

The pNFSv4.2 efficiency improvements over NFSv3 or NFSv2 are massive. A file create in Hammerspace Hyperscale NAS has a single roundtrip to the metadata server and one to the data storage node. That same file create using NFS v3 as well as earlier parallel NFS versions required multiple back and forth round trips to the data storage node. That's a considerable amount of chatter and hops that dragged performance down, especially as the NAS systems scale.

Hammerspace Hyperscale NAS doesn't add any intermediaries in front of either the data or metadata paths. This is in contrast to all other NAS types that run both the metadata and data directly through the same intermediary – storage nodes/controllers. Having no intermediaries eliminates storage node contention, enabling zero cache overlap when different files go to different storage nodes. Separate data and metadata paths, less chatter, much reduced hops, and no storage node contention enormously accelerates linearly scalable performance.

Geographically Distributed Global File Access With Local Performance

Hammerspace Hyperscale NAS also leverages Hammerspace's ability to span multiple on-premises and/or cloud locations with its high-performance Parallel Global File System. Users and applications around the world can access files globally via standard file protocols to storage resources that may be anywhere as if they're local. Such local high-performance is delivered non-disruptively when needed via Hammerspace's automated data orchestration capability.

All users everywhere are accessing files via same global file system. Not file copies, but the same files via this global metadata control layer. Data/files do not need to be mirrored or replicated, even though files are protected with multiple layers of redundancy at both the metadata and data layers. There is no data/file duplication requirement. There is no file synchronization requirement. And most importantly, there is no labor-intensive, error-prone, or even visible data/file migration. This by itself is a game changer.

Exceptional Data Protection and Transparent Archiving

Hammerspace Hyperscale NAS provides outstanding data protection. Leveraging the RAS capabilities of Hammerspace's Global Data Environment, it comes with a series of data services that can be applied to files across all underlying storage resources globally. These include immutable global snapshots across all storage types, undelete, WORM (Write Once Read Many), file versioning, and more.

© 2024 The Cube Research

A key Hammerspace innovation is its ability to centrally manage global snapshots across incompatible storage resources, e.g., other vendor NAS, object stores, or block storage. Snapshots are taken at the share or file level, with highly granular controls. It can offload the snapshots to tier-2 or tier-3 storage locations such as cloud storage, cold/cool storage, or tape. This frees up tier-1 NAS for performance while delivering data protection IT organizations can count on.

Hammerspace Hyperscale NAS supports up to 4,096 snapshots per share, file-level data recovery, snapshots stored on any storage including object storage and/or the cloud. Snapshots are user-accessible through snapshot directory and Windows VSS (Volume Shadow Copy Service). Snapshots are extremely flexible with the capability of being recovered in-place or at a different location.

The automated data orchestration that is such an integral part of Hammerspace can automatically offload cold data from primary storage nodes in background and transparently to cloud cold storage or even tape. This frees up additional headroom in the primary storage to accommodate the increased data output application workloads and users.

Much Reduced TCO and TCO/Performance

Data assimilation without the complexity, personnel, time sink, troubleshooting, and cost enables IT administrators to think differently about their data environments. Much higher data performance without having the replacement cost of all or even any of the current storage. Even when replacing all or part of the current storage, significantly lower cost storage such as JBOFs or JBODs, can be used behind Hammerspace Hyperscale NAS. Lower cost with the same high performance and no storage services loss. All of this adds up to a much lower TCO and TCO/performance.

The unmatched performance acceleration even on legacy storage that Hammerspace Hyperscale NAS enables also translates into lower application response times. Lower application response times turn into much higher productivity, faster time-to-insights, faster time-to-action, faster time-to-market, and faster time-to-unique-revenues/profits.

Summary/Conclusion

The IT world is reeling from the extremely rapid growth of information creation. This has led to an explosion of data and files found to have real heretofore undiscovered informational value by analytical databases, AI, and cloud services, – OLAP (data warehouse), document (JSON), object (XML), spatial/graphic, time series, block chain, vector search, autonomous databases, data lakes, predictive AI, ML, DML, generative AI, and LLMs. These information discovering tools have completely changed the IT landscape. Data lifecycles no longer have a defined endpoint. Data/files are trending towards immortality. This has put enormous pressure on the data/file storage because all of these rapidly expanding data sets have to be stored somewhere. Since most of that data is unstructured, NAS is the primary storage of choice for most IT organizations. It's easy, intuitive, and flexible.

But none of the traditional NAS models work effectively enough for this new actionable information creation paradigm, especially at the enormous performance and capacity scales these workflows demand. All have flaws in meeting this new paradigm, whether it’s slow performance, not enough scalable performance, disruptive data migration, forklift replacements of current NAS systems, inability to extend their global namespace in geographically dispersed regions, poor tiering, or inadequate data protection, every NAS model comes up short.

That explosive information growth has generated requirements for a new type of NAS called Hyperscale NAS. Hammerspace is leading the way by being the first to deliver Hyperscale NAS that consistently meets or exceeds all of these new requirements. No other NAS or any storage system is as good of a fit for the new actionable information-creation paradigm.

Hammerspace Hyperscale NAS eliminates the headaches of data migration with transparent data and metadata assimilation; delivers the HPC-class high-performance demanded for modern data pipelines, but with the ease of use of standard NAS; produces more performance out of less hardware; scales that performance linearly; conveys the necessary exceptional data protection and RAS features needed for Enterprise IT environments; transparently archives cool and cold data; and does all of this with a software-defined solution at an affordable and much reduced TCO.

© 2024 The Cube Research

To find out more on how Hammerspace Hyperscale NAS can smooth the path for valuable and actionable information creation in your organization, go to:

Hammerspace.com

© 2024 The Cube Research