Hammerspace Hyperscale NAS Technology Overview

Silverton Consulting, Inc. StorInt™ Briefing



Introduction

Enterprise IT is undergoing one of the most significant workload changes since the dawn of the internet. That change is due to AI. AI and the cloud arguably have been the two biggest developments for IT in the 21st century. While the cloud has changed where IT work runs, AI is changing *how* IT works.

At first, AI was used in specific domains, such as high-performance computing (HPC). AI has also achieved some success in image/object recognition, language translation and speech transcription.

Since the introduction of generative AI (GenAI), however, AI's applicability for enterprise work has increased a thousand times over. Never has the enterprise seen a technology with so many applications to so great a number of IT activities and use cases.

Today, IT's main challenge in supporting AI is supplying its infrastructure requirements. Obtaining the thousands of expensive GPUs that IT needs to train, tune, and use GenAI can be a major obstacle. Once acquired, keeping these GPUs fully utilized can be a significant challenge, requiring high performance file access and data services.

Data is the other key to AI generally and to GenAI in particular. Training and fine-tuning GenAI models consumes lots of data and requires quick, effective access to that data to keep GPUs working. AI data performance requirements far exceed what's available with today's enterprise storage systems.

One company has created a new data access paradigm to support AI and other data-intensive applications and keep thousands of GPUs busy. That company is **Hammerspace™**.

The paper will highlight some of the implications of the AI workload transition. The paper will also discuss how Hammerspace's **Hyperscale NAS solution** can enable an enterprise's current data storage environment to achieve the performance demanded by extremely data-intensive workloads.

Workload transitions drive data

Over the past decade, business units within the enterprise have been running large-scale scientific workloads, such as big data analytics, high frequency trading back testing, and genomics. HPC has been running similar scientific applications for a long time.

Parallel File Systems (PFS), such as Lustre, have long supplied the data capacity and data performance that HPC data-intensive scientific applications and large-scale supercomputers need. These PFS achieve massive performance scale and efficient use of infrastructure resources by removing metadata from the data path, thereby allowing direct connection to data from compute. PFS also commonly run on specialized networking such as InfiniBand to further accelerate data access.

With PFS, HPC could increase from 100Gbps data bandwidth to 10Tbps or beyond simply by adding storage nodes. (More nodes also increase storage capacity from TB to EB.) Attaining these levels of bandwidth and capacity may require moving from hundreds of storage nodes to thousands. PFS could readily handle all this hardware scaling, and more.

© 2024 SILVERTON CONSULTING, INC. ALL RIGHTS RESERVED

However, adding nodes wasn't enough, as data path flexibility was needed, often from multiple client servers, to support node scaling without impeding data access. PFS use intelligent client software which allows metadata to be taken out of band, but also supplies simultaneous data access to multiple clients. This direct data access from the host to storage was the only way to scale bandwidth and capacity and take advantage of the massive supercomputer powering the workload.

Enterprise IT, with its strict security and software standards, has never seen this level of performance capabilities outside of cloud hyperscale environments – that is, not until Hammerspace Hyperscale NAS.

Data access evolution

At first, data storage was locally attached to servers, called direct-attached storage, or DAS. Systems could easily map one or more of these devices to use as basic block storage or a file system. These systems achieved added efficiency with the introduction of NVMe, which, when used as DAS, removed the RAID controller, significantly increased IOPS rates and reduced latency for data access.

As block storage needs increased over time, IT moved from DAS to shared, network-attached, block storage appliances with new network block storage protocols. In this way, expensive block storage could be amortized across many servers.

File storage underwent a similar transition from being hosted on DAS to being hosted on NAS systems with the emergence of network-oriented, open systems NFS (Network File System) and Windows SMB (Server Message Block) file access protocols.

As with block storage, NAS was created to apportion expensive storage capacity across multiple hosts. Because NAS could improve file performance by adding cache and storage hardware to an appliance, it came to be called **scale-up NAS**. Scale-up NAS performance was limited by the hardware that could fit in an appliance. When a NAS system ran out of hardware space, IT had to purchase more NAS systems or upgrade to faster NAS.

Another alternative, shared file storage architecture emerged, that used proprietary software and a dedicated high speed internal network running on commodity server nodes, called **scale-out NAS**. Scale-out NAS performance could be increased by adding more nodes. NVMe over Fabrics (NVMEoF) was added to some scale-out NAS solutions for increased performance. But front end file servers and their limitations remained. The larger a scale-out cluster grew, the less additional nodes improved performance.

Finally, while HPC had already solved the linear performance scalability problem with PFS, enterprise IT was reluctant to adopt PFS solutions because they all required proprietary client and storage server software and didn't include enterprise services needed to meet data security, compliance, and governance requirements.

PFS primer

HPC scientific applications and supercomputing environments need rapid access to lots of large files with extreme throughput. Hundreds or more servers may require immediate access to the same data. With PFS supplying parallel direct access to data residing on its storage servers, no one host (or multiple

© 2024 SILVERTON CONSULTING, INC. ALL RIGHTS RESERVED

hosts) needed to wait for data access. And nearly 100% of the performance capabilities of the compute, network, and storage servers were delivered to applications as metadata activity was offloaded.

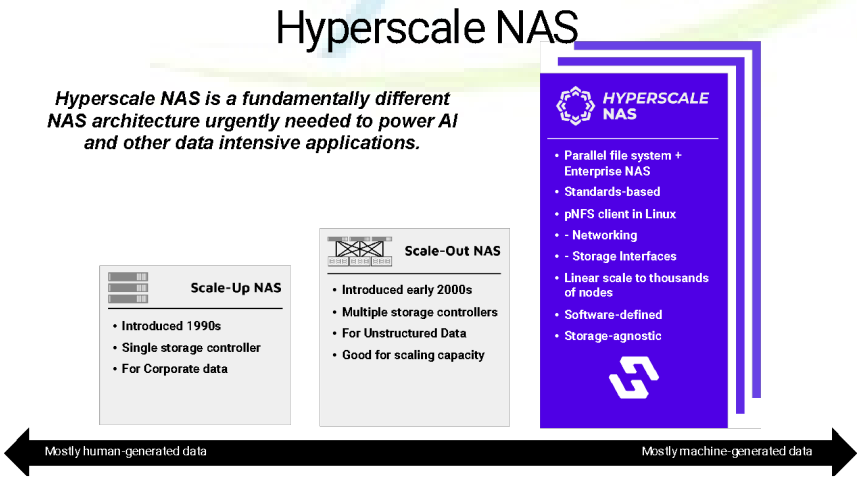
With parallel direct access to data, PFS could take advantage of all the hardware HPC environments could deploy. With storage servers capable of 1Gbps, two servers could sustain 2Gbps of data access bandwidth, 20 could sustain 20Gbps and 200 could sustain 200Gbps. With PFS, budget was the only limitation on data bandwidth and capacity that could be delivered to the most expensive part of the environment, the compute.

HPC introduced other innovations in networking, cluster configurations, server memory and floating-point compute performance. However, PFS remained key to HPC scientific application performance but had limited Enterprise adoption due to not using standard host software and access protocols, or providing ease of use and the RAS (reliability, availability, and serviceability) features they require.

Some of the new technologies PFS added to file storage include the following:

- **Separate metadata servers** – In scale-up and scale-out NAS, metadata was supplied by the same system and over the same path as file data. PFS offloaded metadata services to separate servers, so metadata and file data no longer had to compete for the same data path.
- **Direct access to data** – In enterprise NAS I/O, data often had to pass through many layers of a system's architecture or hop from node to node before being received at the host, which added time and overhead. PFS supplied a direct path for data in storage servers to the compute environment, avoiding unnecessary data transfers and speeding up data access in the process.
- **Extreme scalability** – With direct access to data and separate metadata servers, adding nodes to a PFS cluster multiplied data access bandwidth, significantly improving file access performance.

Over the last decade, the NFS standards body began adding PFS capabilities to NFS. Over time, the NFSv4.2 protocol was defined, that established **parallel NFS (pNFS)** which could support PFS characteristics over NFS.



The standards committee defined several modes of operation for pNFS. One of these modes, **Flex Files**, allowed for the use of standalone NFSv4.2 metadata servers with NFSv3 storage servers and supported direct access to data and file location mobility.

With pNFS Flex Files, the enterprise could achieve the extreme scalability and performance advantages of PFS without using proprietary client software. They could still use their standard Linux servers which host applications and supply the compute environment

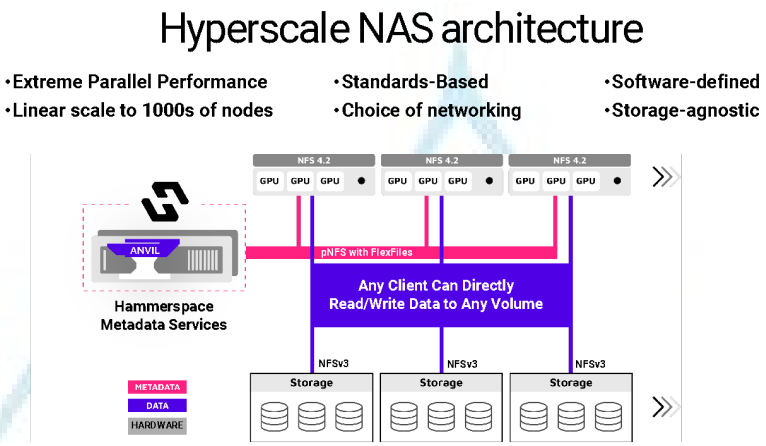
© 2024 SILVERTON CONSULTING, INC. ALL RIGHTS RESERVED

without adding any new proprietary client software or networking. Hammerspace has implemented NFSv4.2 pNFS Flex Files on their Hyperscale NAS.

Hammerspace Hyperscale NAS

A typical Hammerspace Hyperscale NAS cluster consists of Anvil NFSv4.2 metadata server(s), data service (DSX) nodes and existing scale-up or scale-out NAS or commodity-off-the-shelf systems as storage servers. Notably, Hyperscale NAS doesn't require any proprietary client software to be deployed, as pNFS Flex Files functionality is supplied as part of NFS support in the Linux® kernel.

Once Hyperscale NAS is installed, it launches a "data-in-place" migration to incorporate any NAS storage configured for it to use. This **metadata assimilation** process reads file metadata from current storage systems into Anvil servers. Even before metadata assimilation completes, IT can point file shares and mount points to Hyperscale NAS and take advantage of all the capabilities of pNFS Flex Files in their data center, turning existing NAS infrastructure into a high-performance PFS.



Some of the immediate benefits of deploying Hyperscale NAS include:

- **File location flexibility** – File data physical location is completely controlled by Hyperscale NAS and is completely divorced from directory paths and file system constructs.
- **Extreme hardware scalability** – IOPS, bandwidth and capacity can be easily increased by adding more storage server hardware.
- **Increased performance** – Access performance can often improve even with just the current NAS hardware when deployed under Hyperscale NAS.
- **Linear performance scalability** - Performance scales linearly to any performance required.
- **Infrastructure efficiency** - Near 100% of the performance of underlying storage, server, and networking infrastructure are delivered to support data access for the compute environment.

Hammerspace Hyperscale NAS clients can run with any Linux server that supports Linux 5.15+ and 6+ kernels, including RHEL 7.7+ and above. As Hyperscale NAS only uses standard Linux NFS client software, there is no need to install proprietary software on host servers.

Moreover, Hyperscale NAS can use any storage server that supports NFSv3 or NFSv4.1. A Linux system with DAS can run standard NFS server software and be used as a Hyperscale NAS storage server with no additional software required. Network-attached computational storage (SSDs with spare compute) or edge systems running Linux could thus become Hyperscale NAS storage servers.

© 2024 SILVERTON CONSULTING, INC. ALL RIGHTS RESERVED

Hyperscale NAS utilizes NFS4.2 connectivity to the client to provide parallel file system capabilities.

Hammerspace Hyperscale NAS also has NVIDIA® GPUDirect Storage support. With GPUDirect, GPUs can access data directly over the server's PCIe bus without having to pass through CPU memory. Further, Hyperscale NAS offers a container storage interface (CSI) service to enable shared read/write access to all users and applications running in K8S clusters.

For data center networking, Hyperscale NAS can support both Ethernet or InfiniBand and includes RDMA protocol data access. RDMA can speed data transfers across networks from storage to clients that access data.

Hammerspace Anvil metadata servers and DSX systems can be deployed on bare metal servers, on VMs, in containers or in any cloud running Linux. Hyperscale NAS metadata and DSX servers can operate under both VMware and Microsoft Hyper-V virtualization.

Hyperscale NAS storage servers support just about any scale-up or scale-out NAS systems, including Windows File Server. In addition, Hyperscale NAS can run on Amazon® AWS™, Google Cloud Platform™, Microsoft® Azure™, and a wide range of other Cloud offerings, and can use cloud NFS file storage.

Finally, Hyperscale NAS Anvil metadata servers are usually deployed in HA pairs that mirror all metadata between them. Anvil metadata servers can also be deployed and mirrored to remote locations.

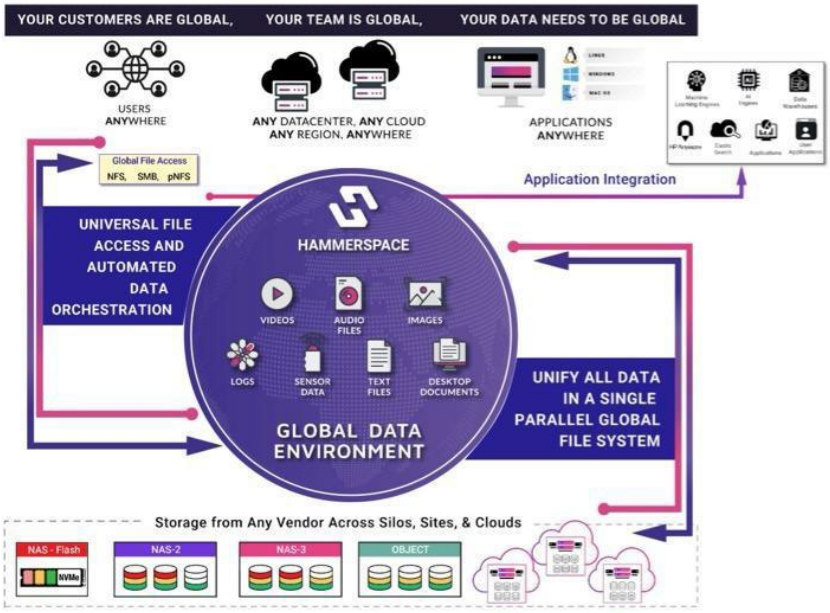
Global Data Environment

The Hammerspace **Global Data Environment (GDE)** is based on Hyperscale NAS and built on a global file system which unifies data from any storage silo into a single file system and presents all data as a single global namespace. Data generated by any user or application, using any protocol (NFS, SMB or S3) can be accessed over pNFS to a Linux client, providing the benefits of PFS performance for data.

This is particularly useful when training GenAI models, collaborating on video editing and rendering, or doing HPC analytics. GDE enables enterprise applications to access data residing in multiple sites and from multiple locations.

© 2024 SILVERTON CONSULTING, INC. ALL RIGHTS RESERVED

HAMMERSPACE HYPERSCALE NAS



One GDE can be distributed across 16 different on-premises or cloud locations, each consisting of a Hyperscale NAS cluster with a pair of Anvil metadata servers, DSX servers, and leverages existing storage nodes. Each Hyperscale NAS cluster can have anywhere from a few storage servers up to thousands of servers per location.

Hyperscale NAS, with its Flex Files pNFS operations and Anvil servers, maintain consistency of changes to data by keeping metadata continuously synchronized, even across multiple sites. When a file is updated, only its metadata needs to be synchronized for users globally to have access to the current version of data.

With a GDE, all locations in the same data center or across the world (e.g., in the cloud, colocation centers or edge clusters) share access to the same file data. Depending on access rights, the same shares, directories, and files can all be accessed from any location.

For example, if data does need to be moved local to a compute cluster, data orchestration policies can move only the data needed, at a file granular level, to where it needs to be processed. This data can continue to be active and in use by users and applications even while being moved between sites. File data can even be pre-staged to different locales and moved off to other locations once local processing is complete. These are not file copies, but instantiations of the same file with the same file metadata.

With environments requiring Hyperscale NAS pNFS v4.2 with Flex Files performance, file data can be located on any NFSv3 or NFSv4.1 storage server in the GDE.

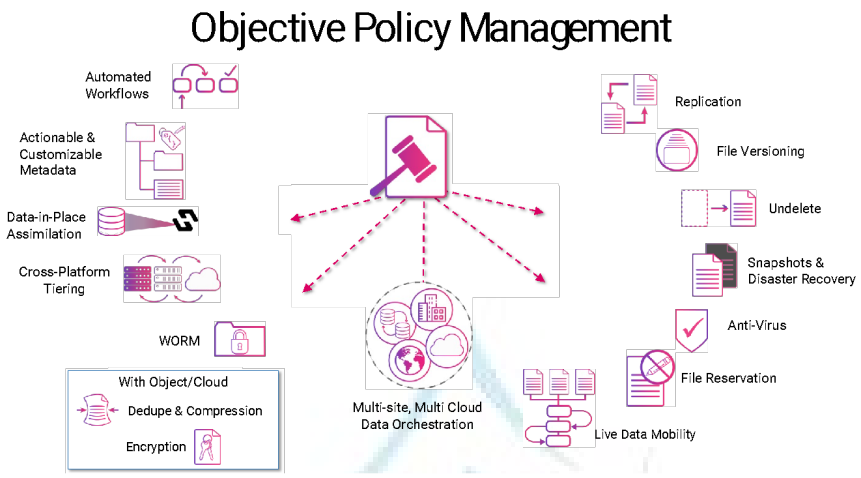
Organizations that have SMB storage deployed in their data center or their own S3 object storage can use Hammerspace DSX servers in front of their SMB or S3 storage and unify that data with other data residing in Hyperscale NAS into a single Global Data Environment. The Hammerspace Global Data Environment incorporates all this data into a single namespace and global file system which includes data generated and accessed over NFS4.2, other versions of NFS, SMB and S3.

With Hyperscale NAS, directories and the files they hold are completely divorced. The GDE takes this flexibility to a whole new level. For instance, a single directory can contain file data residing on more than hundreds to thousands of storage nodes across up to 16 locations.

© 2024 SILVERTON CONSULTING, INC. ALL RIGHTS RESERVED

Objective Policy Management

GDE also offers **Objective-based Policy Management**, which supplies data management, governance, and orchestration services such as tracking file changes/versions, and moving files closer to the servers accessing them.



In addition to the above, Objective-based Policy Management can be used to schedule snapshots, create encrypted-deduped-compressed WORM copies of file data, automate data workflows, and perform anti-virus scans. All this can help to achieve required business policies for compliance, governance, or geographic access constraints of data in the GDE.

Objective-based policy management can also be used to set up data protection for storage servers by mirroring data or using RAID-like erasure coding. With storage server data protection, even multiple storage server failures cannot halt access to data.

Mirroring creates one to two more copies of data to ensure that no single or double storage server failure impacts data availability. Erasure coding is a more sophisticated data protection solution that uses parity and data storage for data protection. It can be configured to support multiple storage server failures without loss of data access. While mirroring doubles or triples storage capacity requirements, erasure coding consumes only a modest amount of additional storage capacity for parity to provide data protection.

In addition, data striping, a new capability offered in an upcoming release, can be used to split file data across multiple storage servers for simultaneous access to data.

Hyperscale NAS performance enhancements

As with any PFS, Hyperscale NAS can support PBs to Exabytes of file storage capacity with billions of files spread across multiple file systems/directories, all with extremely high performance. Performance attributes of Hyperscale NAS include high throughput, fast metadata access, and high-speed read or write I/O. Environments with different performance workloads or mixed I/O demands can consolidate those workloads into Hammerspace to have a single global file system, instead of using separate file systems for each workload.

Linux standard pNFS with Flex Files offers many performance improvements over traditional NAS solutions. For example, direct data access can provide a significant performance boost and improved hardware utilization by removing file system overhead from the data path. Hyperscale NAS works well for both large files and small, highly accessed files. Moreover, with Hyperscale NAS, separate Anvil

© 2024 SILVERTON CONSULTING, INC. ALL RIGHTS RESERVED

metadata servers offer independent access to metadata. While this capability may not make much of a performance difference for large files, it offers a significant performance boost to applications using small files, whose access is metadata intensive.

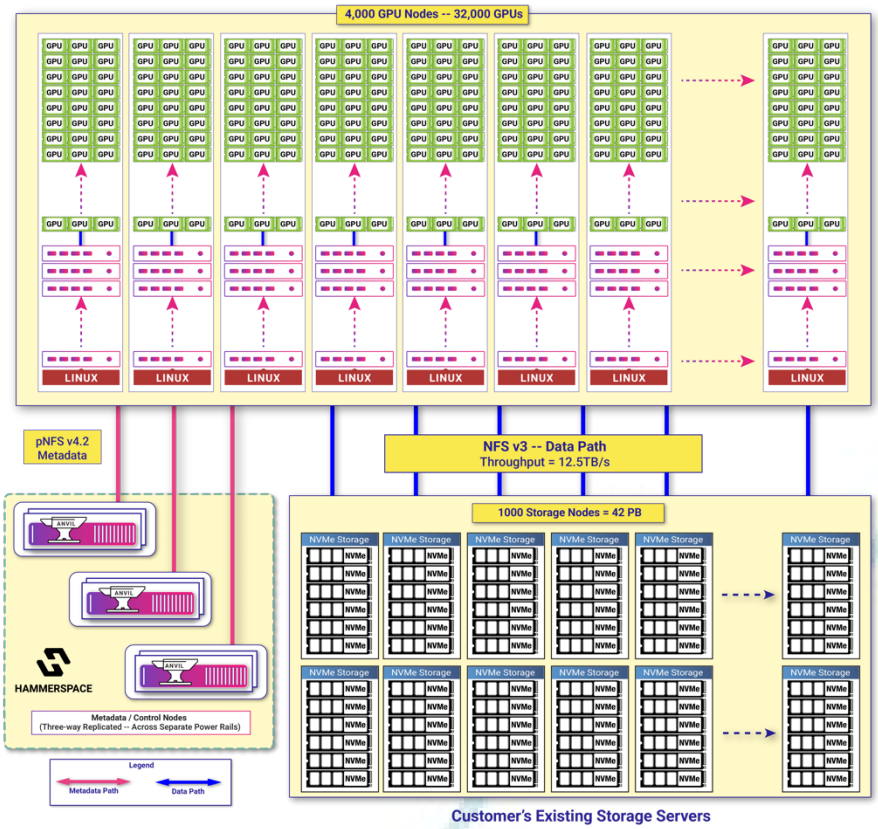
Furthermore, NFSv4.2 has reduced the number of protocol roundtrips required to open a file, which can increase performance for applications using lots of small files. Hyperscale NAS using pNFS v4.2 with Flex Files also offers live file mobility, which means hot files can be moved to hot storage while cold data can be moved to cold storage. This capability works entirely in the background, even on files that are active and in use, and can significantly improve access performance for both large and small files even on legacy storage infrastructure.

Applications can thus process large file read and write access with very high throughput. Small files IOPS rates can be significant as well. These combined features can readily support AI training, fine tuning, and inferencing data access requirements for even the largest organizations.

If an enterprise is fine-tuning a corporate GenAI model, hundreds or even thousands of GPUs and servers can remain busy at all times. However, if an enterprise is simply deploying a pre-built GenAI solution for inferencing, 10 to 100 GPUs and servers can remain active without interruption. In addition, if an enterprise is training their own large language model (LLM) GenAI solution, more than 10,000 GPUs and thousands of servers can remain 100% utilized, all the time.

© 2024 SILVERTON CONSULTING, INC. ALL RIGHTS RESERVED

Use Cases



Large AI training center

A large AI training environment deployed Hyperscale NAS to support file data performance needed to train their GenAI LLM. In this data center, storage nodes are Linux machines running NFS server software with NVMe SSD storage. The customers data center is split up into three separate power and networking zones and Hyperscale NAS Anvil metadata server pairs are replicated three-way across these zones, and all hosts have multiple GPUs attached.

GPUs are this organization's largest infrastructure expense, and they wanted to fully utilize all their GPUs when training large AI models. They had tried other storage systems, but none could sustain the training data access performance needed to keep their GPU farm busy while also fitting in their pre-defined standards-based architectural approach.

For training GenAI LLM, neural networks are split across multiple GPUs and multiple hosts. For most GenAI training environments, host systems commonly have six or more GPUs attached. As such, many GPUs and the servers running them all read the same training data repeatedly.

This organization used Hyperscale NAS to duplicate training data across multiple storage servers, each able to provide simultaneous direct data access to multiple host servers. By deploying Hyperscale NAS, the organization was able to sustain LLM GenAI training throughput across 32,000 GPUs using around 1,000 storage servers on standard routed ethernet with TCP/IP connectivity to achieve sustained throughput of 12TB/s.

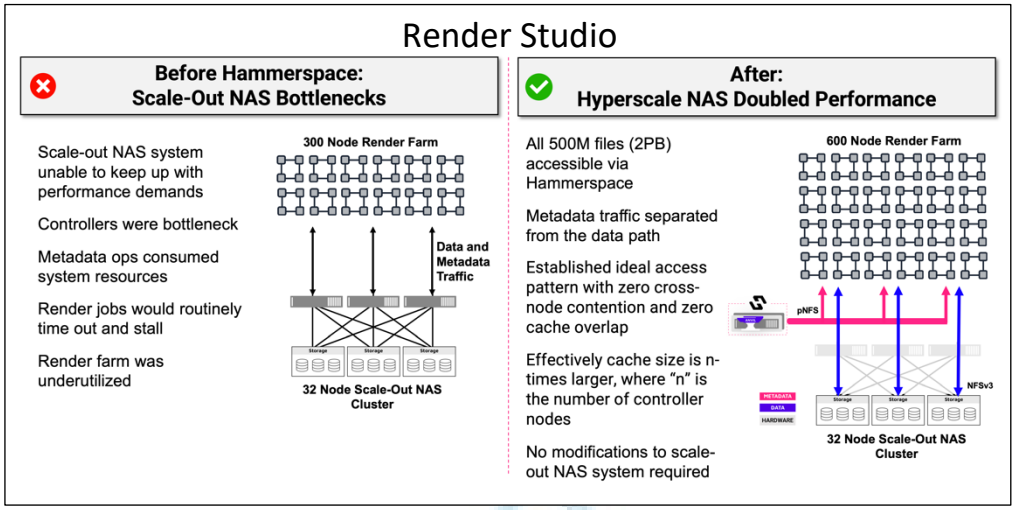
Large rendering farm

A visual effects production company had a 32-node, 2PB-500M file, scale-out NAS solution attached to their 300-server render farm.

© 2024 SILVERTON CONSULTING, INC. ALL RIGHTS RESERVED

Silverton Consulting
Strategy, Storage & Systems

Because rendering jobs were timing out due to lack of I/O performance from existing NAS storage, they had to reduce their render farm node counts and consequently decreased rendering throughput below their data center-designed specifications.



The studio had to decide whether to increase current scale-out NAS node counts, purchase a faster scale-out NAS solution or deploy Hammerspace Hyperscale NAS and pNFS with Flex Files.

They chose to deploy Hyperscale NAS and continue using their existing storage hardware, along with a pair of Hammerspace Anvil metadata nodes and two DSX servers. Within hours (for the metadata assimilation process), data access performance improved, enabling them to expand their render farm back to their original 300-node count without any I/O failures.

With Hyperscale NAS, all data still resided on the same scale-out NAS servers. However, with the Hyperscale NAS architecture in place, each scale-out NAS node became a separate storage server, providing direct access to hosts requesting data.

The studio is currently doubling the size of their render farm to further increase job throughput, all without increasing their scale-out NAS solution. All file access performance improvements resulted from deploying Hyperscale NAS over their current scale-out NAS cluster.

Summary

Hammerspace Hyperscale NAS offers a new paradigm for enterprises that need extreme levels of high-performance file data access. Hyperscale NAS offers the extreme scalability and performance of an HPC-class parallel file system, but without the need for proprietary client or storage server software.

Traditional scale-up and scale-out NAS solutions have performance and scaling limitations when trying to address the HPC-like performance needs for AI/ML and similar workloads. In contrast, performance and capacity scale linearly with Hyperscale NAS, even to extreme use cases such as noted above. The more storage hardware and networking bandwidth is deployed, the more file capacity and I/O performance can be delivered.

In addition, with pNFS v4.2 with Flex Files on Hyperscale NAS, enterprises can remove the hard-coded connection between files and the directories that contain them. Directory and file system structures no longer determine file location. Hyperscale NAS can store file data on any storage server from any

© 2024 SILVERTON CONSULTING, INC. ALL RIGHTS RESERVED

vendor, plus place data throughout a cluster and move data from one node to another, all without application changes or user disruption.

Finally, with Hammerspace's GDE, data location can become global. With up to 16 distinct locations for the global file system, data can be stitched together into one directory or share that spans locations and storage servers. Clients with appropriate permissions can access all that file data in any other cluster location, including on-premises and cloud-based datacenters, as though they were local.

In sum, pNFS with NFSv4.2 direct data access using Hyperscale NAS offers significant file access performance improvements, Flex Files pNFS and Hammerspace's GDE offer data location flexibility and Hyperscale NAS offers extreme scalability. These advantages make Hyperscale NAS the optimal solution for organizations that lack file I/O performance or that provide AI training. Hyperscale NAS can solve enterprise IT's challenges with file performance and capacity limitations today – and for the foreseeable AI-intensive future.

Silverton Consulting, Inc., is a U.S.-based Storage, Strategy & Systems consulting firm offering products and services to the data storage community.

© 2024 SILVERTON CONSULTING, INC. ALL RIGHTS RESERVED

Silverton Consulting
Strategy, Storage & Systems