TECHNICAL WHITE PAPER
August 2024

Transform Your HCI with Next-Gen vSAN 8.0 Architecture:

VMware vSAN ESA with Samsung PM1743 NVMe SSD

vmware by Broadcom

Table of contents

Executive Summary

This white paper provides a comprehensive guide for IT professionals seeking to optimize the performance of their applications deployed on VMware vSAN Express Storage Architecture (ESA). By focusing on the integration of Samsung's PM1743 NVMe SSD devices to vSAN ESA, this study delves into the impact of various configuration parameters on real-world application workloads.

Through a series of rigorous experiments and benchmarks, we demonstrate how adjustments to vSAN ESA settings, such as storage policies, number of disks, Remote Direct Memory Access (RDMA), etc. can significantly enhance performance and responsiveness of applications running on the innovative storage platform.

Key findings of the experiments conducted for this white paper are:

- **Peak Performance:** VMware vSAN ESA combined with Samsung's PM1743 NVMe SSDs demonstrated a maximum performance of 1.49 MIOPS and throughput of 25.45 GB/s in a cluster using three hosts and 100 Gb/s networking.
- **Optimal Block Size:** the optimal block sizes for achieving peak performance were 64 KB and 128 KB.
- **RAID 5 Efficiency:** employing a RAID 5 storage policy with VMware vSAN ESA resulted in a comparable performance to that of RAID 1 while reducing the storage footprint required for object replica capacity by 50%.
- **RDMA Enhancement:** leveraging RDMA with VMware vSAN ESA, performance for write and mixed workloads was improved by 25%, accompanied by an 18% reduction in CPU resource utilization.

Our findings offer valuable insights into the optimal configuration of vSAN ESA with Samsung PM1743 NVMe SSDs, enabling organizations to maximize the potential of their investments and deliver exceptional user experiences. This guide serves as an essential resource for IT administrators and architects seeking to unlock the potential of VMware vSAN ESA and ensure their applications run at peak efficiency.

Introductions & Backgrounds

Automated SMRC Analytic Platform (ASAP) Framework

Samsung Memory Research Center (SMRC) provides customized solutions based on more than 30 years of Samsung's accumulated know-how in memory technology. SMRC aims to accelerate its customers' next evolution through innovative technology collaborations, and as part of this effort, ASAP has been developed, which integrates Samsung's expertise and technology with an automated analytics system.

Specifically, ASAP 2.0 was developed to help customers and IT partners find optimal solutions faster and more efficiently by providing a richer experience of different Software Defined Storage (SDS) offerings in the market. The platform is designed to build on SMRC's technical expertise and experience in delivering solutions tailored to customers' specific needs, and we expect it to automatically produce optimized results to help customers grow their business.

ASAP is capable of supporting diverse test scenarios across various environments., We employed Samsung PM1743 NVMe SSDs for a VMware vSAN Hyper-converged Infrastructure (HCI) Express storage architecture along with hardware featuring Intel's Ice Lake processors and PCIe Gen4 SSD. Benchmarking was conducted using the FIO tool, which is explained following sections, encompassing test scenarios for both vSAN over TCP and RDMA protocols.

While ASAP is designed to automate performance test for various storage platforms, this white paper utilizes a manually executed benchmark test to verify vSAN ESA performance as a reference point.

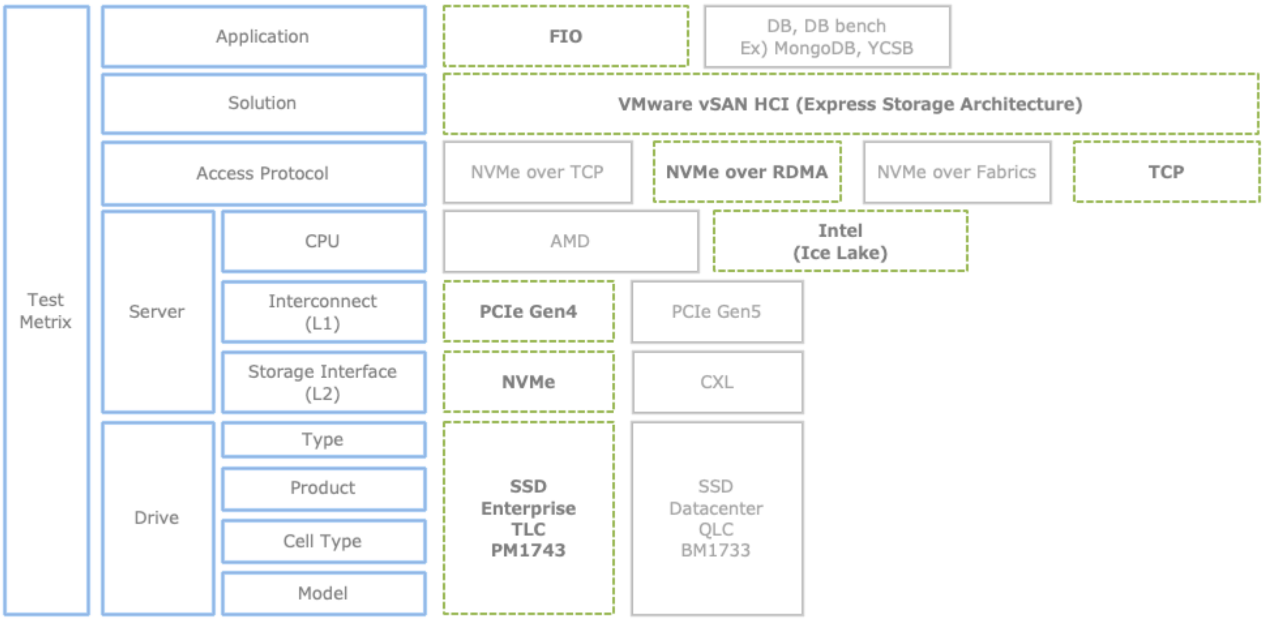

Figure 1. ASAP Supported Target Infrastructure

VMware Cloud Foundation (VCF)

VCF is a hybrid cloud platform that combines the scale and agility of public cloud with the security and performance of private cloud. It provides a full-stack HCI that is made for modernizing data centers and deploying modern container-based applications [1].

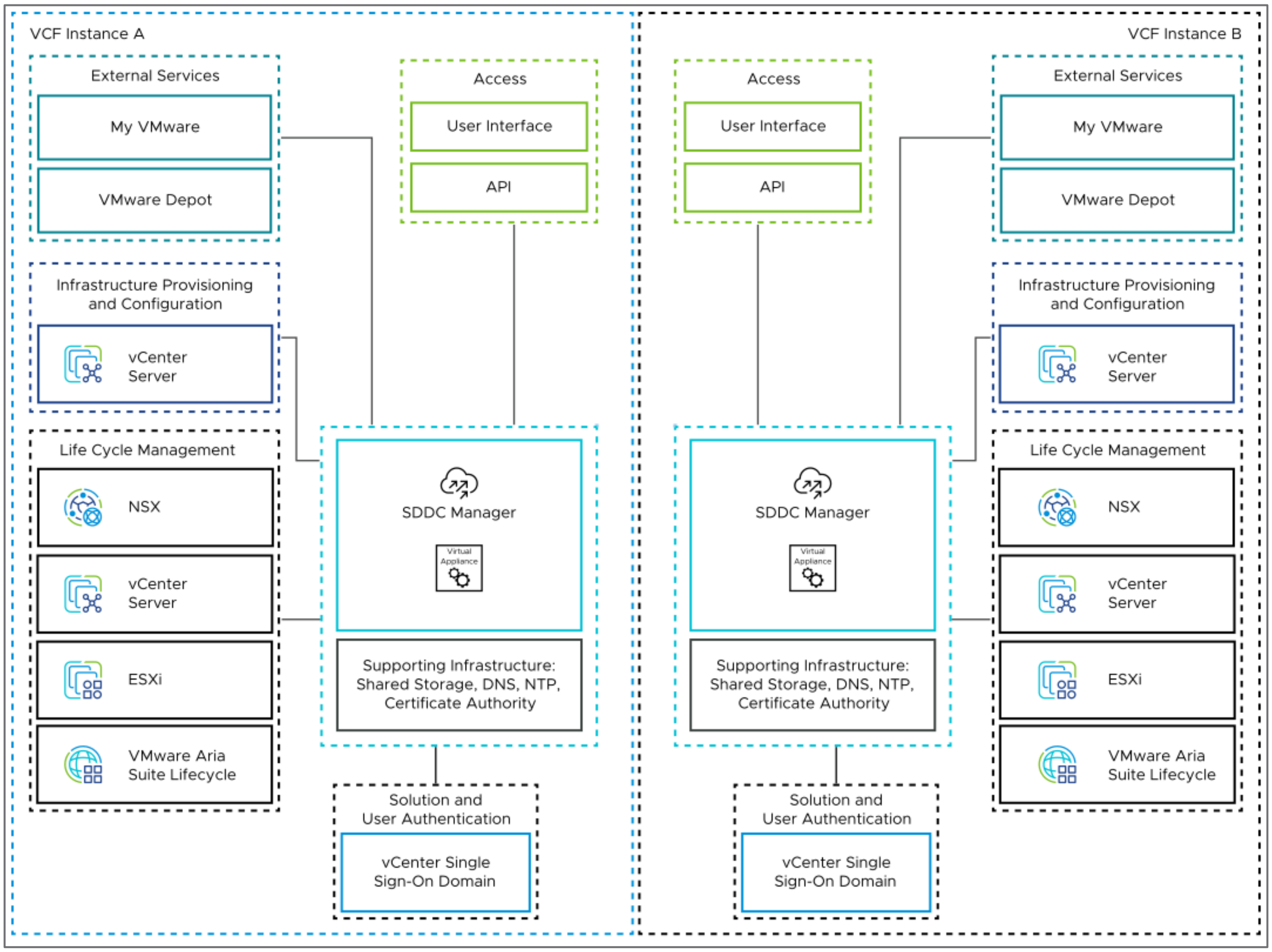

Figure 2. Logical Design of SDDC Manager in VMware Cloud Foundation

Key features and benefits of VCF:

- Integrated Stack: VCF integrates the entire VMware software-defined stack, including compute (vSphere), storage (vSAN), networking (NSX), and cloud management (vRealize Suite), ensuring interoperability and simplifying management.
- Automated Deployment and Lifecycle Management: VCF automates the deployment and lifecycle management of the Software-Defined Data Center (SDDC), reducing complexity and accelerating time to value.
- Support for Traditional and Modern Applications: VCF supports both traditional virtualized workloads and modern containerized applications, offering a unified platform for all types of workloads.
- Enhanced Security: VCF incorporates security features across the stack, including micro-segmentation, encryption, and intrusion detection, to protect applications and data.

VMware vSAN Express Storage Architecture (ESA)

VMware vSAN ESA is a new storage architecture introduced in vSAN 8. It's designed to deliver higher performance, greater efficiency, and simplified management for modern workloads and hardware [2].

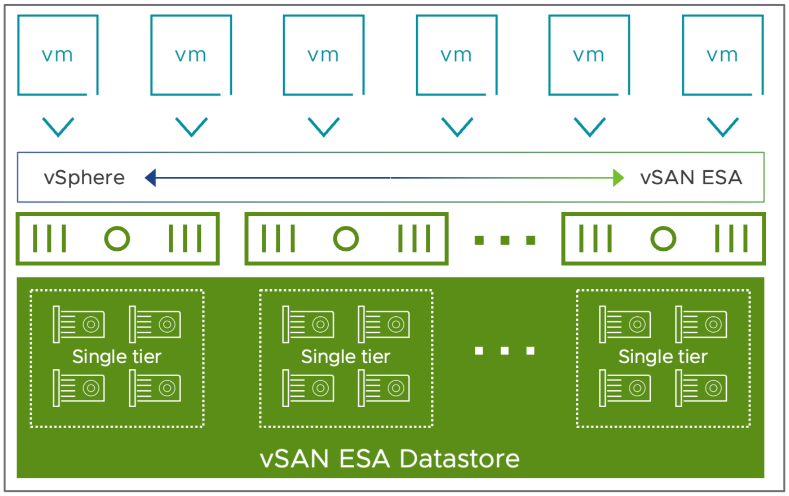

Figure 3. VMware vSAN Express Storage Architecture Overview

Key features and benefits of vSAN ESA:

- High Performance: vSAN ESA is optimized for the latest high-performance NVMe devices, delivering faster and more consistent performance than the Original Storage Architecture (OSA).
- Efficiency: vSAN ESA uses advanced space efficiency techniques to reduce storage space requirements and lower costs.
- Simplified Management: vSAN ESA simplifies storage management by eliminating the need for separate caching and capacity tiers. It also introduces a new native snapshot system that is much faster and more efficient than the previous one.
- Modern Workload Support: vSAN ESA is designed to support modern workloads, such as containers and AI/ML applications, that require high performance and low latency.
- Scalability: vSAN ESA allows for seamless scaling up or down as your storage needs change.

Samsung PM1743 NVMe SSD

The PM1743 represents an evolution in SSD technology that is driven by the fast-changing needs of enterprise customers [3]. Following are the target applications for the PM1743:

- High-end computing servers: the PM1743 is suitable for analysis functions driven by AI/ML needs, as well as core computing.
- Compute servers from Edge to Cloud: as the internet of things (IoT) and host or real-time data operations become more prevalent, there is an increase in demand for servers capable of this level of performance.
- Mixed workload services: these are made up of application servers and file servers and utilize PCIe Gen.5 speed in streaming services, as well as in application stores

vSphere 8.0 Update 2 (Build 22380479) is installed on each node of the vSAN ESA cluster.

Experimental Setup

VMware vSAN ESA Experiment Overall Architecture

Our experimental vSAN cluster environment consists of three hosts, meeting the minimum requirement for a single site vSAN HCI cluster. While supported host configurations may use as few as one Triple Level Cell (TLC) storage device, these hosts use four devices. While the Samsung NVMe PM1743 supports PCIe 5.0, the Dell R750 server limits the interface to PCIe 4.0. For the network configuration, each server is equipped with an Intel E810 network adapter providing 100 Gb/s ports. vSAN traffic utilizes a single VMkernel port configured in an active/active setup with an "Route based on physical NIC load" setting.

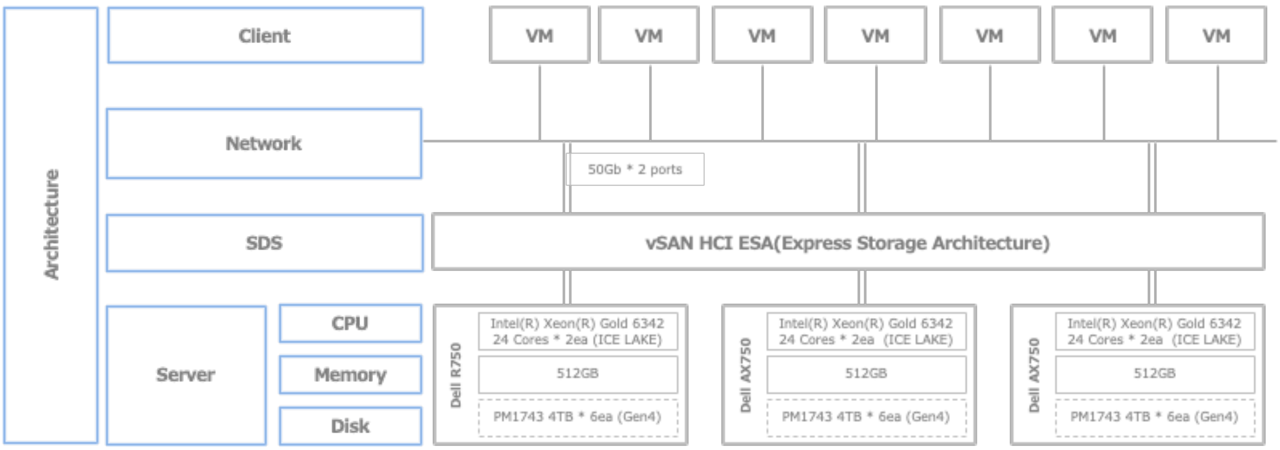

Figure 4. VMware vSAN Test Architecture

Hardware Configurations

The Dell AX750 can be considered a variant of the Dell R750 model, specifically optimized for deployment in Microsoft Azure Stack. While the core hardware specifications may be similar, the AX750 may offer additional features or configurations tailored for Azure Stack environments.

Table 1. Server Specifications

Server Model	Dell R750 / AX750
Processor	Intel(R) Xeon(R) Gold 6342 CPU @ 2.80GHz
Memory Capacity	512GB (32GB * 16)
Storage Controller	No PERC
Network and I/O	Intel(R) Ethernet 100G 2P E810-C Adapter
Drive Bays	SAMSUNG MZWLO3T8HCLS-00A07 PM1743 * 15 ea.
Power Supply and System Cooling	2400W redundant supply

Virtual Machine Configurations

To achieve an optimal resource utilization, a 1:4 CPU-to-memory allocation ratio was chosen. Each host within the cluster is equipped with two 24-core CPUs, enabling 96 virtual cores with hyperthreading. Additionally, each host possesses 16 DIMMs of 32GB RAM, resulting in a total memory capacity of 512GB per host (meaning a CPU-to-memory ratio to be approximately 1:5.33)

Table 2. Virtual Machine Specifications

vCPU	8 vCPU
Memory	32 Memory
Hard Disks	100GB Disk for OS, 200GB Disk for Data

Overview on Flexible I/O (FIO) and Workloads

The Flexible I/O Tester is a versatile benchmark tool which is widely preferred for evaluating disk performance due to its high degree of customization [4]. FIO allows users to define a comprehensive range of test parameters, including read/write ratios, access patterns (sequential/random), block sizes, queue depths, and I/O engine types (synchronous/asynchronous). This granular control enables the creation of test scenarios that closely mimic real-world workloads experienced by the storage system under evaluation. Consequently, the benchmark results allow us to infer into the specific performance needs of the target environment.

Out of the many parameters that could be considered the followings were important for our test scenarios.

Ioengine: defines how the job issues I/O to the file. Libaio is Linux native asynchronous I/O.

Group_reporting: it may sometimes be interesting to display statistics for groups of jobs as a whole instead of for each individual job. This is especially true if numjobs is used.

Direct: if value is true, use non-buffered I/O. This is usually O_DIRECT.

Time_based: if set, fio will run for the duration of the runtime specified even if the file(s) are completely read or written.

Iodepth: number of I/O units to keep in flight against the file.

Numjobs: create the specified number of clones of this job. Each clone is spawned as an independent thread or process.

Readwrite: types of I/O pattern (Based on findings from a previous VMware-published vSAN performance test, a 70/30 read/write workload ratio was selected for the mixed workload evaluation [5])

- Read: sequential reads.
- Write: sequential writes.
- Randread: random reads.
- Randwrite: random writes.
- rw,readwrite: sequential mixed reads and writes.
- Randrw: random mixed reads and writes.

VMware vSAN ESA Experiments

Number of Virtual Machines (VMs)

This test evaluates the maximum achievable performance under diverse workloads by incrementally scaling the number of virtual machines.

Various workload patterns and corresponding to FIO parameters:

- READ workload: each virtual machine issues random READ requests for individual 4 KB data blocks concurrently. A queue depth of 8 and a concurrency level of 8 threads are employed, resulting in 64 outstanding I/O requests per VM.
- WRITE workload: each VM performs random WRITE of 4 KB data blocks with a concurrency level of 8 threads. This configuration results in 64 outstanding I/O requests per VM, as each thread has queue depths of 8.

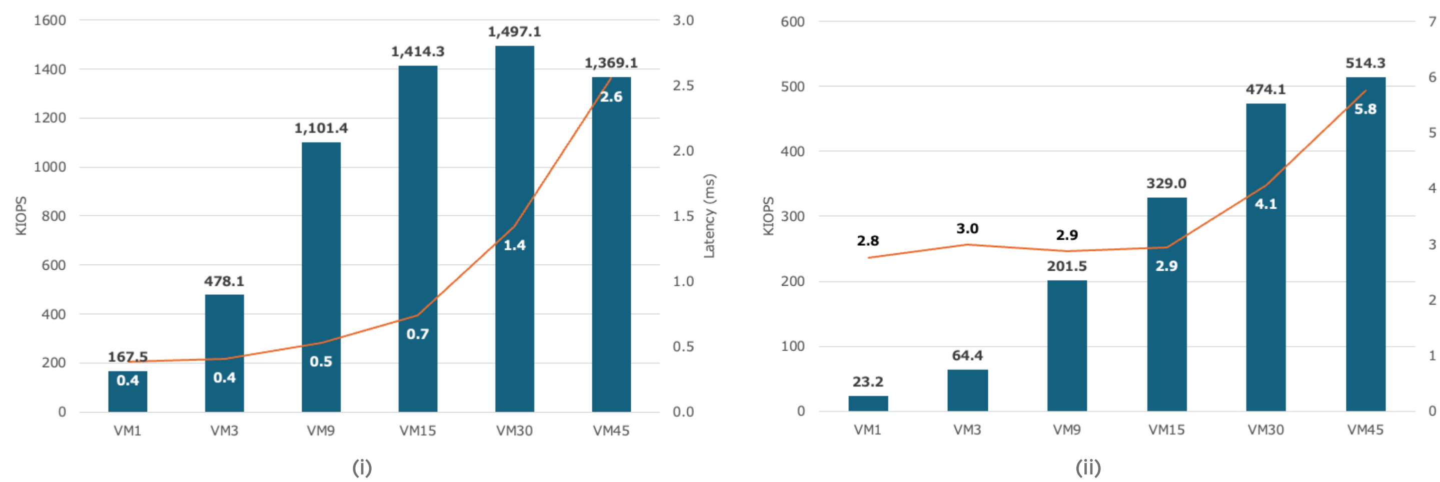

Figure 5. IOPS and latency for Random (i) READ and (ii) WRITE workloads with 4K block size

- Mixed workload: each VM performs sequential READ and WRITE operations in a 70:30 read-to-write ratio. The data block size is set to 128 KB, and the previously configured queue depth as well as thread count (64 outstanding I/O requests) are maintained.

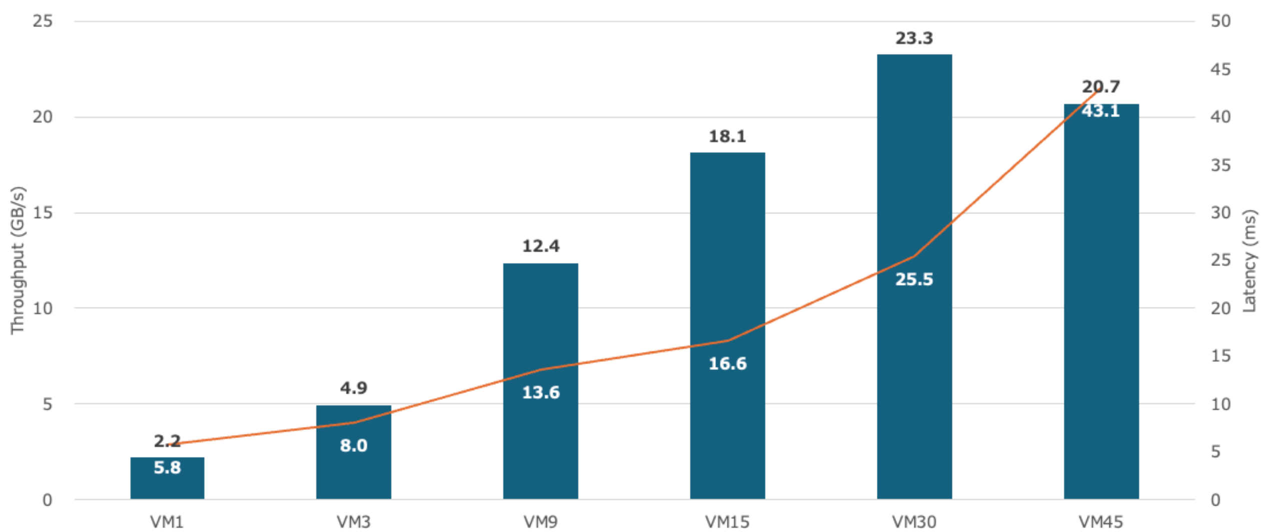

Figure 6. Throughput and latency for Mixed workloads with 128K block size

With the vSAN ESA Cluster which is consist of three hosts that have a 100Gb/s network connectivity, testing revealed that the maximum IOPS (1.49 MIOPS) and bandwidth (25.45 GB/s) were achieved with 30 concurrently running VMs. However, latency significantly increased when the workloads exceeded 15 VMs.

Although increasing the number of VMs to 45 resulted in the highest observed IOPS for random WRITE, the improvement was marginal (only 8%) as compared to 30 VMs. This suggests a peak performance range for IOPS to be between 30 and 45 VMs, likely closer to 30 VMs due to diminishing returns as further workloads increase.

The presented test results indicate that vSAN is best suited for scenarios with multiple workloads. The optimal configuration involves 640 outstanding I/O requests per host (achieved with 10 VMs per host, each issuing 64 outstanding I/O requests). Additionally, larger block sizes notably improve throughput.

Block Size

This investigation explores the optimal I/O block size for enhancing vSAN ESA performance.

The impact of block size on sequential READ performance was investigated by varying the block size from 4 KB to 1 MB. The total throughput was monitored for each block size configuration.

To ensure a consistent testing environment, FIO parameters were configured with queue depth of 8 and concurrency level of 8 threads. Three VMs were run concurrently on the cluster, resulting in 64 outstanding I/O requests per host (for a total of 192 I/O across the cluster).

I/O sizes shown on x axis reflect a doubling of payload with each size noted. Depending on the configuration, this exponential increase in payload can hit physical and logical limits of the stack, including network bandwidth, etc.

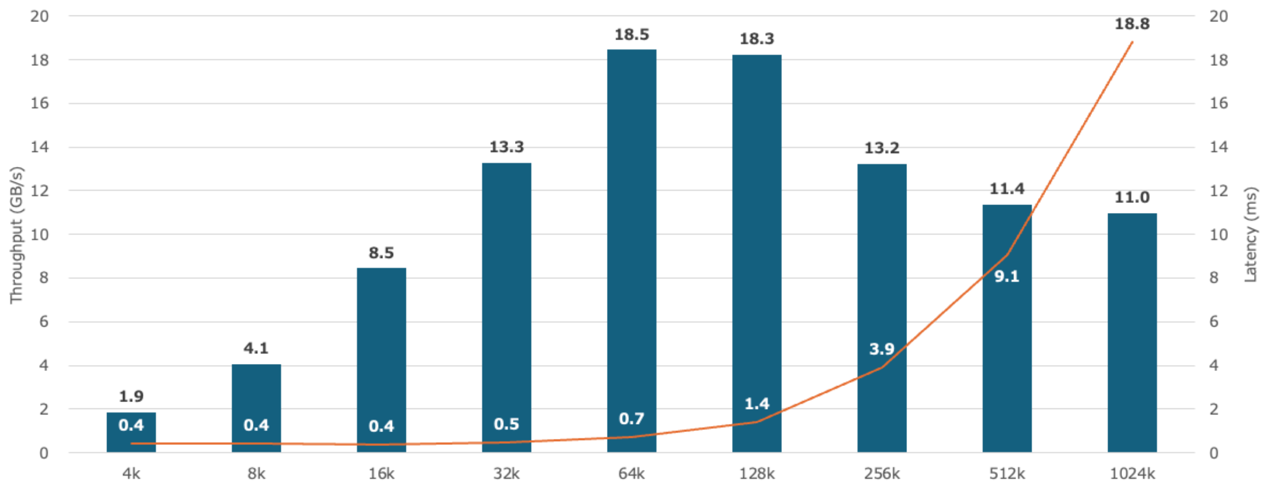

Figure 7. Throughput and latency for Sequential READ Workloads per block sizes

Throughput increased continuously up to a block size of 64 KB (18.4 GB/s), remained steady until 128 KB (18.2 GB/s), and then decreased from the 256 KB block size onwards. Similarly, latency increased significantly from a block size of 128 KB.

Stripe Width

The experiment showed that increasing stripe width within the vSAN ESA demonstrates performance improvements similar to those observed in vSAN OSA.

The storage policy allows configuration of the stripe width, which defines the number of capacity devices across which vSAN distributes each object component. This distribution can enhance I/O parallelism, potentially leading to higher performance.

A fixed configuration was used for consistency, with a 128 KB block size, a queue depth of 8, 8 threads, and 4 VMs per host, resulting in 256 outstanding I/O requests per host.

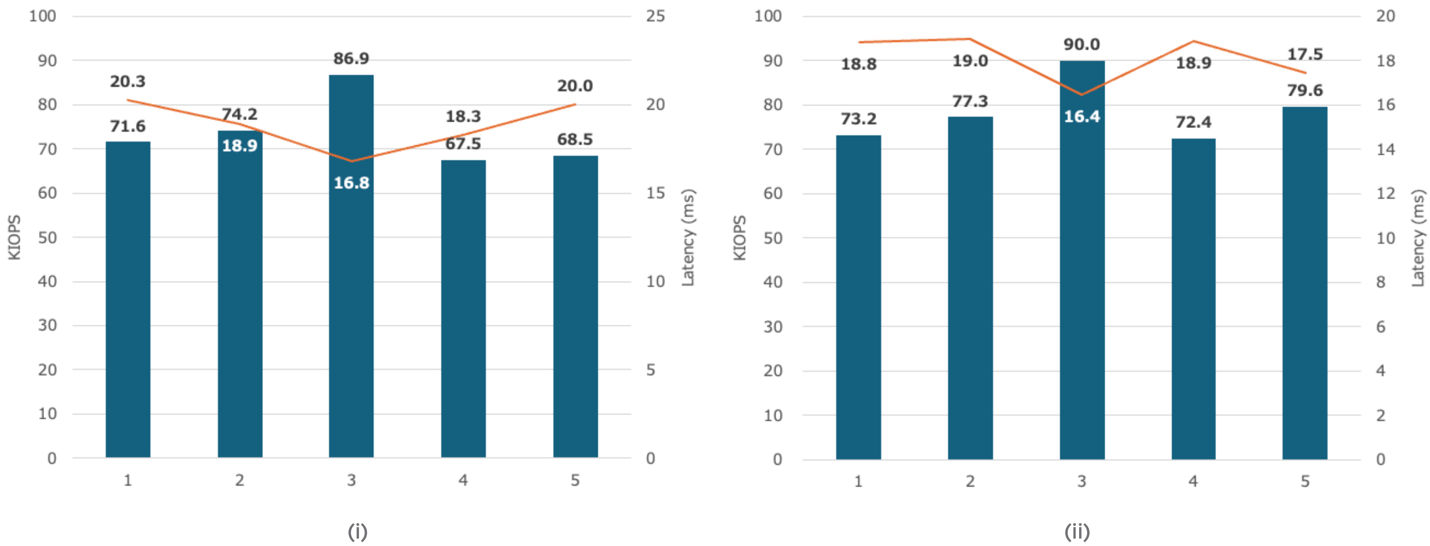

Figure 8. IOPS and latency for (i) Sequential Mixed and (ii) Random Mixed with 128K block size

Contrary to our initial hypothesis, stripe width appears to have an insignificant impact on performance in VMware vSAN ESA. This observation could likely be attributed to the architecture's optimizations for NVMe-based storage devices. The innovative Log Structured Object Manager (LSOM) in vSAN ESA efficiently handles large volumes of I/O in parallel, fully utilizing NVMe capabilities without the need for dividing data into smaller chunks for performance gains. Consequently, increasing the stripe width typically does not boost performance, allowing users to manage their VMware vSAN ESA with simplified storage policies [6].

vSAN Policies (Disk RAID)

Introduced with vSAN 8.0 ESA, VMware's adaptive RAID-5 mechanism dynamically allocates fault tolerance based on cluster size. In clusters with up to five hosts, a 2+1 RAID-5 configuration is employed when RAID-5 is selected. This is further illustrated in the following diagram. The 2+1 configuration incurs a 50% storage overhead, resulting in 150 GB of storage consumed for every 100 GB of data stored [7].

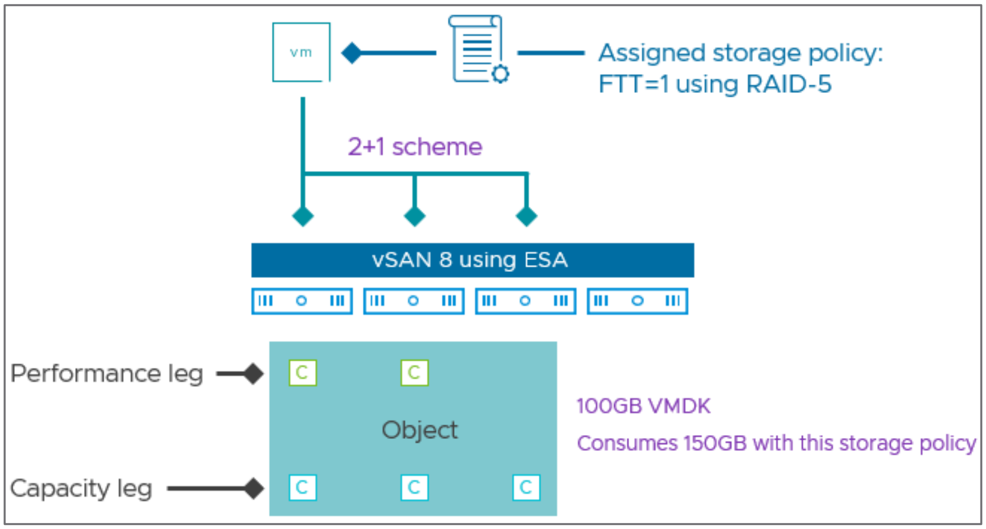

Figure 9. RAID-5 erasure coding when using the vSAN Express Storage Architecture in a small cluster

This investigation evaluates the performance impact of different vSAN storage policies (RAID 1 and RAID 5) in vSAN ESA.

Maintaining consistency with prior stages, this test leverages the same FIO configuration (64k Sequential READ workload, queue depth of 8 and concurrency level of 8 threads). However, it employs a higher VM density per host (6 VMs), resulting in 384 outstanding I/O requests per host.

All objects were configured with thick provisioning to precisely evaluate storage consumption.

The test results did not reveal any significant performance degradation in the RAID 5 configuration, even though it utilizes only half the storage capacity for data redundancy.

- The RAID 1 storage policy achieved a throughput of 17.9 GB/s while consuming 5.28 TB of primary storage and an additional 5.28 TB for replicas.
- The RAID 5 (2+1) storage policy delivered a throughput of 17.7 GB/s, utilizing 5.28 TB of primary storage and an additional 2.64 TB for data redundancy.

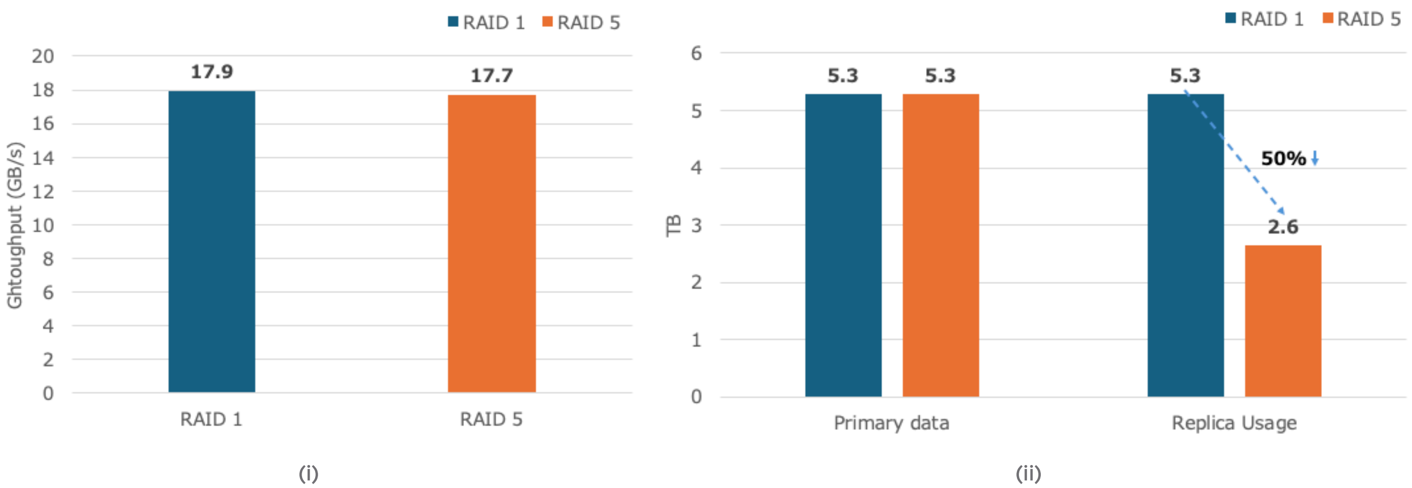

Figure 10. RAID 1 vs. RAID 5 of (i) Throughput with 64k Sequential READ and (ii) Consumed Disk Space

Number of Storage Pools

This experiment assesses how the performance of vSAN ESA is influenced by the number of disks claimed per host, by progressively increasing the number of disks in its storage pools.

The storage policies of VMware vSAN ESA Datastore are FTT=1, RAID=5, Thin as default.

To maintain consistent workloads on target VMs, FIO is configured with each VM issuing 4KB sequential WRITE and 128KB sequential mixed R/W (70:30 ratio) operations, with 8 I/O per job. Therefore, 3 VMs with 8 jobs per VM on a single host yield 192 outstanding I/O.

Test results show a slight (~5%) IOPS improvement for small block sizes with increasing number of disks in storage pool. However, no significant impact was seen on large block sizes.

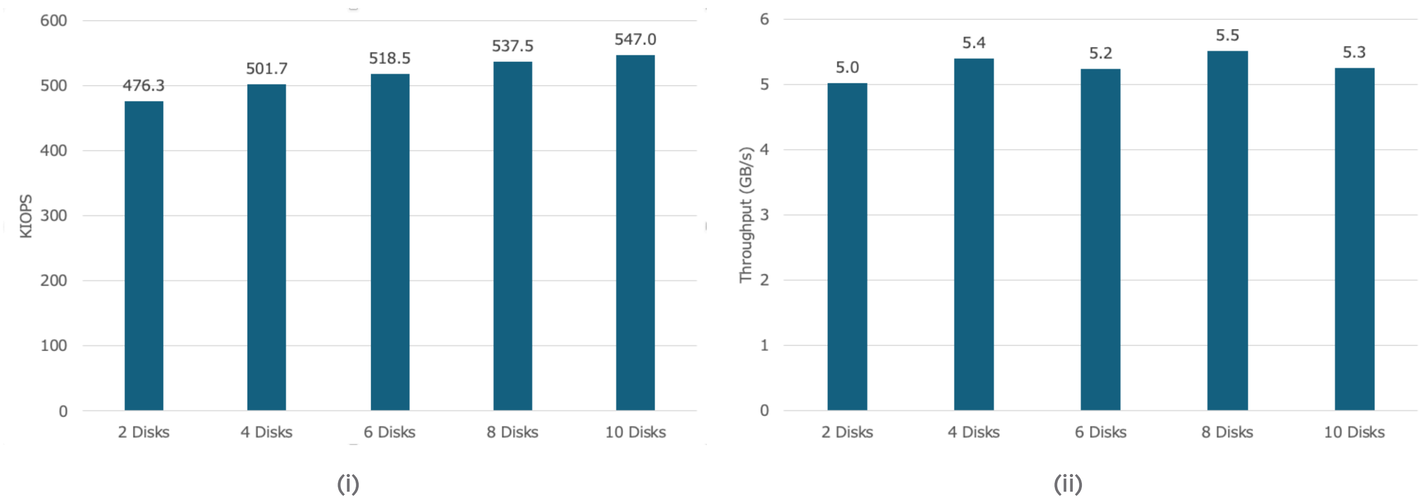

Figure 11. Number of disks for (i) Sequential WRITE IOPS with 4k block size and (ii) Sequential Mixed READ/WRITE Throughput with 128k block size

vSAN ESA with RDMA support

This test assesses the impact of RDMA on VMware vSAN ESA performance.

RDMA accelerates VMware vSAN ESA by facilitating direct memory access between computers, thereby reducing latency, boosting throughput, decreasing CPU overhead, and enhancing data integrity for optimal storage performance.

To accommodate RDMA testing requirements, the vSAN uplink network configuration was modified to an Active/Unused state to establish a lossless Ethernet environment [8].

The benchmark utilized the FIO with the following workload configurations:

- **READ workload:** five VMs per host, each running eight worker threads, issued eight concurrent random 4KB READ requests. This resulted in each host managing 320 outstanding I/O requests.
- **WRITE workload:** mirroring the READ workload, five VMs per host with eight worker threads, issued eight concurrent random 4KB WRITE requests. This also resulted in 320 outstanding I/O requests per host.
- **Mixed workload:** five VMs per host, each with eight worker threads, issued eight concurrent sequential 128KB mixed READ/WRITE requests at a 70:30 ratio. This configuration continued to maintain the 320 outstanding I/O requests per host.

The default storage policy for VMware vSAN ESA is configured with a Failure to Tolerance (FTT) of 1, RAID-5 data protection, and thin provisioning.

During this test, CPU load average was monitored using ESXTOP, a tool provided by the VMware ESXi hypervisor. CPU load average represents the average utilization of processes over 1, 5, and 15-minute intervals. In this analysis, 1-minute intervals were captured from all ESXi hosts at peak load and were averaged to provide a comprehensive system-level assessment of CPU demand.

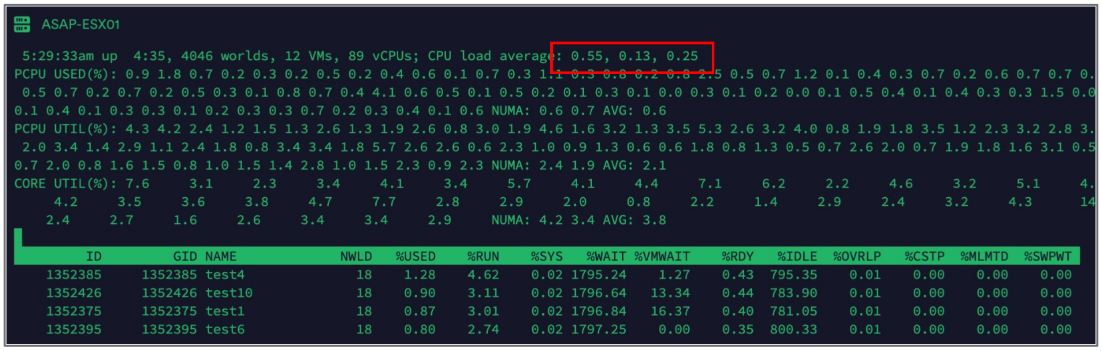

Figure 12. CPU load average monitored with ESXTOP

Considering that the vSAN ESA cluster comprises three hosts with 100 Gb/s network connectivity, The results of this test are as follows:

- **Random 4K READ**: 1,414K IOPS were achieved with TCP, while 1,441K IOPS were achieved with RDMA. Both cases exhibited an average of 99% CPU load.
- **Random 4K WRITE**: 328K IOPS were achieved with TCP, while 345K IOPS were achieved with RDMA. TCP consumed on average of 64.5% of CPU load, while RDMA consumed 54%.
- **Sequential 128K Mixed READ/WRITE**: 18.1 GB/s of throughput was achieved with TCP, while 22.9 GB/s was achieved with RDMA. TCP consumed 75% of CPU load, while RDMA consumed 57%.

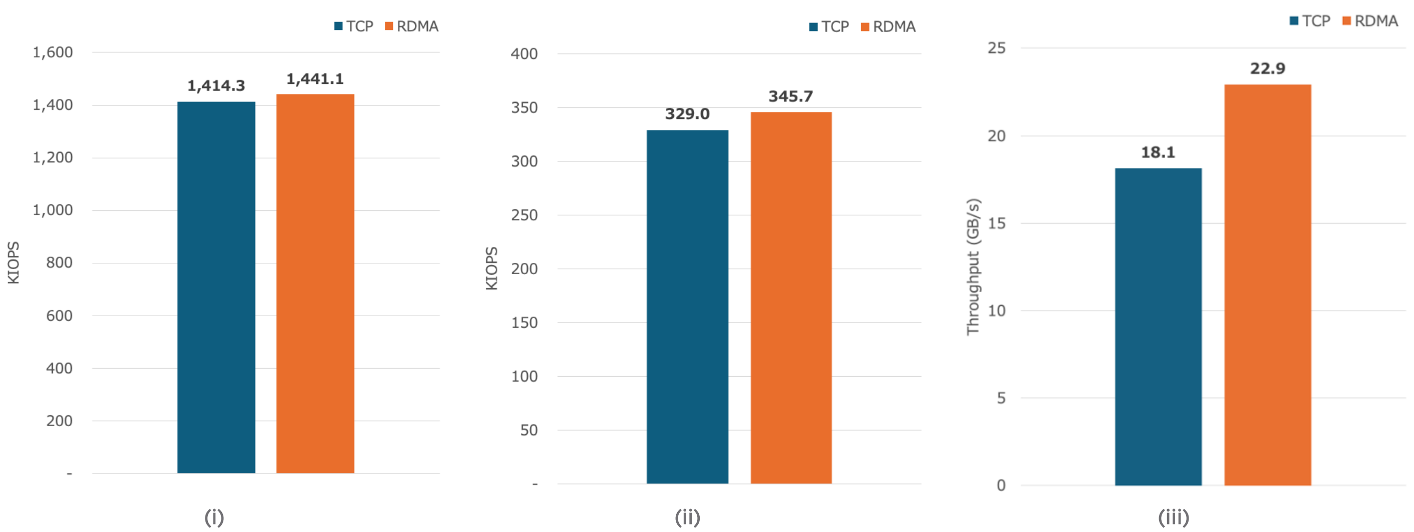

Figure 13. TCP vs. RDMA for Random (i) READ, (ii) WRITE with 4k block size and (iii) Sequential Mixed R/W with 128k block size

Table 3. Performance comparison between TCP and RDMA

4k random read					4k random write					128k sequential mixed R/W				
/	IOPS	BW (MB/s)	Latency (ms)	CPU load avg	/	IOPS	BW (MB/s)	Latency (ms)	CPU load avg	/	IOPS	BW (MB/s)	Latency (ms)	CPU load avg

TCP	1,414,282	5,934	0.74	0.99	TCP	328,966	1,379	2.94	0.65	TCP	134,984	18,127	16.6	0.75
RDMA	1,441,071	6,046	0.35	1	RDMA	345,707	1,450	1.39	0.47	RDMA	170,668	22,916	6.59	0.57

Results consistently demonstrate RDMA's superior performance over TCP across diverse workloads within the VMware vSAN ESA environment. RDMA not only significantly boosts throughput and IOPS, but it also markedly reduces CPU utilization. This CPU efficiency advantage is most pronounced in write-intensive and mixed workloads, highlighting RDMA's potential to optimize resource utilization in demanding storage environments.

vSAN ESA with JBOF

This investigation explores how Samsung Just a Bunch of Flash (JBOF) reference system effectively works with VMware vSAN ESA through NVMe over Fabrics (NVMe-oF) RDMA

Samsung JBOF reference system is a fabric-attached JBOF designed for disaggregated storage. It allows servers to seamlessly access and manage NVMe SSDs over NVMe-oF, supporting both RDMA and TCP transports. This enables servers to utilize SSDs as if they were locally attached, providing flexibility and scalability. The disaggregated storage architecture of the JBOF system decouples storage nodes from compute nodes. This allows data centers to independently scale storage resources to meet growing data demands and rapidly recover from node failures, enhancing overall system resilience and efficiency [9].

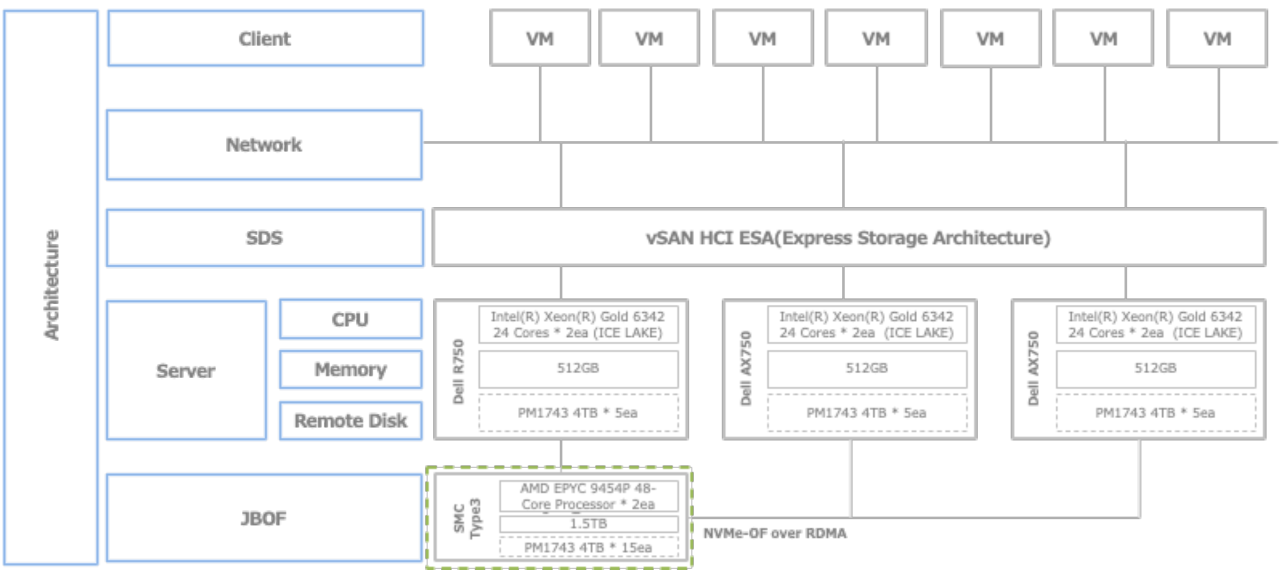

Figure 14. vSAN ESA with JBOF Test Architecture

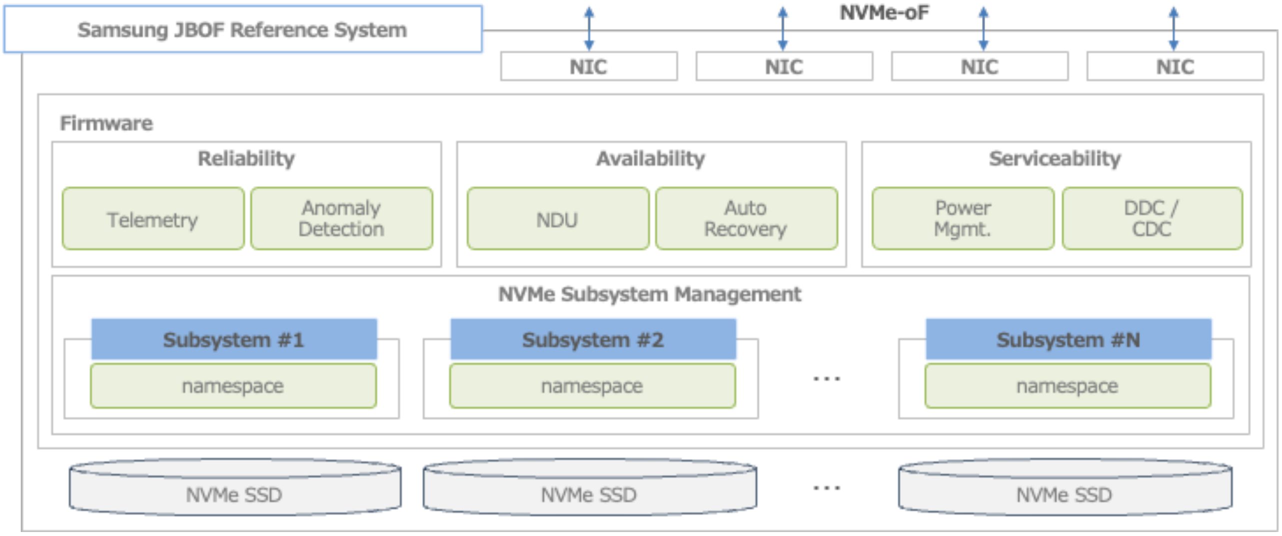

Figure. 15 Samsung JBOF reference system

The FIO parameters are utilized the same configuration of the case 'vSAN ESA with RDMA support' in order to compare between the two cases

The default storage policy for VMware vSAN ESA is configured with a FTT of 1, RAID-5 data protection and thin provisioning.

Results of this test showed the performance gap between the vSAN ESA configured with local NVMe against that of a JBOF reference system (remote NVMe) as a result, we obtained 23% for 4k Random READ, 45% for 4k Random WRITE and 58% for Mixed IO.

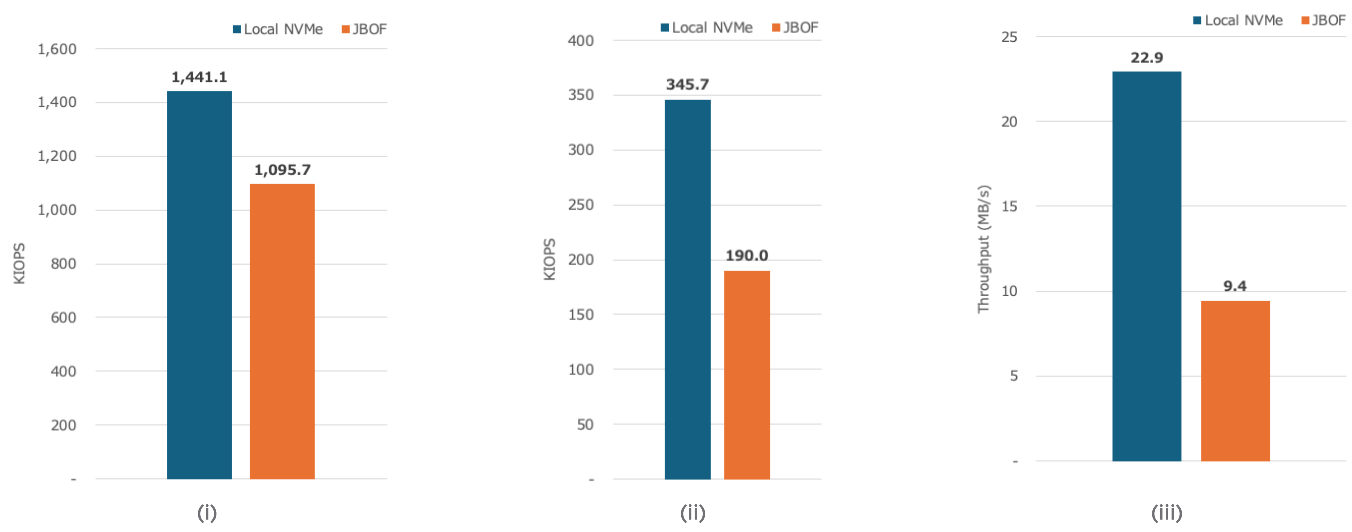

Figure 16. Local NVMe vs. JBOF for Random (i) READ & (ii) WRITE with 4k block size and (iii) Sequential Mixed R/W with 128k block size

Leveraging the estimates presented in a prior vSAN JBOF white paper [9], Figure 16 illustrate the potential performance gains associated with deploying a JBOF configuration. As the referenced research indicated, a 2x performance improvement can be anticipated with each additional JBOF server. This assumption aligns with the expectation that a 1:1 JBOF host-to-compute node ratio will deliver performance comparable to local NVMe storage, while offering superior scalability.

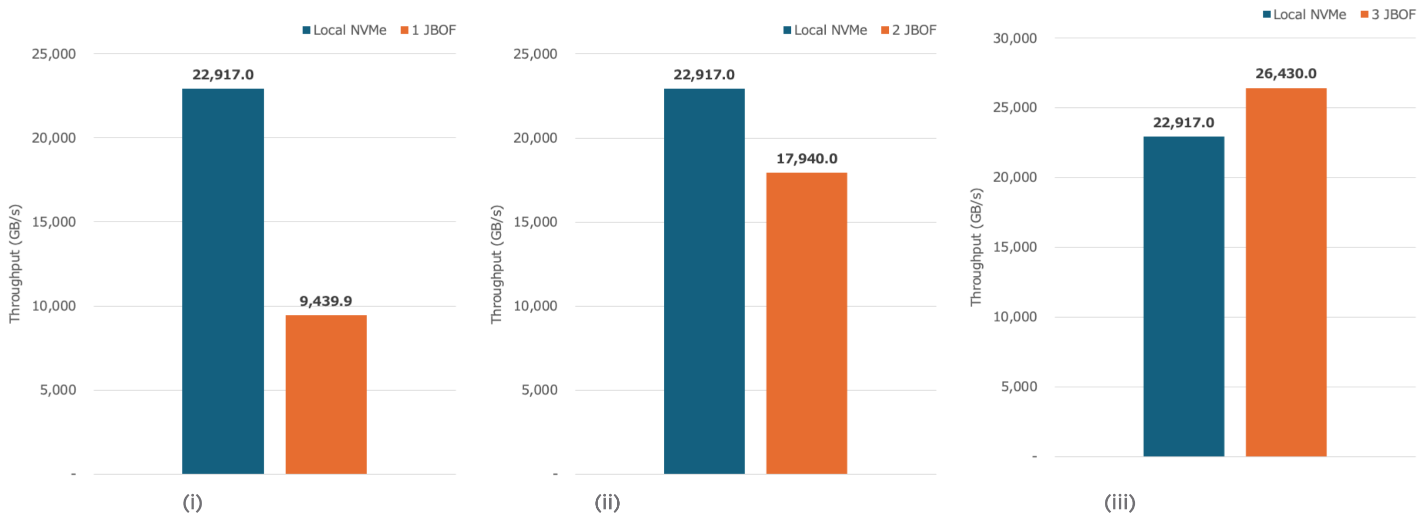

Figure 17. Estimated performance impact as the JBOF configuration scales (i) 1 JBOF server, (ii) 2 JBOF servers and (iii) 3 JBOF servers

Conclusion

In this white paper, we tested VMware vSAN ESA in a variety of configurations, conditions, and variables in order to find what factor would impact the performance of vSAN ESA.

Our findings reveal that vSAN ESA is well-suited for parallel workloads, demonstrating the ability to achieve over 1 million IOPS and 25GB/s throughput with three hosts of vSAN ESA Cluster using 100 Gb/s networking. We observed that vSAN ESA achieves peak performance when handling workloads with block sizes within the 64KB to 128KB range. Our exploration revealed how increasing the number of disks in the vSAN ESA Storage Pool can enhance system scalability and performance. With its new RAID-5 mechanism, vSAN ESA delivers a blend of high performance typically associated with RAID-1 and the storage efficiency of RAID-5. By leveraging RDMA protocols, vSAN ESA achieves significant performance gains with reduced CPU overhead. Our demonstration of the Samsung JBOF reference system as a remote NVMe SSD solution for vSAN ESA highlights its compelling potential, especially in environments requiring flexible disk expansion.

Given the numerous options and factors involved in choosing NVMe SSDs for vSAN ESA, we hope this whitepaper provides necessary insights and guidance to readers for navigating through this complexity and making informed decisions that is aligned with their specific infrastructure requirements.

References

[1] VMware, Inc. (Jun 2024) VMware Cloud Foundation ™ Accelerating Business Outcomes
https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/products/cloud-foundation/vmw-cloud-foundation-datasheet.pdf

[2] VMware, Inc. (Jan 2023) VMware vSAN 8 and vSAN Express Storage Architecture
https://core.vmware.com/api/checkuseraccess?referer=/sites/default/files/vSAN%208%20TCO%20White%20Paper.pdf

[3] Samsung Electronics Co., Ltd. (Feb 2024) Samsung SSD PM1743 Whitepaper
https://download.semiconductor.samsung.com/resources/white-paper/PM1743_SSD_Whitepaper.pdf

[4] FIO (Jul 2024) Flexible I/O Documentation
https://fio.readthedocs.io/en/latest/fio_doc.html

[5] VMware, Inc. (Jan, 2023) vSAN Proof of Concept: vSAN Performance Testing
https://core.vmware.com/resource/vsan-proof-concept-vsan-performance-testing#hcibench-tests-using-liveoptics-data

[6] Pete Koehler (Nov 2022) Stripe Width Storage Policy Rule in the vSAN ESA
https://core.vmware.com/blog/stripe-width-storage-policy-rule-vsan-esa

[7] Pete Koehler (Sep 2022) Adaptive RAID-5 Erasure Coding with the Express Storage Architecture in vSAN 8
https://core.vmware.com/blog/adaptive-raid-5-erasure-coding-express-storage-architecture-vsan-8

[8] VMware, Inc. (Jun 2023) Configuring NVMe over RDMA (RoCE v2) on ESXi
https://docs.vmware.com/en/VMware-vSphere/8.0/vsphere-storage/GUID-F4B42510-9E6D-4446-816A-5012866E0038.html

[9] Samsung Electronics Co., Ltd. Samsung JBOF reference system Whitepaper
https://download.semiconductor.samsung.com/resources/white-paper/Whitepaper_vSAN_JBOF.pdf