Excelero

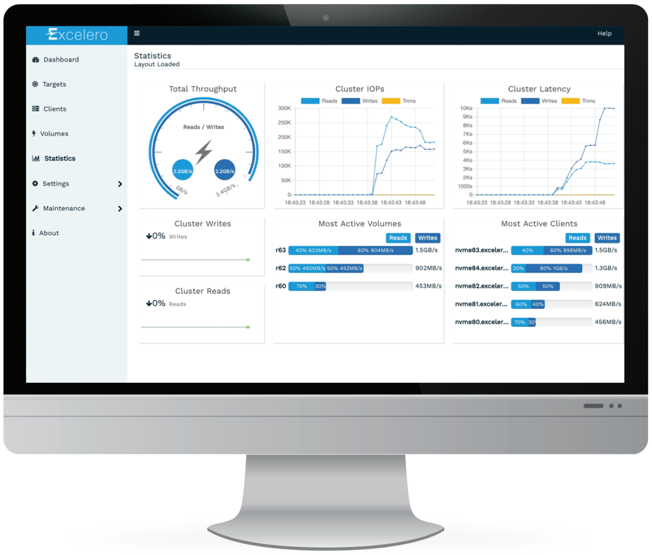
IBM SPECTRUM SCALE ACCELERATED BY NVMesh®

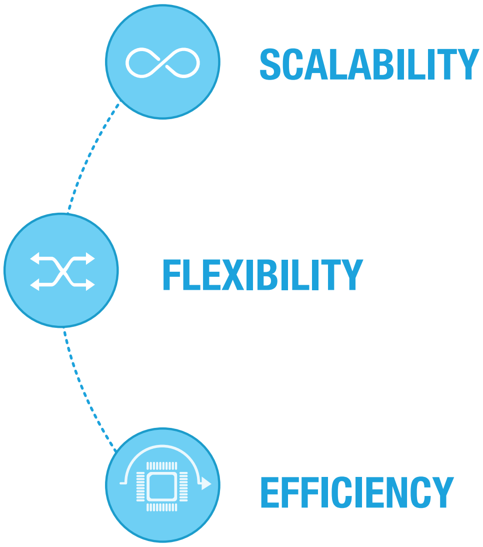
USE CASE

IBM Spectrum Scale

INTRODUCTION

Never before has the amount of data generated and the need for analytics been so high for organizations of all types. Nimble organizations are those that make use of their data via analytics in the most effective, streamlined manner: they have optimized infrastructures and applications to capture, store and analyze higher volumes of data faster. This enables better and faster business decisions. Higher performance infrastructure is required to deliver this massive increase in productivity. Compute, networking and storage all have to deliver more.





SCALABILITY

FLEXIBILITY

EFFICIENCY

TABLE OF CONTENTS

IBM SPECTRUM SCALE

IBM Spectrum Scale (aka GPFS) is a high-performance data storage and management solution that was designed to meet scale-out data challenges. Spectrum Scale features a scale-out, clustered file system that provides massively-parallel, shared read-write access to a global namespace.

Underlying storage nodes, referred to as NSD Servers, support SAN-attached (SAN-Mode, aka MultiAttach), network-attached (NAS), or a mixture of both in either a shared-everything or shared-nothing cluster configuration. These flexible topology options enable high performance access at very large scale to a common set of data supporting scale-out solutions, or to provide a highly-available storage platform.

Spectrum Scale, has been available for over two decades. Initially its primary use cases were mostly in HPC environments. Today, Spectrum Scale has matured into a leading platform for scale-out applications across a growing number of vertical markets. Requirements for Spectrum Scale storage are evolving rapidly as applications need to concurrently process more data faster.

In essence, the key requirements for a fast and successful adoption of Spectrum Scale are:

- Scalability - consistently and simply meet growing data storage requirements
- Performance - process large data sets with consistent high performance
- Simplicity - manage more data with fewer resources

Users of Spectrum Scale are constantly looking to improve their underlying storage solutions to better meet these challenges. Applications utilizing Spectrum Scale benefit significantly from next-generation storage media like NVMe SSDs in the NSD servers that connect directly to the server's PCIe bus. This accelerates Spectrum Scale NSD servers, but the distributed nature of Spectrum Scale clients demands NVMe-based storage be manageable in similar fashion.

EXCELERO NVMesh

Excelero NVMesh enables distributed applications to remotely access pooled-NVMe resources with the same performance of local storage, centrally managed and at data center scale. NVMesh is the only software-defined Server SAN supporting mixed topologies for shared, scale-out, NVMe block-storage: it remains the only solution to date that is 100% software-only.

In this document, we provide an overview of the features and benefits of Excelero's NVMesh for Spectrum Scale. Additionally, we offer two reference architectures of popular NVMesh-Spectrum Scale configurations, complete with performance tests and results.

EXCELERO NVMesh ACCELERATES IBM SPECTRUM SCALE

Scale Flexibly and Efficiently

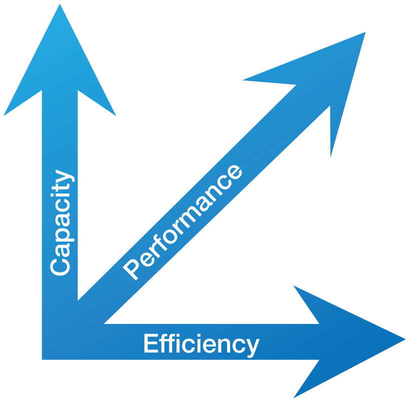
NVMesh allows users to meet their scalability requirements without compromise. Performance and capacity can be scaled as needed, removing bottlenecks typically introduced by traditional SAN array controllers/solutions.

NVMesh is deployed as a virtual, distributed, NVMe-focused Server SAN, and supports both converged and disaggregated storage architectures in the same deployment, giving customers full freedom with their designs. Even after initial deployment, the storage requirements may evolve as capacity or performance demands change. Storage can be added or removed incrementally - without impacting other workloads.

NVMesh allows unmodified applications to utilize pooled NVMe storage across a network at local speeds and latencies. A unified pool of NVMe enables customers to maximize NVMe utilization and avoids legacy concerns for maintaining data locality. NVMesh introduces as little as 5 µs of network round-trip access latency over that of the same media accessed locally...

This provides many benefits:

- No unnecessary upfront investments - scale capacity and performance as needed
- Maintain highest efficiency
- Re-allocate or move flash storage resources when needed
- 100Gb Ethernet as a SAN provides 3 times the bandwidth of Fibre Channel at half the cost



PROCESS MORE DATA FASTER

A key component of Excelero NVMesh is its patented Remote Direct Drive Access (RDDA) transport,which bypasses the CPU on target systems, and eliminates the noisy-neighbor effect often found in traditional software-defined storage solutions. The scale-out architecture of NVMesh shifts data services and IO processing from centralized storage target CPUs out to the distributed storage clients (initiators) instead. This unique approach enables linear scalability, deterministic performance, and the ability for customers to maximize utilization of their flash investments.

RDDA boosts performance in all aspects:

- Lowest latency
- Maximum IOPS
- Highest throughput density

NVMesh INTEGRATION WITH SPECTRUM SCALE

Excelero NVMesh offers low-latency local access to remote NVMe devices. This capability plays well with various Spectrum Scale capabilities. The following sections describe various integration best practices between NVMesh and Spectrum Scale features.

Spectrum Scale Caching Configurations

Spectrum Scale provides multiple cache configurations that are intended to accelerate the overall performance:

Local read-only cache (LROC)	**Reduces the latency of small read IOs on Spectrum Scale clients by caching on a local flash media. LROC is configured on Spectrum Scale client nodes.**
Highly-available write cache (HAWC)	**Reduces the latency of small write IOs by writing first to flash media prior to writing it to the backend spinning media. HAWC is configured on Spectrum Scale NSD servers or client nodes. Requires HAWC pools to be on Highly Available volumes.**

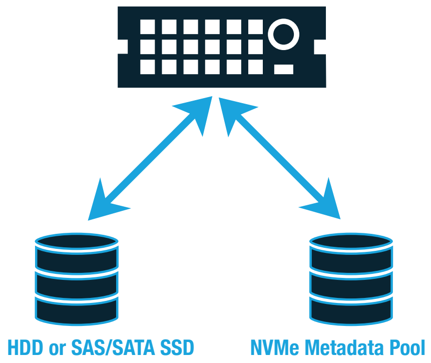
NVMesh eliminates the need to install individual local SSDs for LROC, or SSD pools on external arrays for HAWC. Instead, NVMesh Volumes offer the same performance as local flash but virtually constructed from a pool of remote flash drives. This enables greater flexibility in the configuration of both HAWC and LROC, since volumes can be configured regardless of the availability of flash media on the Spectrum Scale nodes. In addition, caches can be resized and reallocated to other Spectrum Scale nodes when needed, without making any hardware changes.

For more information on Spectrum Scale LROC and HAWC, refer to the IBM Spectrum Scale Knowledge Center.

Spectrum Scale Dedicated Metadata Pools

Spectrum Scale metadata operations are IO intensive. Small file creations, inode scans (such as Linux/Unix “finds” or the Spectrum Scale policy engine), directory traversals and deletions all create a massive amount of random small reads and writes in large scale file systems.

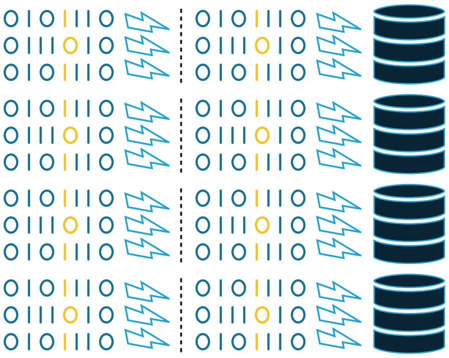
To speed up these operations, especially in a large file system with spinning media, it is recommended to configure dedicated high performance metadata pools. NVMesh enables the configuration of dedicated NVMe flash Metadata Pools across many NSDs without actually installing drives in each node.



Moreover, NVMesh volumes can be mirrored and protected with much lower write latencies than built in Spectrum Scale replication. Since NVMesh provides low latency performance, Spectrum Scale metadata is highly accelerated, providing an overall faster response time for the file system, even if the data itself is stored on spinning storage resources.

Spectrum Scale Burst Buffer Cache

NVMesh ensures Job SLA's and accelerates checkpoints. Spectrum Scale can achieve high throughput writes even with its internal replication pool mechanisms. NVMesh can achieve slightly better throughput in most situations but the real value is in operational management. With NVMesh, Spectrum Scale and the administrators of a cluster need not manage drive failures, replacements nor replication pool adjustments. NVMesh effectively shields Spectrum Scale from these issues with logical volumes where errors, failures and recoveries are handled automatically.



All data is local with NVMesh

When Spectrum Scale is configured in SAN-Mode (multi-attach), Spectrum Scale clients have the ability to access the file system's underlying block devices as if they were local. This means read/write IO does not have to be proxied through the NSD servers and incur latency and performance penalties.

In traditional Spectrum Scale deployments, where the client nodes access the backend storage through the NSD servers, NVMe storage performance cannot be fully harnessed due to NSD proxy overhead.

This is a crucial fact to consider when using small-block (sub 32k), random read/write IO on the Spectrum Scale file system.

An additional NVMesh benefit in multi-attach mode is Spectrum Scale licensing simplification. A minimal number of systems licensed as Servers are required for the Quorum and Manager roles - often just three to five - while the rest of the systems may use the lower-cost Client license.

Easy to install, manage and monitor

- Faster and less complex Spectrum Scale setup
- Simplified backend storage configuration
- Reduced operational complexity
- Reduced cost

NVMesh was designed with ease of use and web-scale deployments in mind. Maximizing operational efficiency requires consolidation of storage silos and the ability to run multiple applications on a single storage platform. By leveraging the capabilities of NVMe, applications can be provisioned with volumes meeting all requirements (scale, performance, availability, reliability, efficiency and cost) ensuring internal or external SLO's.

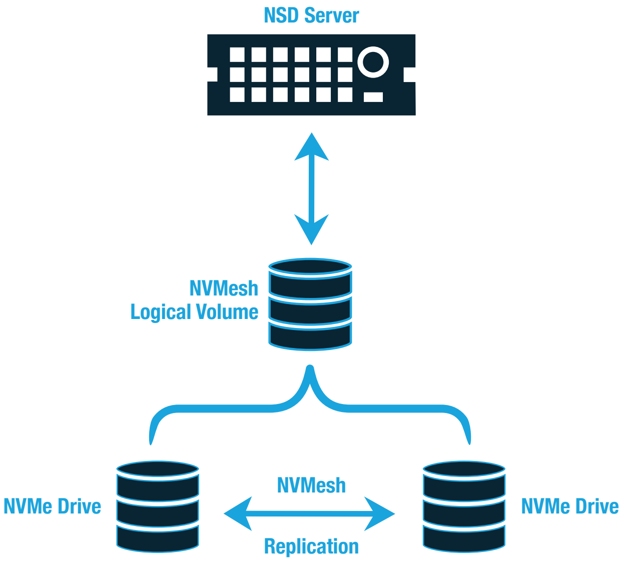
NVMesh is easily managed through the included web interface, or through the RESTful API from which the WebUI was developed. The API allows for seamless integration into pre-existing storage management and provisioning workflow automation tools. If the RESTful API isn't for you, Excelero hosts a community-maintained set of tools on Github for functional examples and scripting. NVMesh offers specific simplicity benefits when used with Spectrum Scale, further reducing complexity and increasing file system efficiency.

Because NVMesh doesn't consume target-side CPU, servers acting as NVMesh targets can also run the Spectrum Scale server processes. This effective eliminates a level of server infrastructure that is necessary with external arrays, yet is much faster, saving money while improving performance and efficiency.

Excelero NVMesh uses either InfiniBand or RDMA over Converged Ethernet v2 (RoCE v2) to enable the sharing of high-performance NVMe flash resources at data center scale. Open-network support enables customers to leverage existing network hardware investments rather than requiring dedicated storage networks.

No Spectrum Scale Replication Groups Required

- Abstracts the complexity of delivering highavailability and eliminates the performance impacts of using Spectrum Scalereplicated volumes.
- More front-end bandwidth available for applications and users
- Lower latency



REFERENCE ARCHITECTURES

The following architectures were used to test the IBM Spectrum Scale performance when configured on NVMesh volumes:

- 8 Clients Multi-Attach
- 12 Clients Multi-Attach

8 CLIENTS MULTI-ATTACH

Hardware

One Supermicro SYS-2028BT-HNR+ BigTwin 4-server system:

Dual Intel Xeon E5-2630 v4 CPUs per server

- 256 GB RAM per server
- NVMesh Storage Targets and Spectrum Scale NSD Servers
- Only for NSD baseline testing – omitted/not configured for Multi-Attach client performance testing
- Six Micron 9100 2.4 TB NVMe SSDs per server – 24 in total
- Two Mellanox ConnectX-4 100Gb Ethernet Adapters per server – 8 in total

Two Supermicro SYS-2028BT-HNR+ BigTwin 4-server systems:

- Dual Intel Xeon E5-2690 v4 CPUs per server
- 128 GB RAM per server
- NVMesh Storage Clients and Spectrum Scale File System Clients
- One Mellanox ConnectX-5 100Gb Ethernet Adapters per server – 8 in total

One 16 port Mellanox SN2100 100Gb Ethernet Switch:

- RoCE v2 configuration, with PFC and ECN configured for Lossless Ethernet QoS

Software

- RedHat Enterprise Linux v7.3
- NVMesh v1.2.1
- IBM Spectrum Scale v4.2.3-6

NVMesh NVMe Target Configuration

- RAID-10 configuration
- 2D + 2P Striping Configuration
- Eight NVMesh Volumes provisioned to all Spectrum Scale servers and clients.

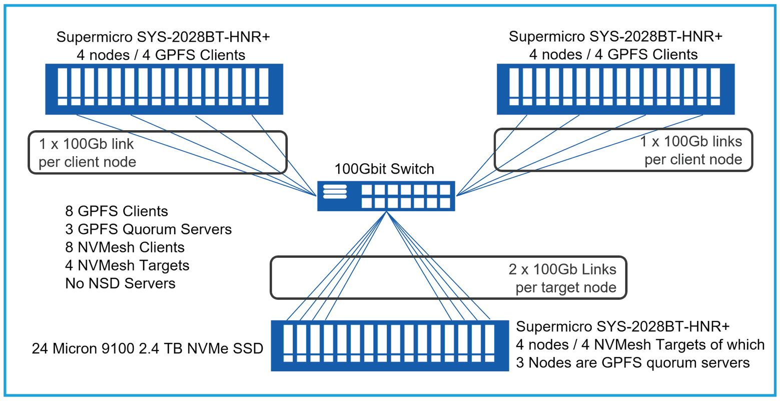


Figure 1 - 8-node Multi-Attach NVMesh/IBM Spectrum Scale configuration

12 CLIENTS MULTI-ATTACH

Hardware

One Supermicro SYS-2028BT-HNR+ BigTwin 4-server system:

- Dual Intel Xeon E5-2630 v4 CPUs per server
- 256 GB RAM per server
- NVMesh Storage Targets and Spectrum Scale NSD Servers
- NVMesh Clients and Spectrum Scale File System Clients
- Six Micron 9100 2.4 TB NVMe SSDs per server – 24 in total
- Two Mellanox ConnectX-4 100Gb Ethernet Adapters per server - 8 in total

Two Supermicro SYS-2028BT-HNR+ BigTwin 4-server systems:

- Dual Intel 2690 per node/server
- 128 GB RAM per node/server
- Used as NVMesh Storage Clients and Spectrum Scale Clients/File System Nodes only
- One Mellanox ConnectX-5 100Gb Ethernet Adapters per server – 8 in total

One 16 port Mellanox SN2100 100Gbit Ethernet Switch:

- RoCE v2 configuration, with PFC and ECN configured for Lossless Ethernet QoS

Software

- RedHat Enterprise Linux v7.3
- NVMesh v1.2.1
- IBM Spectrum Scale v4.2.3-6

NVMesh NVMe Target Configuration

- RAID-10 configuration
- 2D + 2P Striping Configuration
- Eight NVMesh Volumes provisioned to all Spectrum Scale servers and clients

Spectrum Scale Configuration

One filesystem – capacity 21TB

NSD configuration:

usage= dataAndMetadata

servers= not specified/left blank (for SAN-mode)

Filesystem Properties / mmcrfs options:

-A no -B 128K -D posix -E no -i 4096 -k posix -j scatter -S no -Q no -z no --profile=Spectrum ScaleProtocolRandomIO

Spectrum Scale Cluster Tuning:

- RDMA enabled
- RDMA verbs configured for the Mellanox adapters
- workerThreads increased to 2048

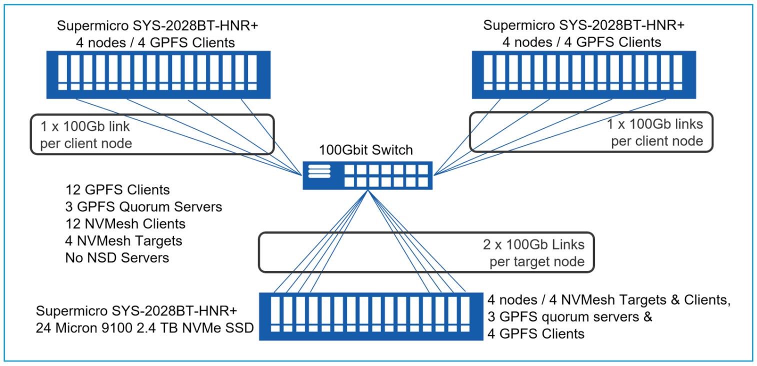


Figure 2 - 12-node Multi-Attach NVMesh/IBM Spectrum Scale configuration

PERFORMANCE TEST RESULTS

Overview

For the purpose of testing the performance of NVMesh combined with IBM Spectrum Scale, a single file system was created with the capacity of 21 terabytes over the 24 2.4TB NVMe SSDs. A total of eight 3.6TB large file sets with 1GB files size were created (36,000 x 1GB files) using fio v3.7. In each test run, each of the clients ran eight jobs per file set in parallel with 4K block size and IO depth of 4.

The following Spectrum Scale configuration was used for the testing:

NSD configuration:

usage= dataAndMetadata

servers= not specified/left blank (for SAN-mode)

Filesystem Properties / mmcrfs options:

-A no -B 128K -D posix -E no -i 4096 -k posix -j scatter -S no -Q no -z no -

-profile=Spectrum ScaleProtocolRandomIO

Spectrum Scale Cluster Tuning:

RDMA enabled

RDMA verbs configured for the Mellanox adapters

workerThreads increased to 2048

GENERAL PERFORMANCE RESULTS

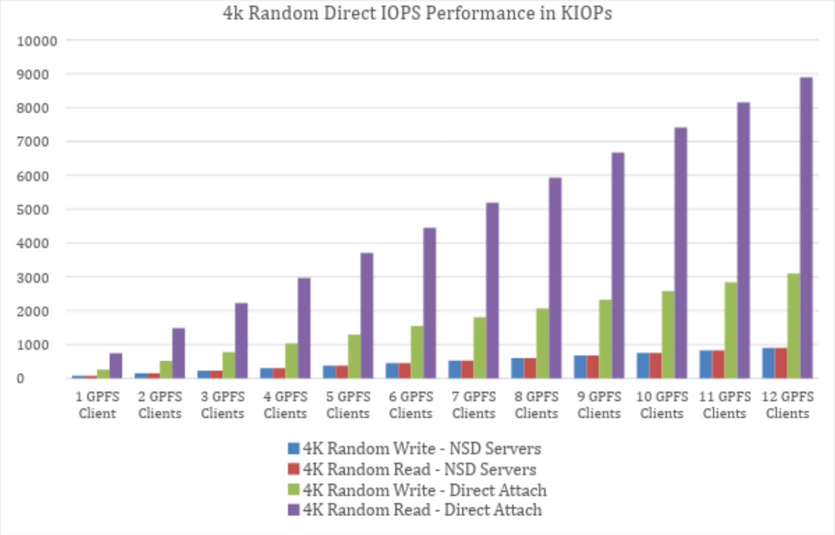
The following Spectrum Scale performance was observed:

	8 Clients		12 Clients	
	Read	Write (to mirrored volume)	Read	Write (to mirrored volume)
Throughput – 1MB Large Sequential direct IO	71.6 GB/sec	20.9 GB/sec	71.6 GB/sec	20.9 GB/sec
Random IO - 4K Random direct IOPS	5.9M IOPS @304 µs	2.5M IOPS @196 µst	8.9M IOPS @310 µs	3.1M IOPS @192 µs

In all tests CPU and memory utilization on the NVMesh Target servers was negligible.

NSD SERVERS VS. MULTI-ATTACH STORAGE ACCESS

Spectrum Scale 4K Random Read/Write IO



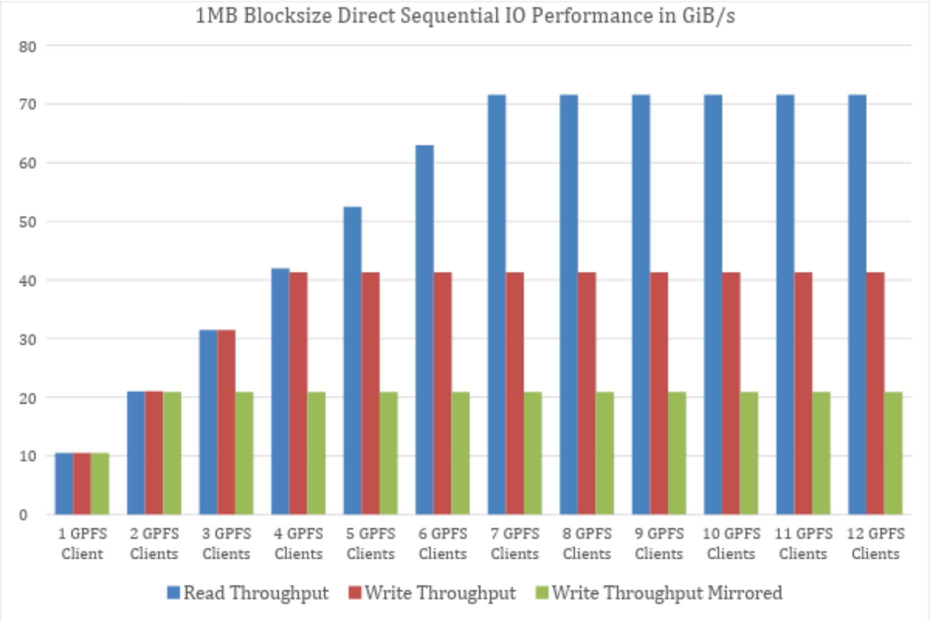
Maximum direct IOPS utilizing 4 NSD servers:

- Read, 600,000 IOPS @468 µs
- Write, 600,000 IOPS @702 µs

Maximum direct IOPs in direct attach mode and not utilizing NSD servers:

- Read, 8.9M IOPS @310 µs
- Write, 3.1M IOPS @192 µs

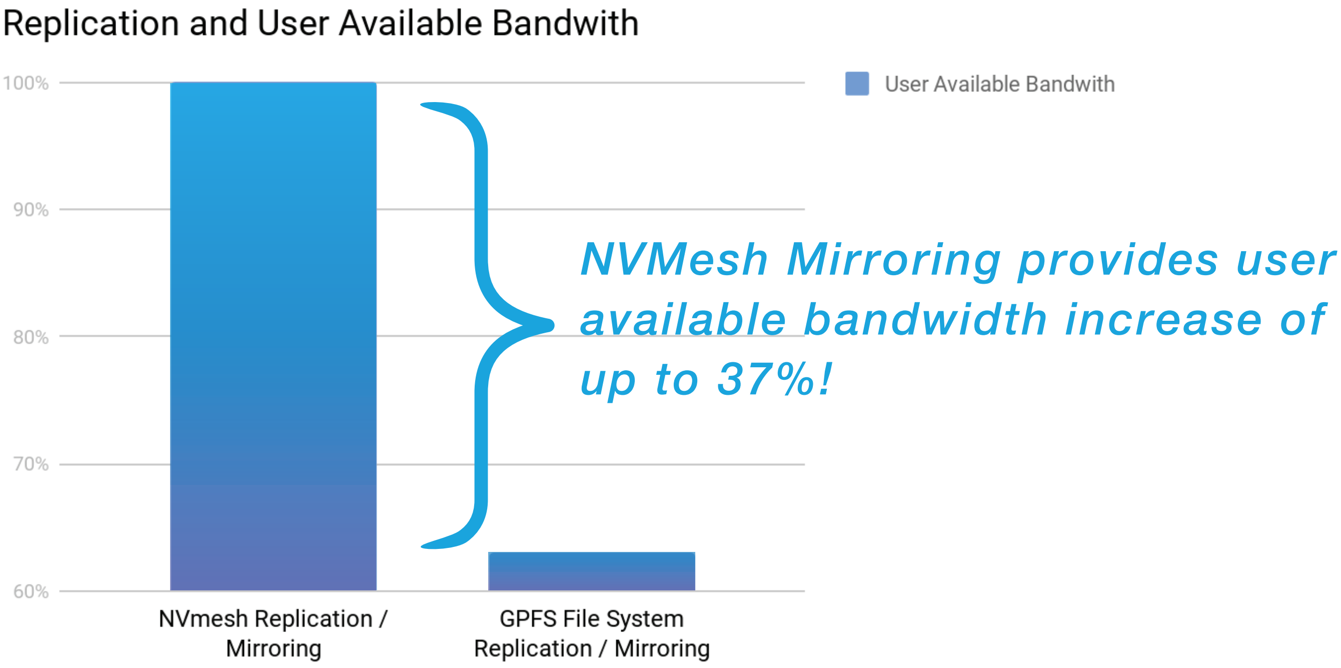
Spectrum Scale Large IO Performance – 8/12 Clients NVMesh Multi-Attach Reference



Maximum Throughput:

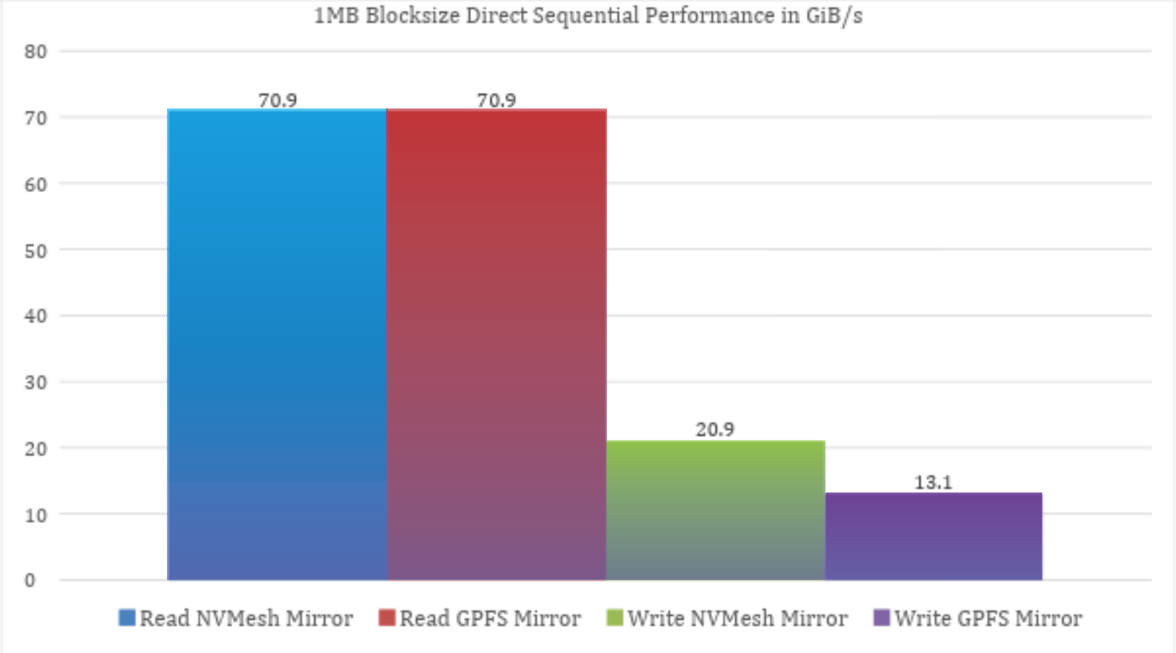
- Read: 71.6 GiB/s
- Write: 42.3 GiB/s
- Write Mirrored: 20.9 GiB/s

NVMesh Mirroring vs. Spectrum Scale Filesystem Replication



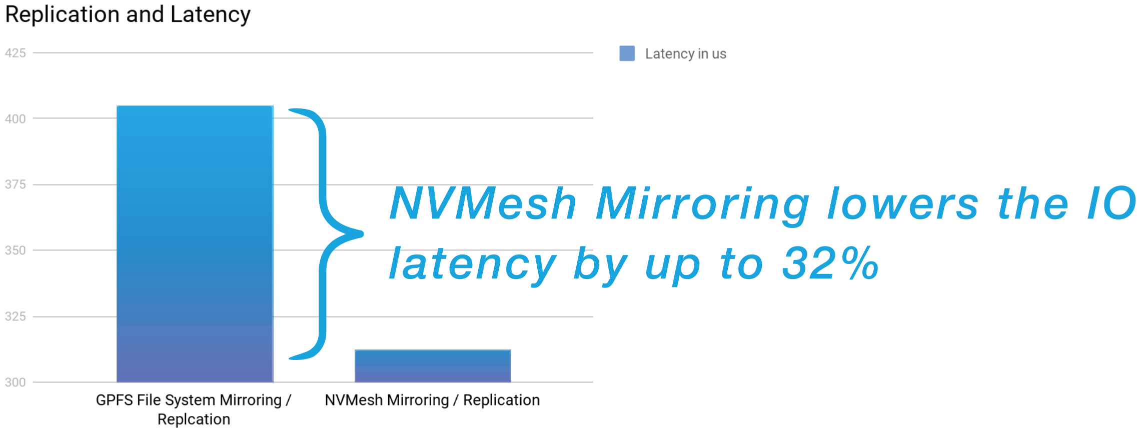
NVMesh Mirroring provides user available bandwidth increase of up to 37%!

Mirrored Large IO Throughput Performance – 8/12 Clients Reference



Maximum Parallel Direct 1MB Blocksize Mirrored Write IO Throughput:

- NVMesh Mirror: 20.9 GiB/s
- Spectrum Scale Mirror: 13.1 GiB/s
- Latency Direct 4K Block size Random Write IO:
- NVMesh Mirror: 192 μs
- Spectrum Scale Mirror: 253 μs



SUMMARY

NVMesh is the world's fastest, lowest latency software SAN product that removes bottlenecks, lowers cost and protects data on commodity NVMe media and servers. The combination of NVMesh and Spectrum Scale allows you to build a parallel, distributed storage architecture that can be tuned for any workload necessary. With this combination, you can achieve high bandwidth reads or writes, high rates of random small-block IO at low latency or both. The solution is readily and linearly scalable. This potent combination allows administrators, application developers and users to stop worrying about limitations of the file system and start thinking about all the new possibilities limitless IO affords.

For more information, refer to Excelero website at http://www.excelero.com

If you have comments about this documentation, submit your feedback to docs@excelero.com

Information in this document is subject to change without notice.

Excelero, Inc.
San Jose, CA
United States
www.excelero.com