VMware Virtual SAN™ 6.0 Performance

Scalability and Best Practices

TECHNICAL WHITE PAPER

Table of Contents

Executive Summary

This white paper investigates how VMware Virtual SAN™ performs and scales for well understood synthetic workloads, Virtual SAN caching tier designs, and Virtual SAN configuration parameters. The goal is to provide guidelines on how to get the best performance for applications deployed on a Virtual SAN cluster.

Tests show that the Hybrid Virtual SAN Cluster performs extremely well when the working set is fully cached for random access workloads, and also for all sequential access workloads. The All-Flash Virtual SAN cluster, which performs well for random access workloads with large working sets, may be deployed in cases where the working set is too large to fit in a cache. All workloads scale linearly in both types of Virtual SAN clusters—as more hosts are added and more disk groups per host, Virtual SAN sees a corresponding increase in its ability to handle larger workloads. Virtual SAN offers an excellent way to scale up the cluster as performance requirements increase.

Introduction

VMware Virtual SAN is a distributed layer of software that runs natively as part of the VMware vSphere® hypervisor. Virtual SAN aggregates local or direct-attached storage disks in a host cluster and creates a single storage pool that is shared across all hosts of the cluster. This eliminates the need for external shared storage and simplifies storage configuration and virtual machine provisioning operations. In addition, Virtual SAN supports vSphere features that require shared storage such as high availability (HA), vMotion, and distributed resource scheduling (DRS) for failover. More information on Virtual SAN can be obtained from the *Virtual SAN 6.0 Design and Sizing Guide* [1].

This paper explores how Virtual SAN performs and scales in several dimensions and describes best practices. This paper also demonstrates the best performance results that can be achieved on a common hardware platform for well understood micro-benchmarks. It provides several guidelines on how a Virtual SAN solution should be sized, and configuration parameters set, depending on the performance requirements of a storage solution.

This paper shows how performance depends on:

- Number of virtual machines (VMs) deployed on the Virtual SAN cluster.
- I/O workload characteristics such as I/O size and outstanding I/Os.
- Virtual SAN disk group configuration, specifically the number of hard disk drives (HDDs) in a disk group.
- Stripe width, which is the number of HDDs across which a single object is striped.
- Failures to tolerate (FTT), which is the number of host failures that a Virtual SAN cluster can tolerate without data loss.
- Size of the Virtual SAN cluster, specifically the number of nodes in the cluster, and the number of disk groups in each node.

Note: Hosts in a Virtual SAN cluster are also called nodes. The terms “host” and “node” are used interchangeably in this paper.

Virtual SAN Cluster Setup

The hardware configuration of our experimental setup is as follows. Virtual SAN can be configured to use some flash drives for the caching tier and HDDs for storage (Hybrid Virtual SAN cluster) or it can be configured to use all flash drives (All-Flash Virtual SAN cluster) [2].

Hybrid Virtual SAN Hardware Configuration

Appendix A provides the detailed hardware configuration of each node in the Hybrid Virtual SAN cluster. Briefly, each node is a dual-socket Intel® Xeon® CPU E5-2670 v2 @ 2.50GHz system with 40 Hyper-Threaded (HT) cores, 256GB memory, 2 LSI MegaRAID SAS controllers hosting 1 400GB Intel S3700 SSD, and 4 900GB 10,000 RPM SAS drives per controller. Each node is configured to use an Intel 10 Gigabit Ethernet (GbE) port dedicated to Virtual SAN traffic. The 10GbE ports of all the nodes are connected to an Arista 10GbE switch. A standard Message Transfer Unit (MTU) is used, which is 1500 bytes. A 1GbE port is used for all management, access, and inter-host traffic.

All-Flash Virtual SAN Hardware Configuration

Appendix B provides the detailed hardware configuration of each node in the All-Flash Virtual SAN cluster. Briefly, each node is a dual-socket Intel Xeon CPU E5-2670 v3 @ 2.30 GHz system with 48 Hyper-Threaded (HT) cores, 256GB memory, 2x 400GB Intel P3700 PCIe SSDs, and 1 LSI MegaRAID SAS controller hosting 6x 800GB Intel S3500 SATA SSDs. Each node is configured to use a 10GbE port dedicated to Virtual SAN traffic. The 10GbE ports of all the nodes are connected to an Arista 10GbE switch. Jumbo frames (MTU=9000 bytes) is enabled on the Virtual SAN network interfaces. A 1GbE port is used for all management, access, and inter-host traffic.

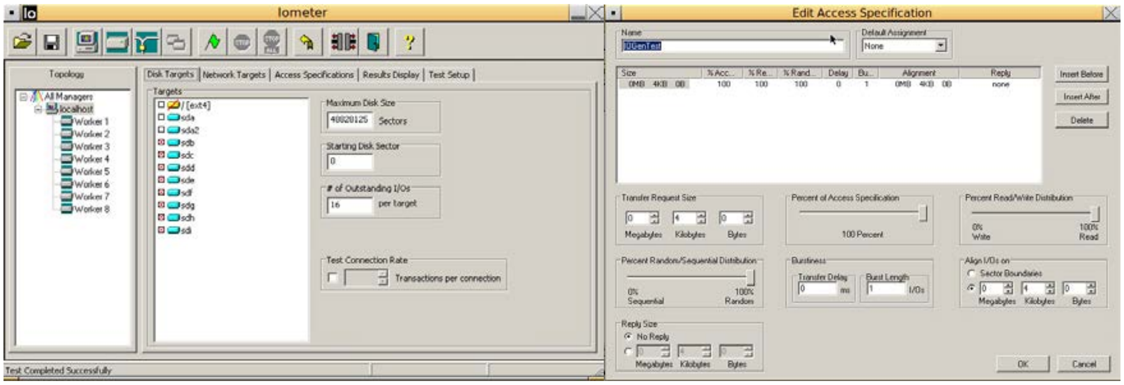
vSphere 6.0 [3] is installed on each node of the Virtual SAN cluster.

Workload

The focus of this paper is to:

- Understand the performance characteristics of a Virtual SAN cluster
- Examine the peak performance of a Virtual SAN cluster
- Emphasize where Virtual SAN performs well or poorly
- Provide best practices on Virtual SAN cluster design and configurable options required to tune the system for the best performance in specific cases.

While there are ample varieties of benchmarks for storage systems [4], a synthetic micro-benchmark, such as well understood traces generated by Iometer, is best suited to meet the above goals of this whitepaper. There are other studies that have examined the performance of commercial applications deployed on Virtual SAN, and readers who are interested in macro-benchmark studies can refer to these papers [5, 6, 7].


Iometer
Disk Targets | Network Targets | Access Specifications | Results Display | Test Setup
Edit Access Specification
Insert Before
Insert After
Delete
OK
Cancel

Topology
All Managers
localhost
Worker 1
Worker 2
Worker 3
Worker 4
Worker 5
Worker 6
Worker 7
Worker 8

Targets
□ / [ext4]
□ sda
□ sda2
☒ sdb
☒ sdc
☒ sdd
☒ sde
☒ sdf
☒ sdg
☒ sdh
☒ sdi

Maximum Disk Size: 40020125 Sectors

Starting Disk Sector: 0

# of Outstanding I/Os: 16 per target

Test Connection Rate: □ Transactions per connection

Test Completed Successfully

Name: IOGenTest

Default Assignment: None

Size	% Acc...	% Re...	% Rand...	Delay	Bu...	Alignment	Reply
0MB 4KB 0B	100	100	100	0	1	0MB 4KB 0B	none

Transfer Request Size: 0 Megabytes 4 Kilobytes 0 Bytes

Percent of Access Specification: 100 Percent

Percent Read/Write Distribution: 0% Write – 100% Read

Percent Random/Sequential Distribution: 0% Sequential – 100% Random

Burstiness: Transfer Delay 0 ms; Burst Length 1 I/Os

Align I/Os on: ○ Sector Boundaries ● 0 Megabytes 4 Kilobytes 0 Bytes

Reply Size: ● No Reply ○ 0 Megabytes 4 Kilobytes 0 Bytes

Figure 1. Iometer screenshots for All Read workload

Five different I/O workloads are considered in this study:

1. **All Read workload:** Each Iometer worker thread is configured to access random reads across the entire volume. Unless otherwise noted, the I/O size is 4KiBs (1KiB = 1024 bytes). This workload can be used to understand the maximum random read I/Os per second (IOPS) that a storage solution can deliver.
2. **Mixed Read/Write (Mixed R/W) workload:** Each Iometer worker thread is configured to do a mixed R/W access with a 70%/30% ratio. All accesses are random. Unless otherwise noted, the I/O size is 4KiB. Most applications use a mix of reads and writes; therefore, this trace comes closest to representing the performance that can be expected from a commercial application deployed in a Virtual SAN cluster.
3. **Sequential Read workload:** Each Iometer worker thread is configured to do sequential read access with 256KiB I/O size. This trace is representative of scanned read operations, such as reading a video stream from a storage solution.
4. **Sequential Write workload:** Each Iometer worker thread is configured to do sequential write access with 256KiB I/O size. This trace is representative of scanned write operations, such as copying bulk data to a storage solution.
5. **Sequential Mixed R/W workload**: Each Iometer worker thread is configured to do sequential R/W access in a 70%/30% ratio with 256KiB I/O size.

Figure 1 shows Iometer screen snapshots for the All Read workload experiment.

Virtual Machine Configuration

Iometer version 2006.0727 is deployed on an Ubuntu 12.0 VM. This package is also available as an I/O Analyzer fling from VMware Labs [8]. Each VM is configured with 4 vCPUs and 4GB of memory. The VM is configured with three PVSCSI controllers: one for the OS disk and the other two equally share the data disks. The queue depth for the PVSCSI controller is also increased to 254 as specified in VMware KB 1038578 [9]. Eight eager-zeroed-thick VMDKs per disk group per VM are created. Each VMDK is available to the VM as a block mount. All the experiments are conducted after the VMDK is written to at least once. This is to prevent zeroed returns on reads. An Iometer I/O worker thread is configured to handle each VMDK independently.

Metrics

In the All Read and Mixed R/W experiments, there are two important metrics to follow: I/Os per second (IOPS) and the mean latency encountered by each I/O operation. In each VM, the IOPS are measured and latency is observed for each I/O request for a 60-minute steady-state duration on each running Iometer instance. The mean of these values is added across each node in the cluster, to get cumulative cluster-wide IOPS. Similarly, the mean latency is calculated across each node in the cluster, to achieve the cluster-wide latency. The latency standard deviation is also noted, which gives us confidence on our mean latency measure. These cluster-wide metrics of cumulative IOPS and latency help understand and characterize the Virtual SAN performance. For the Sequential read and write workloads, we are interested in monitoring the read and write bandwidth achieved in the Virtual SAN cluster.

Virtual SAN Configuration Parameters

Several Virtual SAN configuration parameters are varied in the experiments. Unless otherwise specified in the experiment, the Virtual SAN cluster is designed with the following configuration parameters:

- Disk group with 1 Intel S3700 SSD and 4 HDDs for the Hybrid Virtual SAN cluster, and 1 Intel P3700 PCI-E SSD and 3 S3500 SSDs for the All-Flash Virtual SAN cluster
- Stripe width of 1
- FTT of 1
- Default cache policies are used and no cache reservation is set.

Hybrid Virtual SAN Cluster Experiments

Our experiments start with a 4-node Virtual SAN cluster in which each host is configured with a single disk group created with 1 SSD and 4 HDDs connected to the same controller.

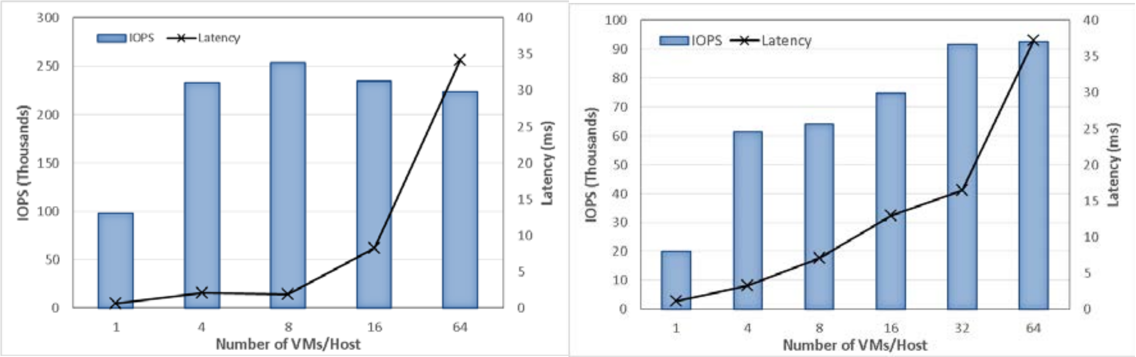
Number of VMs

The cluster is scaled up from 1 VM to 64 VMs in the cluster. As mentioned previously, each VM accesses 8 VMDKs created on the Virtual SAN. VMs are placed to ensure uniform distribution across the hosts. For example, in the case of 8 VMs, 2 VMs are deployed on each cluster node. A constant working set of 200GiB per host is ensured (1GiB = 2^{30} bytes). The working set size is chosen to ensure that Virtual SAN can completely cache it. Note that the VM volumes can be located anywhere in the cluster as assigned by the Virtual SAN Cluster Level Object Manager (CLOM). The remaining parameters of the experiment are as follows:

- All Read workload: Each VM issues 4KiB sized reads, with 16 I/Os per I/O worker thread. Thus, every VM with 8 worker threads issues 128 outstanding I/Os.
- Mixed R/W workload: Each VM issues 4KiB sized reads and writes, with 4 I/Os per I/O worker thread. Thus, every VM has 8 worker threads with 32 outstanding I/Os. A lower outstanding I/O is chosen for the mixed R/W workload because the write latency increases significantly when I/Os are queued.

Figure 2 shows the plot of IOPS and latency as the virtual machines are scaled up from 1 to 64. IOPS is the cumulative IOPS in the Virtual SAN cluster, while latency refers to the mean latency across all VMs reported by I/O Analyzer. For the All-Read workload, the peak IOPS (60 thousand per host) can be reached with 8 VMs (2 VMs per host). As the VMs are increased to 64, IOPS remain flat (the change seen in the graph is within the noise margin). The steady IOPS as the VMs are scaled from 8 to 64 shows that Virtual SAN does not exhibit any performance penalty at higher consolidation ratios and with higher numbers of VMDKs. The increase in latency is expected due to queuing delays for higher outstanding I/O requests from the larger number of deployed VMs. For the Mixed R/W workload, peak IOPS (22 thousand per host) is reached with 32 VMs (8 VMs per host). Note that in these experiments, only one disk group per node is used. Performance could scale further if there were more disk groups per node.

The remainder of experiments described in this white paper use 1 VM per node (4 VMs across the Virtual SAN Cluster), each accessing 8 VMDKs.



(i) (ii)

Figure 2. IOPS and Latency vs. number of VMs for (i) All Read and (ii) Mixed R/W workload

Outstanding I/Os

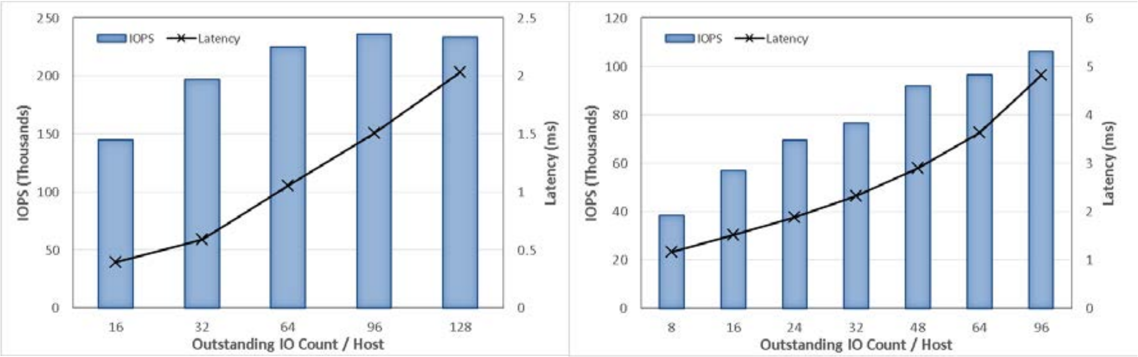
On a single VM, the outstanding I/O is scaled up from 1 to 16 per I/O worker thread (16 to 128 outstanding I/Os per VM) for the All Read workload, and from 1 to 12 per I/O worker thread (8 to 96 outstanding I/Os per VM) for the Mixed R/W workload. The I/O access size is kept constant at 4KiB. The working set size is maintained at 200GiB per VM. The results are shown in Figure 3. As expected, lower outstanding I/O shows lower latency. Latency increases with higher outstanding I/Os because of queuing delays, while IOPS increases because more throughput can be achieved by the caching tier. The cumulative IOPS scales to 240 thousand (60 thousand per node) for the All Read workload, and 106 thousand (26 thousand per node) for the Mixed R/W workload. Note that these numbers are again per disk group. In the remaining experiments of this white paper, 128 outstanding I/Os per VM per disk group are considered for the All Read workload, and 32 outstanding I/Os per VM per disk group for the Mixed R/W workload.

***Best Practice:** It is evident that Virtual SAN can deliver 1 millisecond latencies at the application tier, as long as it is not overloaded with outstanding I/Os. All Read workloads showed one-third the latency compared to a Mixed R/W workload. Increasing outstanding I/Os helped scale from 140 thousand to 240 thousand (1.7 times) for the All Read workload, and from 39 thousand to 106 thousand (2.71 times) for the Mixed R/W workload for the 4-node cluster using 1 disk group. Latency increases with an increase in outstanding I/Os due to queuing delays.*

I/O Size

In this experiment, we vary the I/O size from 4KiB to 256KiB. The number of outstanding I/Os is 128 per VM for the All Read workload, and 32 per VM for the Mixed R/W workload. The working set size is maintained at 200GiB per VM. At an I/O size of 4KiB, the workload access pattern is random. At the other end of the spectrum, a 256KiB I/O size offers a mixed sequential random workload. Figure 4 shows the cluster-wide bandwidth as the I/O size is increased. For the All Read workload, the bandwidth increases until an I/O Size of 32KiB, and then remains steady. With an I/O size of 32KiB, 450MB per second average read bandwidth is achieved per host, which is near the 500MB per second sustained sequential read bandwidth available from the Intel S3700 SSD [11]. Therefore, increasing the I/O size does not drive additional bandwidth on the Hybrid Virtual SAN Cluster. For the Mixed R/W workload, a ramp up in bandwidth is observed until the message size of 32KiB, and then a gradual increase beyond that. In this case, a more sequential workload reduces the number of times the Virtual SAN needs to de-stage and move data from the caching tier to the HDDs. At an I/O size of 256KiB, 475MB per second bandwidth per disk group is achieved for the All Read workload, and 140MB per second bandwidth per disk group is achieved for the Mixed R/W workload. Note that a 64 times increase in I/O size leads to a 2 times increase in bandwidth in the All Read workload, and 2.5 times in the Mixed R/W workload. This shows that the Virtual SAN cluster caching tier effectively masks the poor performance of the SAS HDDs for small message random access workloads. On the other hand, the caching tier can also limit the maximum sequential read I/O bandwidth. If higher sequential read bandwidth is desired per disk group, the All-Flash Virtual SAN Cluster would be a better choice.

***Best Practice:** The message from this experiment is that a reasonable I/O size is sufficient to get great bandwidth from a Virtual SAN cluster. Our experiments show that an I/O size of 32KiB is sufficient to get good bandwidth.*



(i) (ii)

Figure 3. IOPS and latency vs. outstanding I/O per node for (i) All Read and (ii) Mixed R/W workload

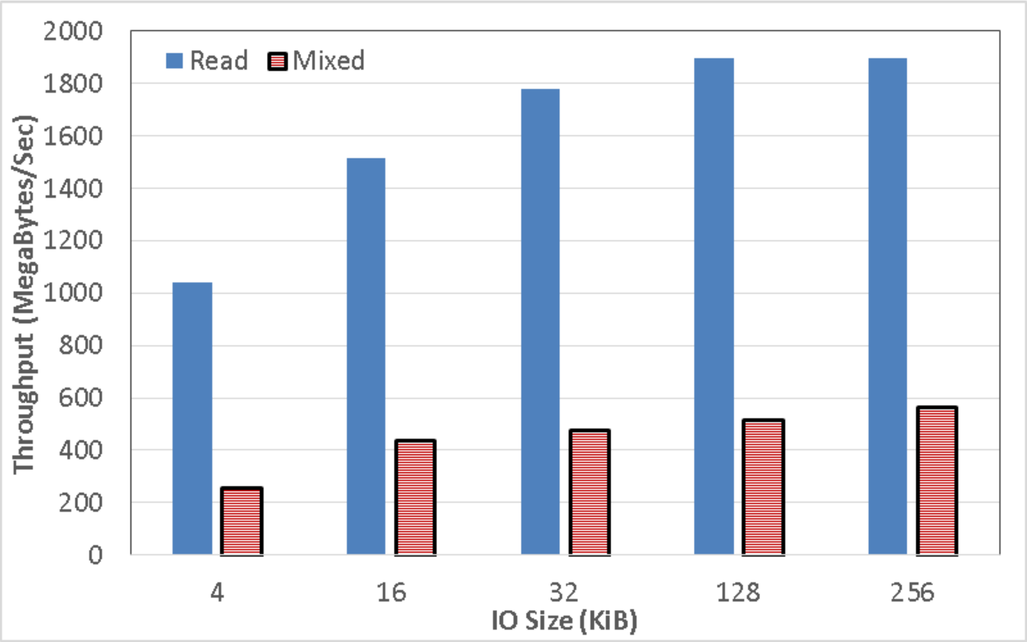


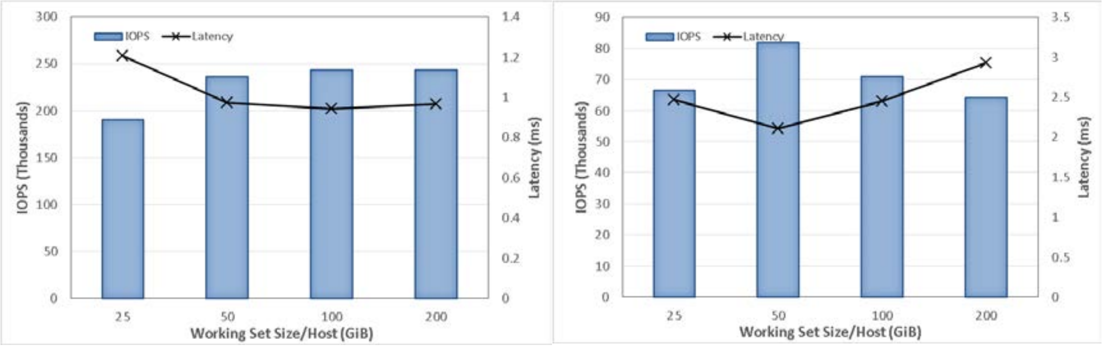
Figure 4. Cluster-wide I/O bandwidth vs. I/O size for All Read and Mixed R/W workload

Working Set Size

The working set is the portion of data storage that is regularly accessed. The previous experiments had a working set of 200GiB per host, per disk group. In this experiment, the working set is varied from 25GiB to 200GiB per host for experiments with random I/O access by increasing the number of VMDKs from 1 to 8 (each VMDK is 25GiB). The working set is limited to a 200GiB for random access workloads because that is roughly half the size of the Intel S3700 caching tier (400GB). Note that Virtual SAN splits the available cache in a 70% read and 30% write region. A 200GiB working set can therefore be fully cached in the read cache region. For the experiments with sequential data access, the working set is varied from 200GiB to 800GiB per host, while keeping 8 VMDKs per VM.

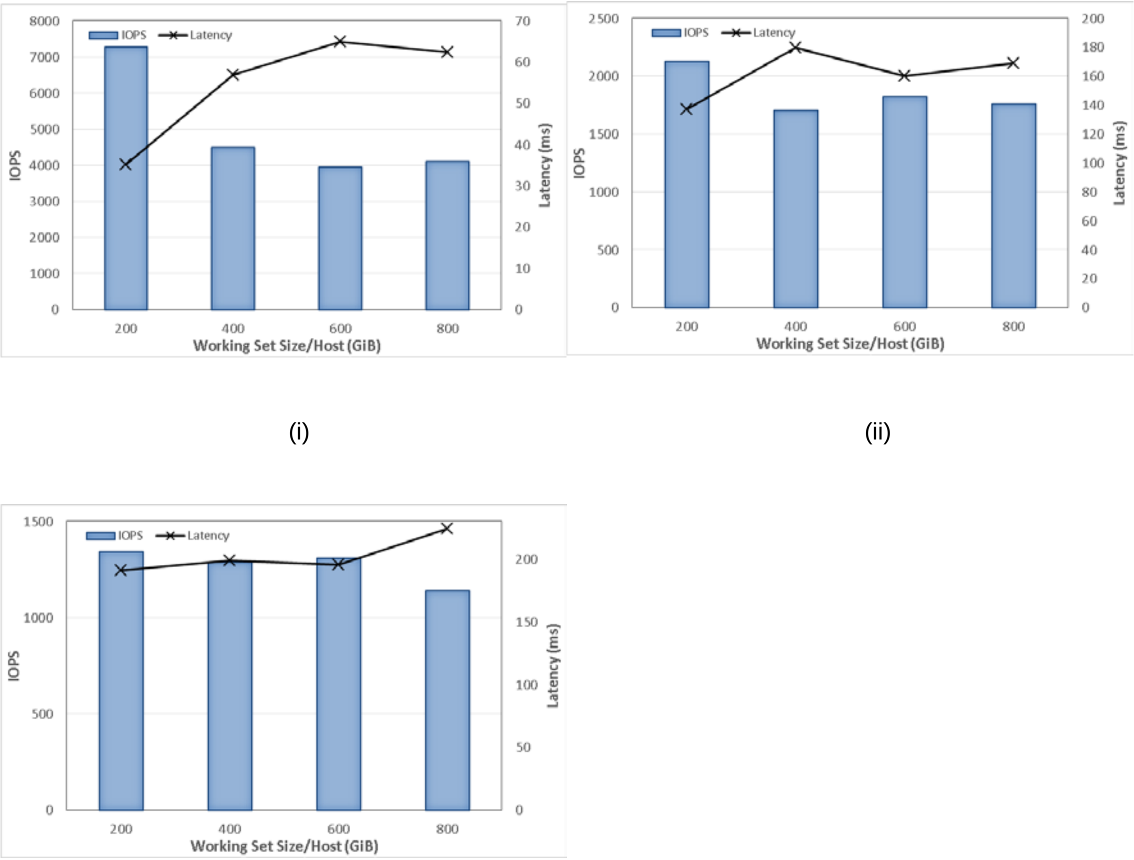
For random I/O access experiments, the outstanding I/O is kept constant at 64 outstanding I/Os per VM for the All Read workload, and 32 outstanding I/Os per VM for the Mixed R/W workload. The results are presented in Figure 5. It is observed that the IOPS and latency are steady for the All Read workload across different working set sizes (except for the case where a single VMDK limits scalability and achieves lower IOPS). For the Mixed R/W workload, a 20% drop is observed in IOPS, and a similar increase in latency is seen, when the working set is increased from 50GiB to 200GiB. A larger working set increases the frequency at which the Virtual SAN needs to de-stage written content from the caching tier to the HDDs. Any de-staging activity causes simultaneous read/write activity on the caching tier, which leads to a decrease in performance as illustrated in Figure 5 (ii).

For sequential access experiments, the outstanding I/O is kept constant at 64 outstanding I/Os per VM while the I/O access size is 256KiB. Figure 6 shows the IOPS and latency at different working set sizes for the sequential read, write, and mixed R/W workloads. Although the working set is no longer fully cached, sequential I/O access benefits from the large sequential I/O bandwidth of the HDDs. For the All Read and Mixed R/W workloads, an IOPS drop of 37.5% and 22.5% respectively is observed when the working set increases from 200GiB to 400GiB. This is mainly because the working set of the 400GiB size does not completely fit in the read cache any more. Steady performance is observed when the working set is increased from 400GiB to 800GiB.



(i) (ii)

Figure 5. IOPS and latency with working set size for (i) All Read and (ii) Mixed R/W workload



(i)

(ii)

(iii)

Figure 6. IOPS and latency vs. working set size for sequential (i) Read, (ii) Mixed R/W, and (iii) Write workload

Having a smaller caching tier size than the active working set may be acceptable for sequential I/O workloads; however, it is not desirable for completely random I/O access workloads. With complete random I/O access, a significant chunk of I/Os could be queued up on the HDDs, slowing down the performance of the system. Figure 7 illustrates this point. The performance of the All Read workload drops when the working set size is larger than 250GiB. Please note that for the Intel 400GB SSD, the read cache would be provisioned for 70% of 400GB (260.7GiB). To sum up, if the application environment has completely random I/O access, you must design the hybrid Virtual SAN cluster with a caching tier that is 2 times larger than the working set size, or alternatively consider the All-Flash Virtual SAN cluster, which can deliver large random IOPS at large working sets.

Best Practice: *Sizing of the caching tier is an important consideration during Virtual SAN cluster design. It is important to have sufficient space in the caching tier to accommodate the I/O access pattern of the application. In general, for each host, the caching tier should be at least 10% of the total storage capacity. However, in cases where high performance is required for mostly random I/O access patterns, it is recommended that the SSD size be at least 2 times the working set. For sequential I/O access workloads, the sequential I/O bandwidth of HDDs is an important concern.*

Disk Group Configuration

In all previous experiments, each node had a single disk group with 1 SSD and 4 HDDs. In this experiment, the number of HDDs in the disk group is varied from 2 to the maximum value of 7. All Read workloads do read accesses from the caching tier, and therefore the performance of the All Read experiment is not affected by the number of HDDs. Figure 8 presents the results for the Mixed R/W experiment. The working set size is 200GiB per disk group, I/O size is 4KiB, and the number of outstanding I/Os is 32 per VM. It is observed that the IOPS reach 62 thousand with 4 HDDs and then marginally increases to 64 thousand with 7 HDDs. The latency gets better with a similar trend. The small change from 4 to 7 HDDs suggests that having a large number of HDDs may have very incremental benefit for random I/O accesses on cached working sets. Another way to get a performance benefit from a higher number of HDDs is to increase the stripe width. The performance impact of this is discussed in the next section.

Figure 9 presents the results for sequential access workloads. For all three access patterns, much higher IOPS and lower latency is observed with a higher number of HDDs. This increase is mainly due to a higher parallelism of non-cached HDD accesses offered by the higher number of HDDs.

Please note that some I/O controllers may get bottlenecked for I/O bandwidth or transactions per second when a large number of HDDs are installed on them. This may particularly affect the performance of the caching tier if the SSD is placed on the same controller as the HDDs. In such situations where a large number of HDDs may cause an I/O controller to become bandwidth bottlenecked, it is recommended that a PCI-E flash SSD be installed, instead of an on-controller SSD, to maintain good performance from the caching tier.

Best Practice: *Keep in mind the working set when designing the disk group. With a cached working set, having a number of HDDs beyond 4 does not seem to give added benefit. It is recommended that you use a higher stripe width when more than 4 HDDs are present in a disk group. For random workloads that are not expected to fit in cache or for sequential workloads, a larger number of HDDs will give better performance.*

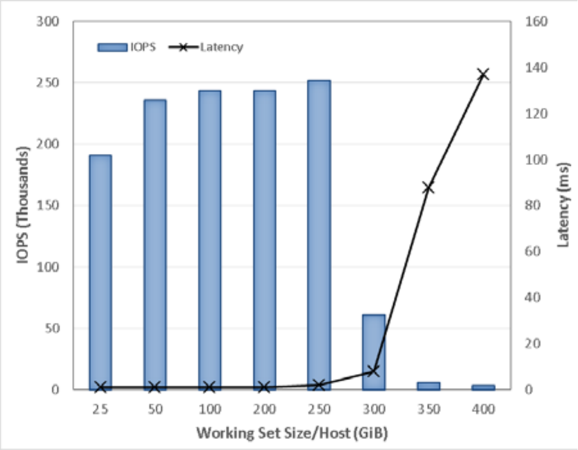


Figure 7. IOPS and latency vs. working set sizes larger than the cache size for an All Read workload

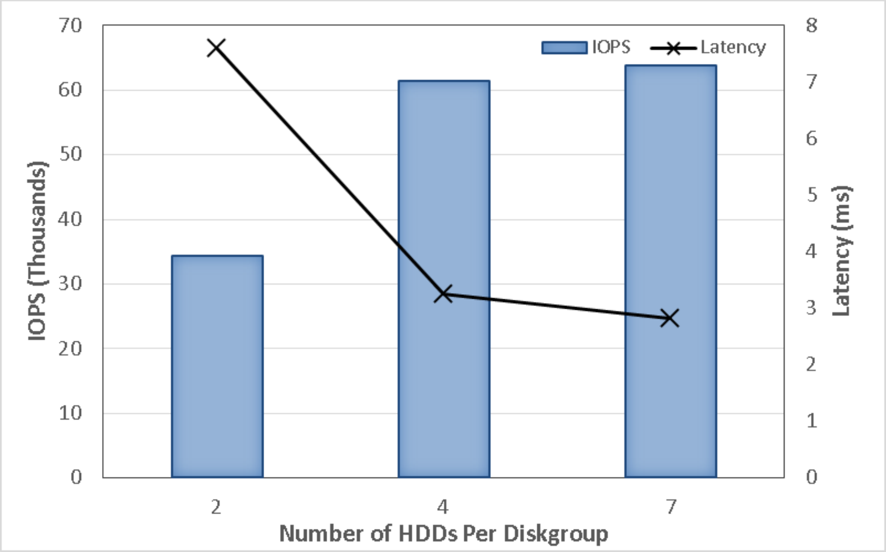
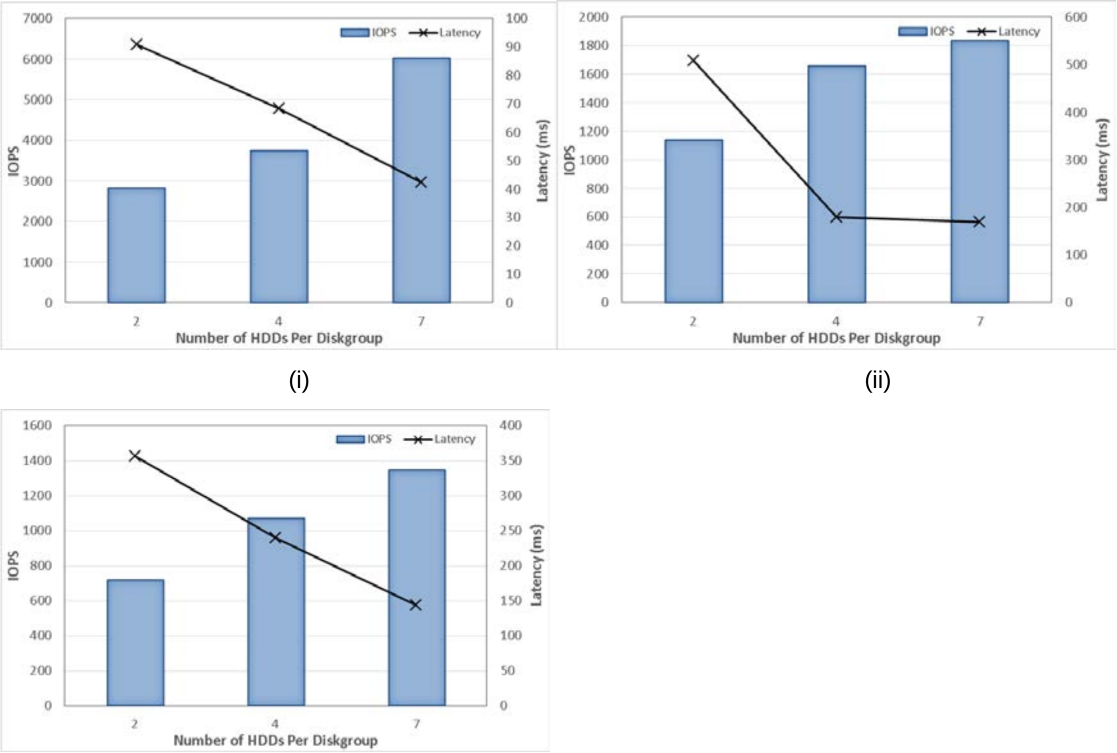


Figure 8. Variation of IOPS and latency with disk group configuration for Mixed R/W workload

Stripe Width

All experiments until now used a stripe width of 1 as a policy while creating the VMDKs. With a higher stripe width setting, each VMDK is striped across multiple HDDs. Figure 10 shows the benefit of a higher stripe width. The Mixed R/W workload is similar to what was used in the previous experiment; 200GiB working set, 4KiB message size, and 32 outstanding I/Os per VM with 1 VM per node A higher stripe width helps lower latency because the random I/O accesses to a VM volume get striped across multiple HDDs. A higher stripe width is unlikely to help sequential access workloads.

Best Practice: *A higher stripe width setting may be used with applications with random I/O access patterns in situations where: (i) latency is an important requirement, (ii) there are very few VMDKs, (iii) there are more than 4 HDDs in the disk group, or (iv) there is a 2x or higher difference in the IOPS or latency from different HDDs in a disk group. Performance metrics may be monitored using Virtual SAN observer.* [12]



(i)

(ii)

(iii)

Figure 9. IOPS and latency with disk group configuration for sequential (i) Read, (ii) Mixed R/W, and (iii) Write workload

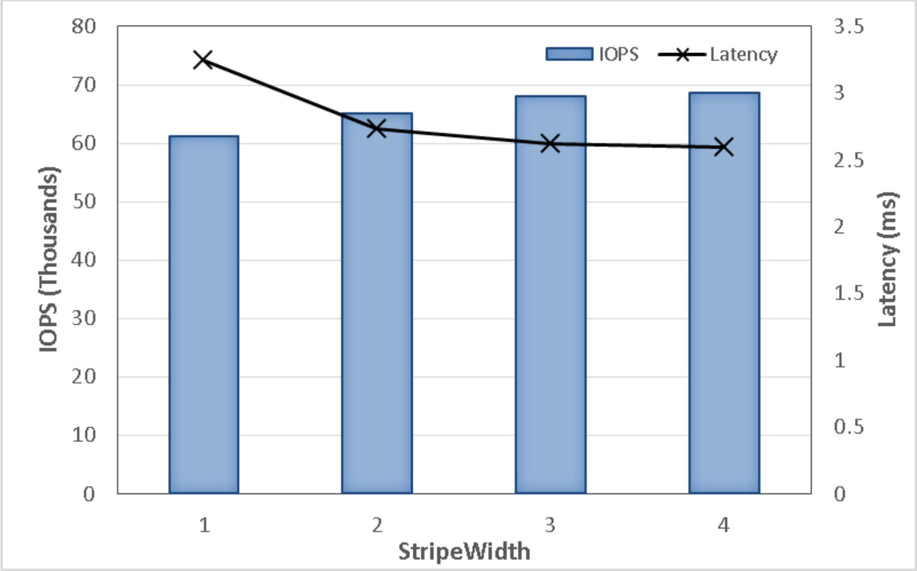


Figure 10. IOPS and latency with stripe width for Mixed R/W workload

Failures To Tolerate (FTT)

The Failures To Tolerate (FTT) feature allows you to set redundancy in a Virtual SAN cluster. By default, FTT is set to 1, which implies that the cluster is designed to tolerate a 1-node failure without any data loss. A higher level of redundancy can be set to protect from multiple nodes failing concurrently, thereby ensuring a higher degree of cluster reliability and uptime. However, this comes at the expense of maintaining multiple copies of data and thereby impacts the number of writes needed to complete one transaction. Moreover, a larger cluster size is required for a higher degree of redundancy; the minimum number of cluster nodes required is *2 × FTT + 1*. Figure 11 demonstrates the impact of higher levels of FTT on the Mixed R/W workload on an 8-node Virtual SAN cluster. Note that an 8-node cluster was chosen because 7 is the minimum number of nodes to support FTT=3. The Mixed R/W workload is similar to what was used in the previous two experiments: 1 disk group per node, 200GiB working set, 4KiB message size, and 32 outstanding I/Os per VM, 1 VM per node, and a stripe width of 1. Note that the performance of the All Read workload will not be affected by a higher FTT setting because all of the I/Os are served from the caching tier.

***Best Practice:** FTT may be set higher than 1 for higher levels of redundancy in the system and for surviving multiple host failures. Higher FTT impacts performance because multiple copies of data must be ensured on write operations.*

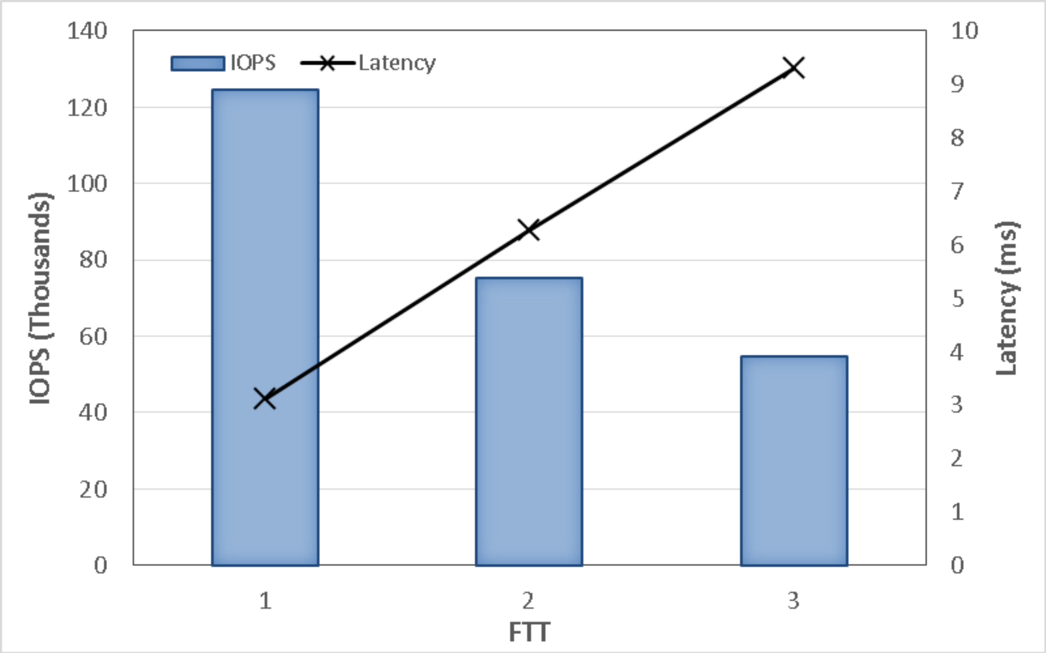


Figure 11. IOPS and latency with FTT for Mixed Read/Write workload

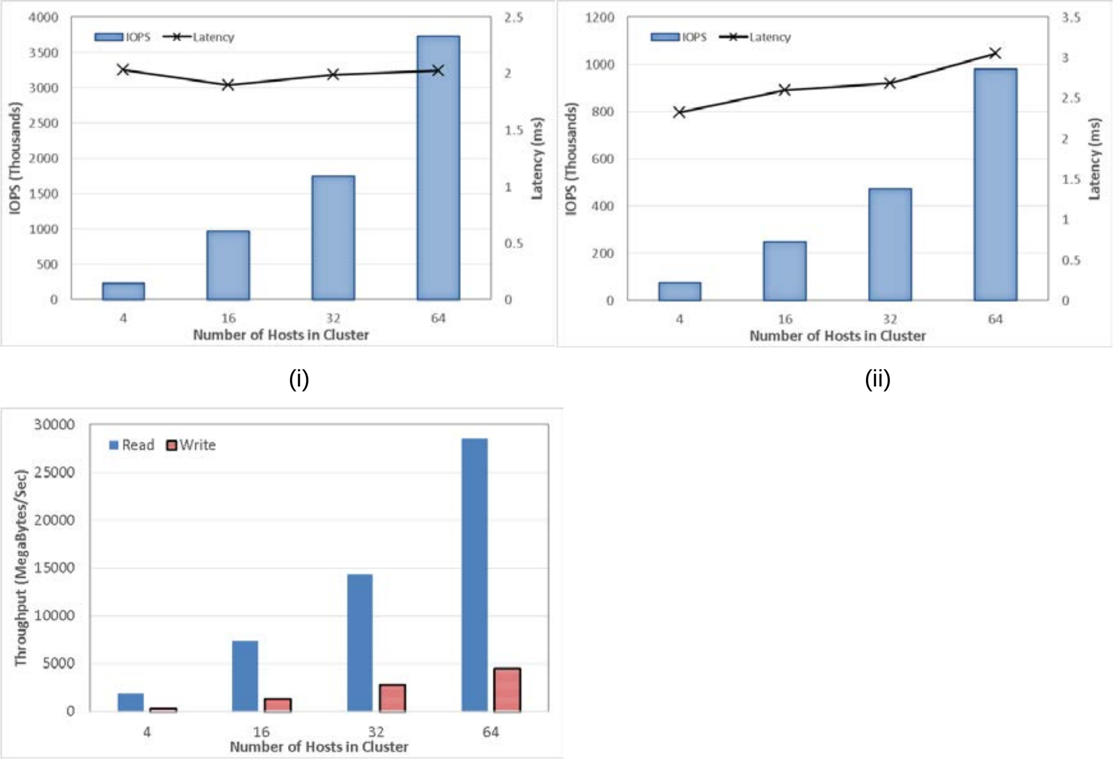
Size of the Virtual SAN Cluster

In this experiment, the cluster size is scaled up from 4 nodes to 64 nodes. The performance scalability of IOPS and latency is studied for the: (i) All Read, (ii) Mixed R/W, (iii) Sequential read, and (ii) Sequential write workloads. All workload parameters are depicted in Table 1. We maintain 1 VM per host, 1 disk group per host, 200GiB working set per VM, and 8 Iometer workers per VM.

Figure 12 shows the scalability results for IOPS and latency for the All Read and Mixed R/W workloads, and for bandwidth for the sequential workloads. One of the strong features of Virtual SAN is the close-to-linear scalability seen in all of these workloads. As an example, the All Read workload scales from 240 thousand IOPS (60 thousand IOPS per node) in a 4-node cluster to 3.73 million IOPS (58 thousand IOPS per node) in a 64-node cluster. This shows a 3% decrease in performance per node for a 16 times increase in cluster size. The latency does not vary by less than 5% as the size of the cluster is scaled up. The Mixed R/W workload scales from 62 thousand IOPS in a 4-node cluster (15.5 thousand IOPS per node) to 980 thousand IOPS (15.3 thousand IOPS per node) in a 64-node cluster. This shows a 1% decrease in performance per node for a 16 times increase in cluster size. Similar scalability improvements are observed with the sequential read and sequential write workloads.

Workload/Attribute	All Read	Mixed R/W	Sequential Read	Sequential Write
IO Size	4KiB	4KiB	256KiB	256KiB
Outstanding I/O per Iometer worker	16	4	8	8

Table 1. Workload parameters



(i)

(ii)

(iii)

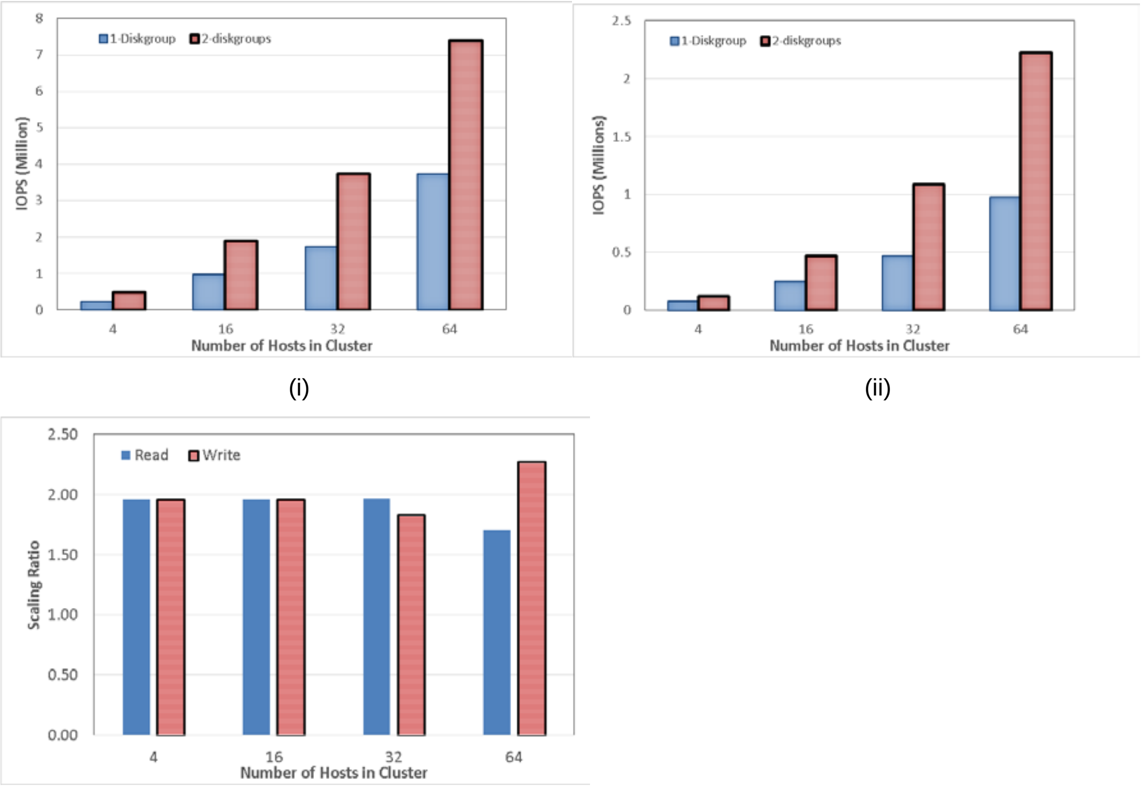
Figure 12. IOPS, latency, and throughput vs. number of hosts in Virtual SAN cluster for (i)All Read, (ii) Mixed R/W, and (iii) Sequential Read and Write workloads.

***Best Practice**: Virtual SAN scales extremely well with the size of the cluster. This gives Virtual SAN users the flexibility to start with a small cluster and increase the size of the cluster when required.*

Number of Disk Groups

In this experiment, the number of disk groups in the cluster is doubled, and scaling performance is measured from a 4-node cluster to a 64-node cluster. Two disk groups are created on each host of the Virtual SAN cluster. Each disk group consists of 1 SSD and 4 HDDs picked from the same I/O controller. The same workload profiles described in Table 1 are used. Figure 13 shows the detailed results comparing the performance of 1 disk group with 2 disk groups. The All Read workload scaled to 7.4 million IOPS on a 64-node cluster with 2 disk groups (115.6 thousand IOPS per node, 57.8 thousand IOPS per disk group per node).The Mixed R/W workload scaled to 2 million IOPS (31 thousand IOPS per node, 15.5 thousand IOPS per disk group per node). Both these results are a 100% increase compared to the single disk group. Similar scalability is observed in the sequential read and write workloads. Note that the Maximum Transfer Unit (MTU) was set as 9000 to the Virtual SAN network interfaces to get maximum performance for the two disk group configuration for the All Read workload. This is mainly to reduce the CPU utilization consumed by the vSphere network stack at such high loads.

Best Practice: *Virtual SAN scales well with the number of disk groups. To ensure the best performance, please make sure that the SSDs and HDDs of different disk groups are placed on different storage controllers. Setting a higher MTU (for example, 9000) may help to get maximum performance for all cached read workloads when using more than one disk group.*



(i)

(ii)

(iii)

Figure 13. IOPS, latency, and throughput scaling ratio with number of nodes for single and two disk groups in Virtual SAN cluster for (i) All Read, (ii) Mixed R/W, and (iii) Sequential Read and Write workloads.

All-Flash Virtual SAN Cluster Experiments

The All-Flash Virtual SAN cluster uses flash devices for both the caching and capacity tiers. With the use of SSDs in the capacity tier, Virtual SAN no longer caches read data in the caching tier. The caching tier SSD is now used purely as a write buffer. All guest writes are absorbed by the caching tier SSD and reads are satisfied from either the write cache or from the SSD layer. With the overhead of the read-cache eliminated and SSDs replacing the HDDs, improved IOPS and latency are expected with an All-Flash Virtual SAN cluster. The second impact of removing the read cache is that workload performance should stay steady as the working set size is increased beyond the size of the "caching" tier SSD. Test results demonstrate that both these expectations are met with the All-Flash Virtual SAN cluster. We conducted the All Read and Mixed R/W workload experiments on an 8-node All-Flash Virtual SAN cluster. We used one and two disk groups per host, keeping the number of capacity disks per disk group constant at three disks. Please note that unlike the Hybrid Virtual SAN cluster, a single I/O controller was used to host all the disks in the two disk groups case.

Since the capacity tier uses a SSD and can support higher random IOPS, outstanding I/Os were increased to 10 per Iometer I/O worker thread for the Mixed R/W workload (80 per VM per disk group).

All Read Workload

Figure 14 shows the impact of working set size per host of an All Read workload. The key point to note here is that even though the working set size exceeds total cache capacity by 50% (400GiB cache capacity vs. 600GiB working set for 1 disk group and 800GiB cache capacity vs. 1.2TB working set for 2 disk groups), both IOPS and latency stay relatively flat. We also observe that the IOPS show linear scalability with the number of disk groups. Another key point is that an All-Flash configuration is not expected to significantly outperform the Hybrid configuration when the working set size fits in the cache. This is borne out when the numbers are normalized in Figure 14 and Figure 3. The IOPS per disk group is 80 thousand for All-Flash and 60 thousand for Hybrid; latency is 1.2 milliseconds for All-Flash and 1 millisecond for Hybrid.

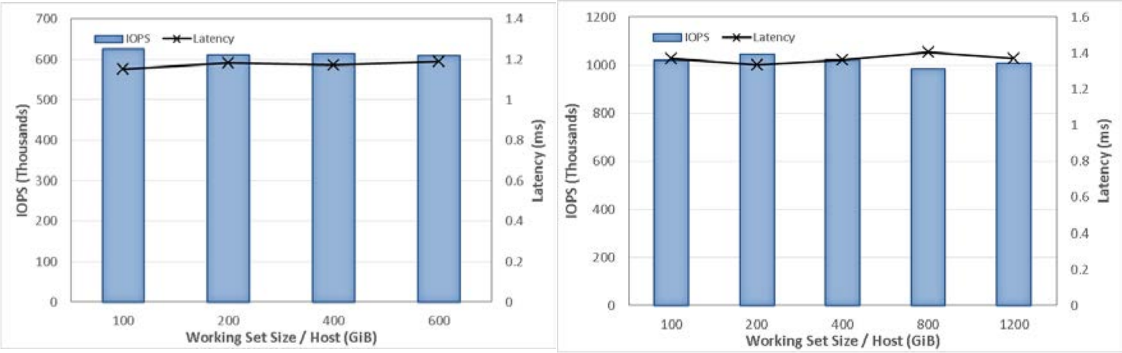


Figure 14. IOPS and latency for All Read workload on an All-Flash Virtual SAN Cluster with (i) single disk group and (ii) two disk groups.

Mixed R/W Workload

Since All-Flash Virtual SAN does not have a read-cache, reads of un-modified data come from the capacity tier. So the overhead of maintaining a read cache is removed. For workloads with random write patterns, moving the data from the cache tier to the capacity tier becomes more efficient, because the capacity tier is very likely better at random writes than HDDs. Because of these two reasons, the All-Flash Virtual SAN cluster is expected to significantly outperform the Hybrid Virtual SAN cluster for a Mixed R/W workload.

Because a 100% random I/O workload was running with very large working sets, the default ESXi settings for two parameters that are used to cache metadata were changed. For a majority of real-world workloads, the default size settings are expected to work fine. Since we were running a 100% random I/O workload with very large working sets, we had to increase the size of slabs used to cache metadata to get the maximum performance. For a majority of real-world workloads, we expect the default size settings to work fine. The settings are available as advanced configuration options and can be changed using esxcli or the SDK (KB 1038578) [9]. Please note that the settings cannot be changed from the Web Client or vSphere Client GUI. The configuration options, default values, and tuned values are listed below. For reference, the default settings can use up to a maximum of 1.3GiB of memory per disk group, whereas the tuned settings can use up to 10.4GiB of memory per disk group.
* /LSOM/blPLOGCacheLines, default=16384, tuned=131072
* /LSOM/blPLOGLsnCacheLines, default=1024, tuned=32768

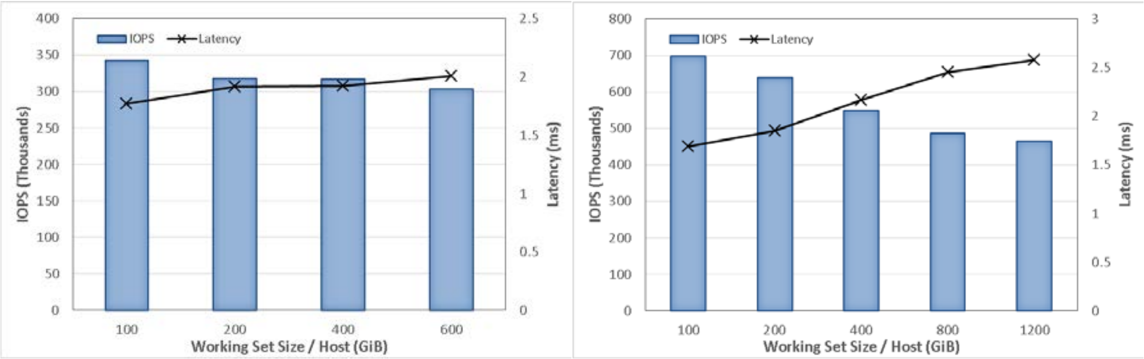


Figure 15. IOPS and latency for Mixed R/W workload on an All-Flash Virtual SAN Cluster with (i) single disk group and (ii) two disk groups.

Figure 15 plots the IOPS and latency for the single and two disk group cases. We can see that the IOPS per disk group can be as high as 45 thousand and latency is only 1.7 milliseconds. Even when the working set is 50% larger than the cache size, the performance is still at a very high level of 29 thousand IOPS per disk group and latency is at 2.5 milliseconds. On the other hand, in the case of the Hybrid Virtual SAN cluster, a cached working set can achieve 17 thousand IOPS per node at 2 milliseconds latency. At 28 thousand IOPS per node, the latency can go as high as 5 milliseconds. Please refer to Figure 2 for the data. This illustrates the advantages of the All-Flash configuration when lower latencies and larger working sets are desired. Table 2 shows a summary of comparison between the two clusters.

	Hybrid	All-Flash small working set, single disk group	All-Flash large working set, two disk groups
Working Set size	200GiB	100GiB	1.2TiB
Size of the caching tier	400GiB	400GiB	800GiB
Highest IOPS per disk group	28,000	45,000	29,000
Latency	5ms	1.7ms	2.5ms

Table 2. Summary of Hybrid Cluster performance vs. All-Flash Cluster performance for Mixed R/W workload

Best Practice: *Use the All-Flash Virtual SAN configuration for large working sets.*

Conclusion

This paper analyzed how Virtual SAN performed in different, well understood synthetic workloads, different disk group designs, and different configuration settings of Virtual SAN design. Several best practices were described to get the best performance out of a Virtual SAN cluster. It was observed that Virtual SAN scales extremely well with an increase in the cluster size and an increase in the number of disk groups. Excellent performance was demonstrated for cached random access workloads and for sequential access workloads on the hybrid Virtual SAN cluster. The All-Flash Virtual SAN cluster shows excellent performance and outperforms the Hybrid Virtual SAN cluster for random access workloads with very large working sets. The best practices explored in this white paper will help you design your Virtual SAN environment to get the best performance for your workloads.

Appendix A. Hardware Configuration for Hybrid Virtual SAN Cluster

Our servers had the following configuration.

- Dual-socket Intel® Xeon® CPU E5-2670 v2 @ 2.50GHz system with 40 Hyper-Threaded (HT) cores
- 256GB DDR3 RAM @1866MHz
- 2x LSI / Symbios Logic MegaRAID SAS Fusion Controller with driver version: 6.603.55.00.1vmw, build: 4852043
- 2x 400GB Intel S3700 SSDs
- 8x 900GB Western Digital WD9001BKHG-02D22 HDDs
- 1x Dual-Port Intel 10GbE NIC (82599EB, fiber optic connector)
- 1x Quad-Port Broadcom 1GbE NIC (BCM5720)

Appendix B. Hardware Configuration for All-Flash Virtual SAN Cluster

Our servers had the following configuration.

- Dual-socket Intel® Xeon® CPU E5-2670 v3 @ 2.30GHz system with 48 Hyper-Threaded (HT) cores
- 256GB DDR3 RAM @1866MHz
- 1x LSI / Symbios Logic MegaRAID SAS Fusion Controller with driver version: 6.603.55.00.1vmw, build: 4852043
- 2x 400GB Intel P3700 PCIe SSDs
- 6x 800 GB Intel S3500 SSDs
- 1x Dual-Port Intel 10GbE NIC (82599EB, fibre optic connector)
- 1x Quad-Port Broadcom 1GbE NIC (BCM5720)

References

[1] Cormac Hogan, VMware, Inc. (2015, March) Virtual SAN 6.0 Design and Sizing Guide. http://www.vmware.com/files/pdf/products/vsan/VSAN_Design_and_Sizing_Guide.pdf

[2] VMware, Inc. (2015, February) VMware Virtual SAN Datasheet. http://www.vmware.com/files/pdf/products/vsan/VMware_Virtual_SAN_Datasheet.pdf

[3] VMware, Inc. (2015, February) What's New in VMware vSphere 6.0? http://www.vmware.com/files/pdf/vsphere/VMware-vSphere-Whats-New.pdf

[4] A. Traeger, E. Zadok, N. Joukov, and C.P. Wright, "A Nine Year Study of File System and Storage Benchmarking," in *ACM Transactions on Storage (TOS)*, 2008, pp. 4, 5.

[5] VMware, Inc. (2014, July) VMware Horizon with View and Virtual SAN Reference Architecture. http://www.vmware.com/files/pdf/techpaper/vmware-horizon-view-virtual-san-reference-architecture.pdf

[6] Jinpyo Kim, Tuyet Pham, VMware, Inc. (2014, October) Microsoft Exchange Server Performance on VMware Virtual SAN. http://www.vmware.com/files/pdf/techpaper/Vmware-exchange-vsan-perf.pdf

[7] VMware, Inc. (2014, October) VMware Virtual SAN Performance with Microsoft SQL Server. https://communities.vmware.com/docs/DOC-27645

[8] VMware Labs. I/O Analyzer. https://labs.vmware.com/flings/io-analyzer

[9] VMware, Inc. (2015, March) Configuring Advanced Options for ESXi/ESX (1038578). http://kb.vmware.com/kb/1038578

[10] Intel Solid-State Drive DC S3700, Prodcut Specifcation. http://download.intel.com/newsroom/kits/ssd/pdfs/Intel_SSD_DC_S3700_Product_Specification.pdf

[11] (2014, August) Monitoring VMware Virtual SAN with Virtual SAN Observer. http://blogs.vmware.com/vsphere/files/2014/08/Monitoring-with-VSAN-Observer-v1.2.pdf

[12] VMware, Inc. (2015, January) Large-scale workloads with intensive I/O patterns might require queue depths significantly greater than Paravirtual SCSI default values (2053145). http://kb.vmware.com/kb/2053145

About the Authors

Amitabha Banerjee is a staff engineer in the I/O Performance Engineering group at VMware. Lenin Singaravelu is a senior staff engineer in the I/O Performance Engineering group at VMware. Their work strives to improve the performance of networking and storage products of VMware.

Acknowledgements

The authors would like to acknowledge Shilpi Agarwal, Maxime Austruy, Radu Berinde, Julie Brodeur, Emre Celebi, Cormac Hogan, Chuck Hollis, Christos Karamanolis, Yuvraaj Kelkar, Jinpyo Kim, Srinath Premachandran, Sandeep Rangasawamy, and Zach Shen for their contributions to Virtual SAN performance improvements, and for their valuable contributions to this white paper.

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com