VMware Virtual SAN™ Stretched Cluster

Performance and Best Practices

TECHNICAL WHITE PAPER

Table of Contents

Executive Summary 3

Introduction....3

Virtual SAN Stretched Cluster Setup.... 4

- Workloads....5
- Virtual Machine Configuration.... 6
- Metrics 6
- Virtual SAN Configuration Parameters 6

Performance of Stretched Cluster vs. Regular Cluster....7

- Mixed Read/Write (Mixed R/W) Workload with 100GiB Working Set 7
- Sequential Write Workload with 200GiB Working Set 8
- DVD Store Workload 9

Performance During Resource Failure....10

- Single Disk Failure 10
- Site Failure 12

Best Practices....13

- Impact of Inter-Site Latency and Bandwidth 13
- Resource Sizing for Stretched Virtual SAN Cluster....14
- Performance During Failure Recovery 14

Conclusion 14

Appendix A. Hardware Configuration for Hybrid Virtual SAN Cluster 15

Appendix B. DVD Store Virtual Machine and Workload Configuration Detail 15

Bibliography....16

Executive Summary

One of the most distinguishing features of VMware® Virtual SAN™ 6.1 is the availability of stretched cluster deployment. The stretched cluster allows the Virtual SAN customer to configure two geographically located sites, while synchronously replicating data between the two sites. This provides high availability and protection against a single site failure.

This white paper examines the performance aspects of a Virtual SAN stretched cluster deployment. Specifically, it examines the overhead of synchronously replicating data across two geographical sites by benchmarking against a regular, single site Virtual SAN cluster deployment. Failure scenarios of a single hard disk and entire site failure are considered. The Virtual SAN stretched cluster can handle both failure scenarios robustly.

Introduction

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 VMware vSphere® High Availability (HA), VMware vSphere® vMotion™, and VMware vSphere® Distributed Resource Scheduler™ (DRS) for failover. More information on Virtual SAN design can be obtained in the Virtual SAN design and sizing guide [1].

Virtual SAN stretched cluster refers to a deployment where a user sets up a Virtual SAN cluster with two active/active sites (hosts evenly split between the two sites) that are connected by a high bandwidth/low latency link, as well as a third site hosting a Virtual SAN Witness Host which is connected to both of the active/active data sites by a low bandwidth/high latency link. The Witness Host contains only the metadata components of all objects. The Virtual SAN stretched cluster extends the Virtual SAN cluster from a single site to two sites for a higher level of availability and inter-site load balancing. The stretched cluster is typically deployed in environments where the distance between data centers is limited, such as metropolitan or campus environments, and tolerance to failure is desired between the two data centers. The stretched cluster provides automatic handling of either the active/active site failing, or the network link between them failing. If either site fails, the storage on the other would be up to date and be able to continue, and vSphere HA [2] would restart any virtual machine that needs to be restarted.

The Virtual SAN stretched cluster architecture is different from how the regular (non-stretched, single fault domain) Virtual SAN cluster behaves. The following are the main differences.

(1) Write latency: In a regular Virtual SAN cluster, mirrored writes incur the same latency. In a stretched Virtual SAN cluster, the write operations need to be prepared on the two sites. Therefore, one write operation needs to traverse the inter-site link, and thereby incur the inter-site latency. The higher the latency, the longer it would take for the write operations to complete.

(2) Read locality: The regular cluster does read operations in a round robin pattern across the mirrored copies of an object. The stretched cluster does all reads from the single object copy available at the local site.

(3) Failure: In the event of any failure, recovery traffic needs to originate from the remote site, which has the only mirrored copy of the object. Hence, all recovery traffic traverses the inter-site link. In addition, since the local copy of the object on a failed node is degraded, all reads to that object are redirected to the remote copy across the inter-site link.

The focus of the experiments in this white paper is to:

- Understand the performance characteristics of a Virtual SAN stretched cluster with respect to a regular Virtual SAN cluster deployment (performance characteristics expected from a regular Virtual SAN deployment are described in detail in the Virtual SAN scalability and best practices performance study [3]). The test results emphasize how the active/active site configuration and the latency between the two active sites affects Virtual SAN stretched cluster performance.

- Examine how a Virtual SAN stretched cluster performs in the case of network or site failure and hard disk failure. The degradation in performance during the failure scenario and the recovery traffic that starts during failure recovery are measured.
- Provide best practices on Virtual SAN stretched cluster design and options required to tune the system for the best performance in specific cases.

All the experiments and performance comparisons in this whitepaper are done for a hybrid Virtual SAN setup consisting of SSDs in the caching tier and HDDs in the capacity tier.

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

Virtual SAN Stretched Cluster Setup

For performance investigations, a Virtual SAN stretched cluster was simulated in a lab environment. The hardware configuration of the experimental setup is as follows.

Two Virtual SAN stretched cluster setups were used for the experiments. Each stretched cluster contained two sites and one Witness Host. Setup 1 had seven Virtual SAN nodes, while Setup 2 had nine virtual SAN nodes. Each node had a similar hardware configuration; details on the hardware are available in Appendix A. Nodes were split equally into two fault domains (sites), leaving one node to be the witness node. Thus, Setup 1 was deployed in a 3+3+1 Virtual SAN stretched cluster configuration (three hosts on both sites and one Witness Host), while Setup 2 was deployed in a 4+4+1 configuration (four hosts on both sites and one Witness Host). Setup 1 was employed for all Iometer experiments described in this whitepaper and Setup 2 for the DVD Store experiments. A vSphere 6.0 Update 1 hypervisor [4] with Virtual SAN 6.1 [5] was installed on each node of the Virtual SAN cluster.

A metropolitan area network was replicated in a lab environment as follows. Figure 1 shows a diagrammatic layout of the setup. The two sites were placed on different VLANs. Then, a bridge was created between the two VLANs by using a native Linux machine. Netem [6] was used on the native Linux machine to simulate an appropriate inter-site latency between the two VLANs and thereby the two fault domains. The witness node was connected directly to the bridge machine. Since, conceptually, the witness node can be located in a far remote office for a stretched cluster deployment, the latency to the witness node may be much larger than inter-site latency. A Linux virtual machine with Netem support was deployed on the witness node to introduce the additional delay. For the experiments in this whitepaper, there was a witness delay of 200ms.

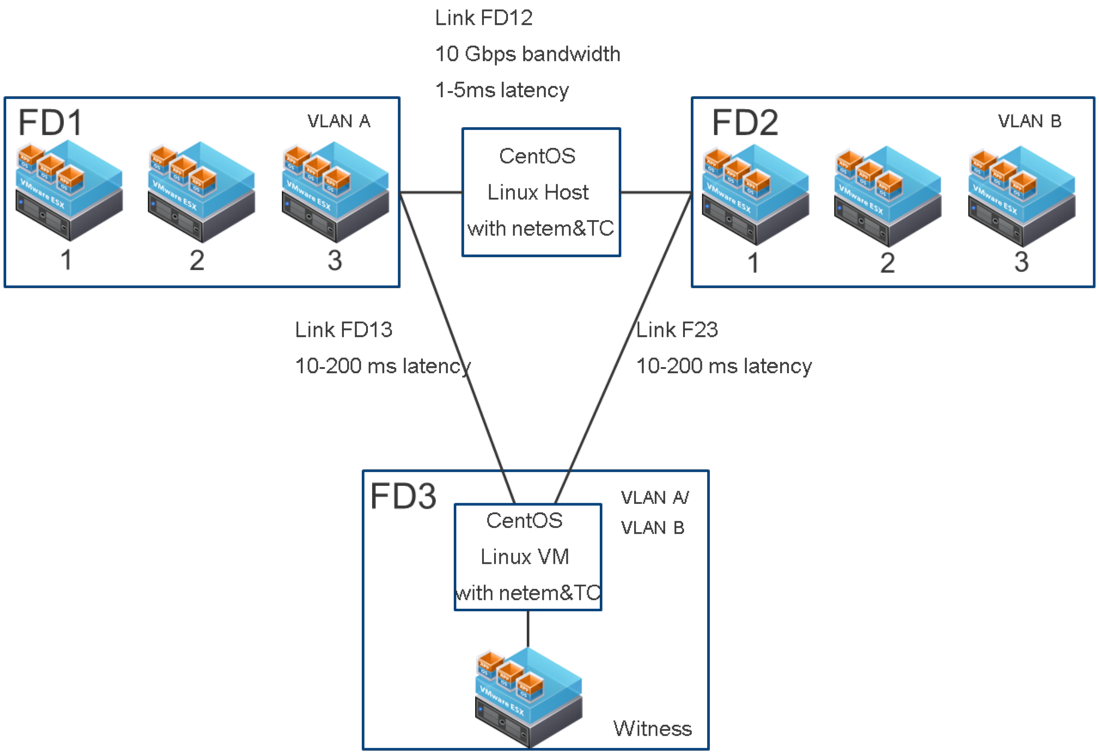

Figure 1. Simulation of 3+3+1 stretched cluster in a lab environment.

Workloads

The workloads for this performance study were Iometer version 2006.0727 (available in VMware IOAnalyzer [7]), and the DVD Store [8] applications.

The following I/O profiles were used for Iometer. For Iometer, the write-related workload is of more interest, hence:

(1) Mixed Read/Write (Mixed R/W) workload: Each Iometer worker thread was configured to do a mixed R/W access pattern with a 70%/30% ratio. All accesses were random. The I/O size was 4KiB. Eight Iometer worker threads were used, and eight outstanding I/O requests were maintained on each worker thread. 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 stretched cluster deployment.

(2) Sequential Write workload: Each Iometer worker thread was configured to do sequential write access with 256KiB I/O size. Eight Iometer worker threads and eight outstanding I/O requests on each worker thread were maintained. This trace is representative of scanned write operations, such as copying bulk data to a storage solution.

For the DVD Store workload, the open source DVD Store version 2.1 was used as a benchmark. DVD Store simulates an online ecommerce DVD store, where customers log in, browse, and order products. The benchmark tool is designed to use a number of advanced database features, including transactions, stored procedures, triggers, and referential integrity. The DVD Store workload used a database size of 100GB with 200 million customers and 1 million products within each virtual machine. For details of workload configuration, see Appendix B.

For the purpose of the Virtual SAN stretched cluster, emphasis is mainly on the performance overhead of I/O write operations with respect to the regular Virtual SAN cluster. For both workloads, how the write I/O performance is different for a Virtual SAN stretched cluster deployment is investigated, with respect to a regular Virtual SAN deployment.

Virtual Machine Configuration

The Iometer workload was deployed on an Ubuntu 12.0 virtual machine. Each virtual machine was configured with 4 vCPUs and 4GB of memory. The virtual machine was configured with three PVSCSI controllers: one for the OS disk; the other two equally shared the data disks. The queue depth for the PVSCSI controller was also increased to 254 as specified in KB 1038578 [9].

For the Iometer experiments, one virtual machine was deployed on each node of the cluster. Eight VMDKs per disk group per virtual machine were created. Each VMDK was available to the virtual machine as a block mount with 100% object space reservation. All the experiments were conducted after the VMDK was written to at least once. This prevented zeroed returns on reads. An Iometer I/O worker thread was configured to handle each VMDK independently.

For the DVD Store experiments, a Microsoft Windows Server 2008 R2 virtual machine is used. Besides the OS disk, two VMDKs per virtual machine were created to accommodate the database and log files, respectively. 4 virtual machines per node were used for comparing stretched cluster with regular cluster, and 1 virtual machine per node was used for site failure study.

Metrics

In the Iometer 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 virtual machine, the IOPS were measured and latency was observed for each I/O request on each running Iometer instance. All the experiments started from a state where the cache was dropped and was completely empty. This was done to achieve run-to-run consistency. The experiments were run for a duration long enough to ignore the cache warm-up behavior. In the case of the Mixed R/W workload, the experiment was run for a duration of three hours; the steady state duration was considered between the second and third hour mark. In the case of the Sequential Write workload, the experiment was run for a duration of two hours; the steady-state duration was considered between the first and second hour marks. The mean of the IOPS was added across each node in the cluster to get cumulative cluster-wide IOPS. Similarly, the mean latency was calculated across each node in the cluster to achieve the mean cluster-wide latency. The latency standard deviation was also noted, which gives confidence to the mean latency measure. These cluster-wide metrics of cumulative IOPS and latency help understand and characterize the Virtual SAN performance. For the sequential write workloads, the write bandwidth achieved in the Virtual SAN cluster was studied. The mean bandwidth on each node of the cluster was added together to achieve the cumulative cluster-wide bandwidth.

The primary performance metric of the DVD Store benchmark is orders per minute (OPM). The DVD Store benchmark driver outputs a moving average of orders per minute and a cumulative number of transactions every 10 seconds. The database size used in the tests are is 100GB per virtual machine. For the DVD Store workload, the OPM was compared, as well as guest read and write transactions, and the latency of each transaction.

Virtual SAN Configuration Parameters

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

- Single disk group with 1 Intel S3700 SSD and 4 HDDs for the hybrid Virtual SAN cluster
- Stripe width of 1
- Failures to Tolerate (FTT) of 1
- Default cache policies were used and no cache reservation was set. Each object was created with 100% object space reservation.

Performance of Stretched Cluster vs. Regular Cluster

Testing shows that the Virtual SAN stretched cluster retains acceptably high IOPS and bandwidth when compared with the regular Virtual SAN cluster in Iometer tests. In addition, the Virtual SAN metro cluster performs relatively well with an acceptable increase of application latency for a Tier 1 application, DVD Store.

Mixed Read/Write (Mixed R/W) Workload with 100GiB Working Set

The aforementioned Iometer workload was run with one virtual machine on each Virtual SAN host. A working set size of 100GiB per virtual machine was maintained, therefore, each VMDK was 12.5GiB in size. Figure 2 shows the guest IOPS and latency as described in the measurement metrics section.

Two inter-site, round-trip-time configurations were examined: 1ms inter-site latency and 5ms inter-site latency. As mentioned before, the two different round trip delays were simulated in the lab environment using Linux netem. The mean IOPS and latency for each node was recorded after the cache warm-up period, and an average of these numbers across the nodes in the cluster was taken to report the IOPS and latency.

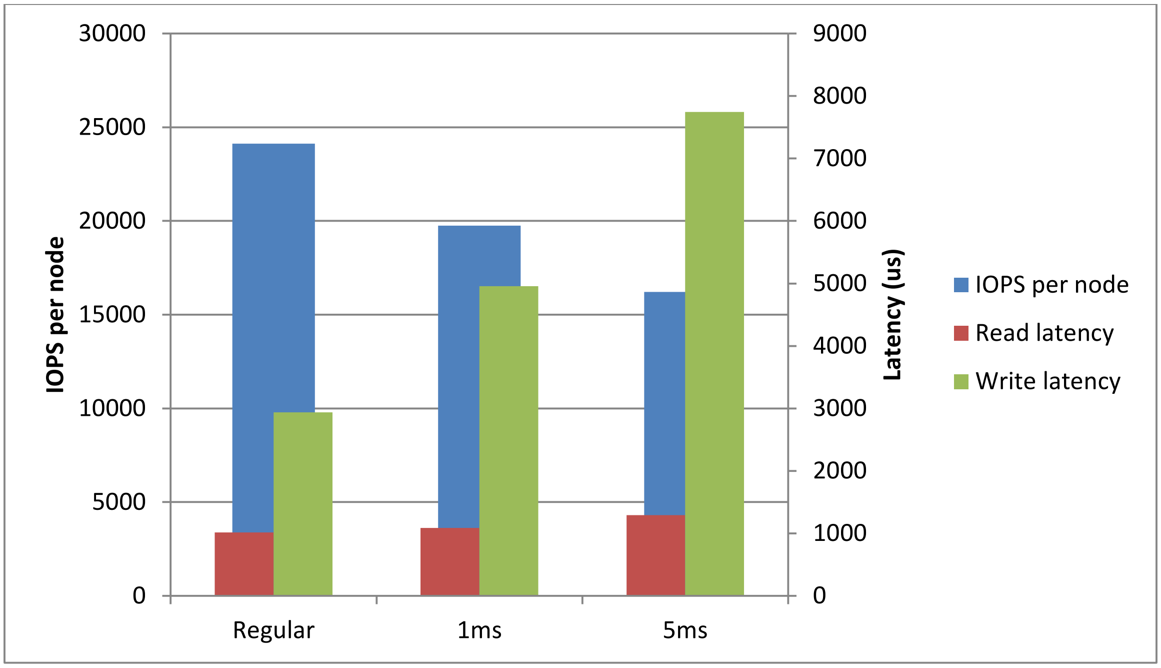

Figure 2. Comparison of IOPS and latency for a mixed read/write workload between the regular Virtual SAN cluster and a Virtual SAN stretched cluster with 1ms and 5ms inter-site latency.

It was observed that the IOPS per node is close to 20% lower for the Virtual SAN stretched cluster configuration with 1ms inter-site latency when compared with a regular Virtual SAN deployment. When the inter-site latency is changed to 5ms, the IOPS per node is 35% lower for the Virtual SAN stretched cluster configuration. There was also an increase in latency of write operations. From Figure 2, the latency for write operations are 2.9ms for the regular Virtual SAN cluster, and 4.9ms and 7.7ms for the Virtual SAN stretched cluster with 1ms and 5ms inter-site configurations respectively. The read latency numbers are 1.0ms for the regular Virtual SAN cluster, and 1.1ms and 1.3ms respectively for the Virtual SAN stretched cluster with 1ms and 5ms inter-site configurations respectively.

The increase in write latency, which causes a subsequent decrease in IOPS, can be explained. For every write operation issued from a guest, the underlying mechanism of the Virtual SAN stretched cluster pushes the write data to both local and remote sites to maintain data replication and provide for recovery during the event of site failure. Once the I/O operation is prepared on both sites, the write operation is acknowledged to the guest virtual machine. This imposes an extra inter-site latency for each write. Read I/O operations are programmed to read only from the local copy of the data; that is, read operations are not designed to traverse the inter-site link, unless there is a failure. Hence, the read latency in a Virtual SAN stretched cluster deployment is no different from a regular Virtual SAN cluster.

The Iometer workload used is an extremely aggressive micro-benchmark workload that drives very high I/O rate with a large number of outstanding I/Os per second. As such, the overhead of a stretched cluster in terms of an active/active site deployment and inter-site latency is reasonable. Experiments with real workloads such as DVD Store (described later) show that most applications do not encounter these overheads of a Virtual SAN stretched cluster and perform similarly with respect to a regular Virtual SAN cluster.

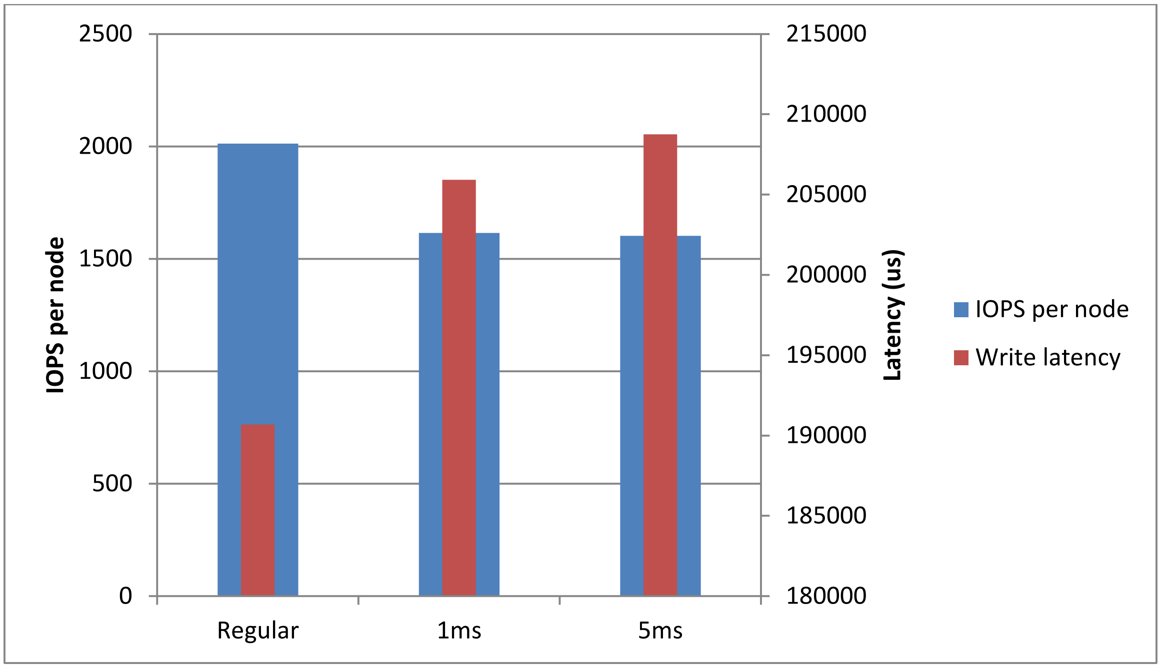

Figure 3. Comparison of IOPS and latency for a sequential write workload between the regular Virtual SAN cluster and a Virtual SAN stretched cluster with 1ms and 5ms inter-site latency.

Sequential Write Workload with 200GiB Working Set

Figure 3 shows the guest bandwidth and latency for a sequential write workload. The lower bandwidth seen in the stretched cluster can be again explained by the added latency due to write commits across the inter-site link. Both the Virtual SAN stretched cluster configurations with 1ms and 5ms inter-site latency can sustain a write bandwidth of close to 70 megabytes per second, per Virtual SAN stretched cluster node.

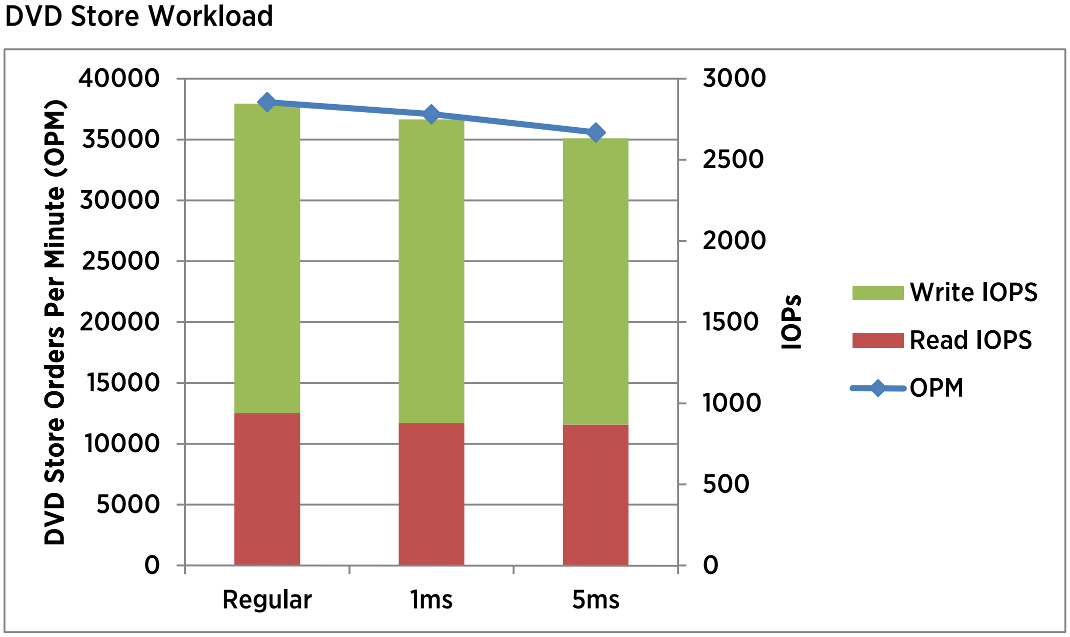

Figure 4(a). DVD Store OPM in the cluster and guest IOPs comparison.

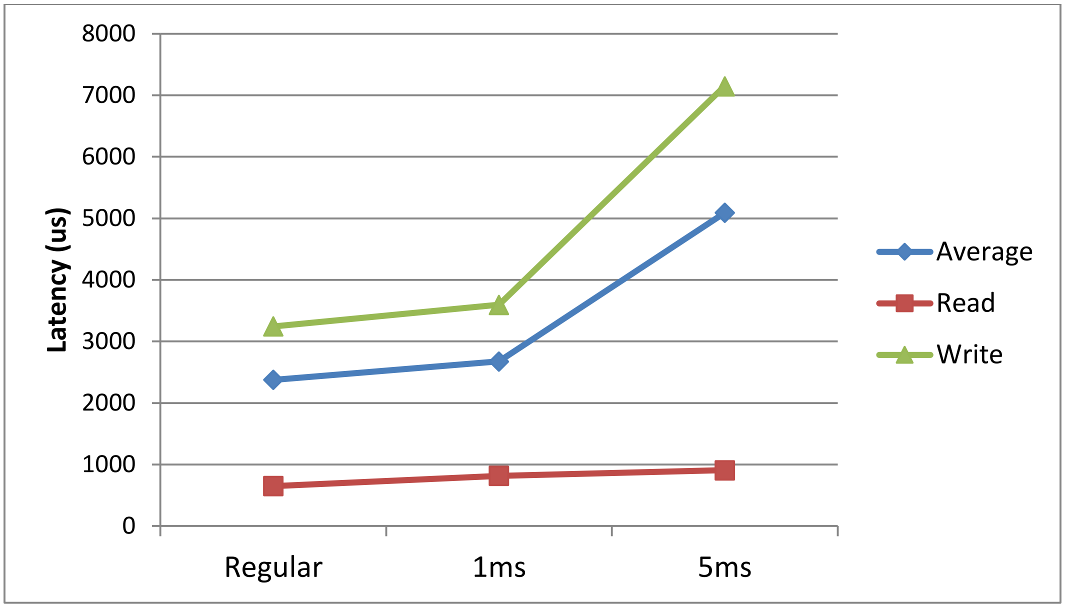

Figure 4(b). DVD Store latency comparison.

In this experiment, the DVD Store benchmark is executed on four virtual machines on each host of the nine-node Virtual SAN stretched cluster (Setup 2), and the results are compared with a similar workload on the regular Virtual SAN cluster. Figure 4 shows that DVD Store performance metrics of cumulated orders per minute in the cluster, read/write IOPs, and average latency. Clearly, DVD Store performs well in the Virtual SAN stretched cluster setup. The orders per minute (OPM) is lower by 3% and 6% for the 1ms and 5ms inter-site latency stretched cluster compared to the regular Virtual SAN cluster.

Guest read/write IOPS and latency were also monitored. The read/write mix ratio for the DVD Store workload is roughly at 1/3 read and 2/3 write. Write latency shows an obvious increase trend when inter-site latency is higher, while the read latency is only marginally impacted. This is because of the same reasons described in the previous Iometer mixed read/write section. As a result, the average latency increases from 2.4ms to 2.7ms, and 5.1ms for 1ms and 5ms inter-site latency configuration. The overheads of a Virtual SAN stretched cluster deployment have marginal performance impact on a commercial workload like DVD Store.

Performance During Resource Failure

Single Disk Failure

The most common failure event expected in a Virtual SAN cluster deployment is that of single hard disk drive (HDD) failing. In the Virtual SAN stretched cluster deployment, a copy of data is maintained on each site. Therefore, in the event of a HDD failure, a new replica needs to be created for all the objects on the failed HDD. The source of all the replicas is in the remote site; therefore, recovery traffic must flow through the inter-site network from one site to another. For this reason, the inter-site network must have sufficient bandwidth to facilitate a high volume of recovery traffic that can occur during a failure scenario. At the time of failure, all objects on the HDD are marked as degraded; therefore, all read and write operations to that object have to traverse the inter-site link to the remote site where the mirrored copy of the object exists. Moreover, since read cache at the remote site does not contain the degraded objects, read I/O operation hit the capacity tier during the cache warmup phase.

Figure 5 and Figure 6 analyze the effect of a single HDD failure on a Virtual SAN stretched cluster with 1ms of inter-site latency. Similar results were observed for a stretched cluster with 5ms inter-site latency. An Iometer Mixed Read/Write workload was running on all the nodes (on both sites) of the Virtual SAN stretched cluster. At steady state, the performance is about 19.5 thousand IOPS per Virtual SAN node. This is similar to the performance shown in Figure 2.

At around the time of 80 minutes, a single HDD was failed on one of the Virtual SAN nodes that had 4 HDDs, and the HDD was removed from the Virtual SAN data store. The failed HDD may have multiple objects on it. This triggered the recovery operation in which Virtual SAN recreated replicas of all the objects on the failed disk. Recovery traffic lasted for a duration of 54 minutes, the average recovery traffic was 15 megabytes per second, and the peak recovery traffic was 33 megabytes per second. The Iometer workload continued during the recovery phase; however, the average IOPS dropped to 16.3 thousand IOPS per Virtual SAN. Average read latency remained about the same, but average write latency increased from 4.9ms to 7.5ms. The performance drop is due to two reasons:

(1) All read and write operations to the degraded objects traverse the inter-site link.

(2) Recovery operations are proceeding in parallel with regular virtual machine I/O operations. Virtual SAN maintains independent queues for recovery and virtual machine traffic, and gives priority to this traffic. However, the volume of data that needs to be committed to the capacity tier (HDDs) is much higher with recovery traffic. This can slow down virtual machine I/O traffic.

At the time of 134 minutes, the recovery phase was complete, and all the replicas that were earlier lost due to the HDD failure were active. Performance of the Iometer workload returned to approximately 17.1 thousand IOPS per node; this is lower than steady state performance because the Virtual SAN data store was missing one HDD. The Virtual SAN elevator, which commits data from the caching tier (SSD) to the capacity tier (HDDs), executes slower on the host with one less HDD.

Figure 5 shows the guest I/O traffic and recovery traffic across the duration of the entire 3 hour run.

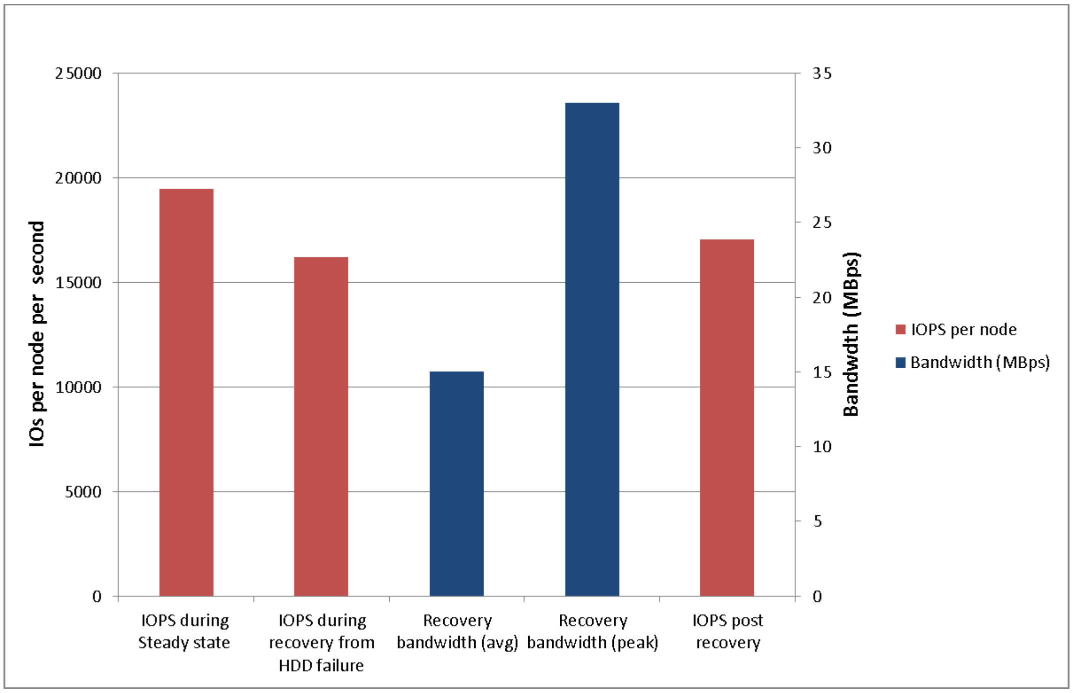

Figure 5. Performance during a single HDD failure.

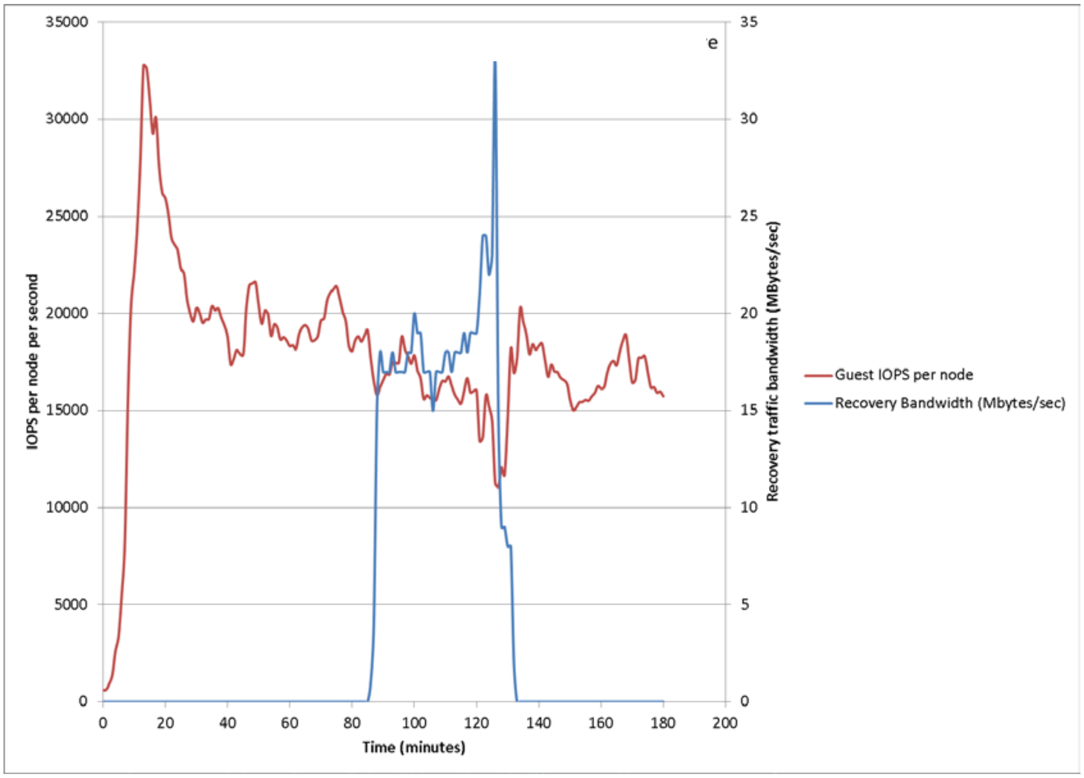

Figure 6. Performance of guest IOPS and recovery traffic during recovery from a single HDD failure.

Site Failure

This experiment demonstrates one of the most powerful features of the Virtual SAN stretched cluster: its ability to maintain data availability even under the impact of complete site failure. This experiment measures the impact of site failure on the DVD Store workload running on one virtual machine per host on the nine-node Virtual SAN stretched cluster with an inter-site latency of 1ms (Setup 2).

vSphere HA and vSphere DRS were enabled on the cluster. After the workload performance achieved steady state, an entire site (Site 1) of four nodes was turned down. This affected four virtual machines on the site that had been powered off. vSphere HA restarted the virtual machines on the other site (Site 2) and distributed one affected virtual machine on each node of the remote site. The site outage did not affect data availability because a copy of all the data of Site 1 existed on Site 2. Thus, the affected virtual machines were automatically restarted by vSphere HA without any issues. Once the virtual machines had restarted, the DVD Store benchmark was restarted on the affected virtual machines. The performance of the DVD Store virtual machines was monitored in terms of OPM and guest read and write latency, and these metrics were compared before and after failure.

The results shown in Figure 7 are interesting. After the site failure, the workload started from a cold cache for the virtual machines that were restarted on Site 2. Figure 7 shows the OPM per virtual machine and read and write latency at steady state after the cache warm-up period. DVD Store OPM per virtual machine was maintained, while the average write latency was reduced by 30% after Site 1 failed. The benchmark delivered better application latency performance after site failure than before the failure. This may be explained as follows. After site failure, data redundancy could not be maintained because the backup site was no longer active. As a result, there were faster write operations, which earlier were bottlenecked by write I/Os being synchronously prepared on the remote site. However, the performance improvement comes at a cost: none of the newly written data since the site failure has any backup. The stretched cluster is designed to tolerate only a single failure at a time. A second failure subsequent to the site failure will cause data loss.

Site 1 was brought back up after 5 minutes. The Virtual SAN stretched cluster started the process to restore the backup of the changed components on Site 2. This initiated recovery traffic from Site 1 to Site 2. Figure 7 shows the effect of this recovery traffic on DVD Store performance. Recovery traffic was measured at 310 megabytes per second and recovery took 8 minutes to complete. During the recovery process, after the cache warm-up period, a slight drop in performance was observed on DVD Store with 20% lower OPM per virtual machine, 3.5ms greater read latency, and 3ms greater write latency. This is due to the competition between the recovery traffic and virtual machine traffic on the same storage system. Virtual SAN tries to strike a balance between these two types of traffic with the aim to finish recovery as soon as possible, as well as have minimal impact on guest performance. After the recovery was complete, the benchmark metrics returned to what they were before site failure.

To sum up, this experiment shows that the Virtual SAN stretched cluster, retained performance during site failure for the DVD Store workload. Virtual SAN also kept a healthy trade-off between recovery traffic and the running workload once the failed site was brought back.

Note: The performance after site failure depends on adequate resources such as CPU, cache, and memory to be available to accommodate the virtual machines that are restarted by vSphere HA. In the above experiment, Site 2 had sufficient resources to accommodate two DVD Store virtual machines per host. Therefore, customers who require sustaining performance during site failure must ensure that their site configuration is overprovisioned and site resource utilization does not exceed 50%. This can be done using the admission control feature in vSphere HA.

Note: In the event of site failure, instead of bringing up the failed Virtual SAN hosts one by one, it is recommend to bring all hosts online approximately at the same time, within a span of 10 minutes. The Virtual SAN stretched cluster waits for 10 minutes before starting recovery traffic to a remote site. This avoids repeatedly resynchronizing a large amount of data across the sites. Once the site recovers, it is recommended to wait for the recovery traffic to complete before migrating virtual machines to the recovered site. If any virtual machines were to be migrated before recovery is complete, some read I/O operations would still traverse the inter-site link because the components in the failed site are not fully recovered. For the same reason, it is recommend that the vSphere DRS policy be changed from fully automated to partially automated in the event of site failure.

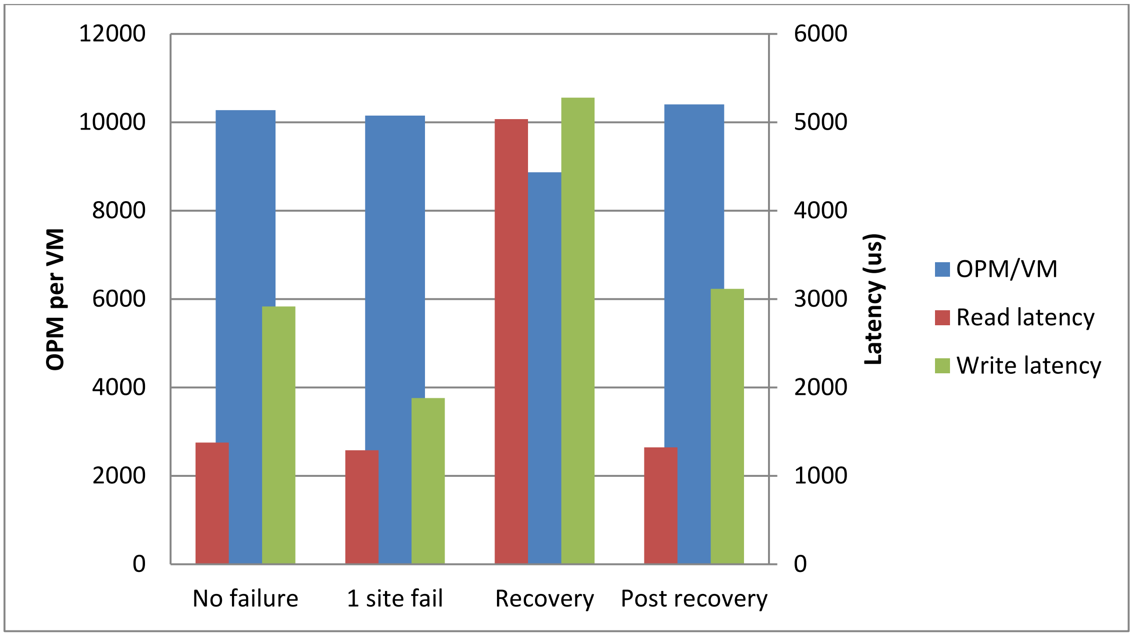

Figure 7. DVD Store result during network/site failure.

Best Practices

This section summarizes the best practices for a Virtual SAN stretched cluster deployment documented in the different sections of this white paper.

Impact of Inter-Site Latency and Bandwidth

One of the major distinguishing factors of a Virtual SAN stretched cluster deployment is the active/active site configuration and a metropolitan area link between the two sites. The performance of the Virtual SAN stretched cluster largely depends on bandwidth and latency available on this inter-site link.

It is recommended that the inter-site link have preferably 10 Gbps bandwidth. Higher inter-site bandwidth may be required for larger sized clusters. This bandwidth is required mainly to accommodate recovery from a failure scenario. In the event of any failure, the replicas in the failed node must be recreated from the data available on the other site. As displayed in Figure 4, even in the case of a single HDD failure, this recovery traffic could be very high. In the experiments with site failure, recovery peak traffic was measured at over 2 Gbps.

In terms of inter-site latency, higher inter-site latency affects the latency of write I/O transactions. The experiments show that with DVD Store, the Virtual SAN stretched cluster can sustain the overhead of 5ms latency without much impact to the benchmark performance. However, a 5ms inter-site latency does impact the write latency manifold when compared to a regular Virtual SAN deployment. It is recommended to limit inter-site latency to the order of 1ms, unless customer applications can tolerate high write latency.

Resource Sizing for Stretched Virtual SAN Cluster

In the event of site failure, all virtual machines start running on one site. Therefore CPU, memory, and cache resources must be overprovisioned to accommodate the failure event. Customers who require sustaining performance during site failure must ensure the admission control policy in vSphere HA is set so that the site resource utilization does not exceed 50%.

Performance During Failure Recovery

Recovery from a failure imposes significant stress on the Virtual SAN stretched cluster because the cluster needs to handle virtual machine I/O requests and recovery traffic at the same time. Virtual SAN design aims to balance the twin goals of fast recovery to ensure no data loss, and minimal impact to virtual machine I/O traffic. Testing shows that in the most common failure scenarios, such as HDD failure and site failure, the recovery traffic does not affect virtual machine I/Os significantly. Striking a balance between the two may depend on the customer's specific setting. While the Virtual SAN software stack is designed to achieve an excellent balance, fine tuning the priorities of the recovery traffic and virtual machine I/O traffic might be necessary in very specific cases to achieve the best performance when the customer desires faster recovery or lower impact to virtual machine I/Os during recovery. The following configurable tuning option is provided for this purpose.

```
#esxcfg-advcfg -g /VSAN/DOMResyncDelayMultiplier

Value of DOMResyncDelayMultiplier is 6

#esxcfg-advcfg -s <New Value, Min:1, Max: 256> /VSAN/DOMResyncDelayMultiplier
```

A higher value of DOMResyncDelayMultiplier delays recovery traffic, making the recovery process longer but less intrusive to virtual machine I/O traffic. Caution must be exercised while changing this flag because higher values may make recovery very time consuming. Similarly, a lower value may make recovery traffic very disruptive.

There could be extreme failure scenarios such as disk group failure (caused by SSD failure). Such a failure could cause a heavy volume of recovery traffic because of a large volume of data objects in the failed disk group; these objects now need to be backed up. HDD performance is on the critical path of these scenarios. Therefore, if such failure scenarios are foreseen, one remedy is to design the cluster with higher RPM HDDs that can sustain higher random I/O performance.

Conclusion

This white paper examines the various performance overheads that can exist in the Virtual SAN stretched cluster design. Testing shows that the Virtual SAN stretched cluster provides protection from site failure without introducing a significant performance penalty. This paper also describes the performance of the cluster under several failure scenarios. Testing shows that the Virtual SAN stretched cluster can adequately recover from site failure and balance recovery traffic with virtual machine I/O traffic.

Appendix A. Hardware Configuration for Hybrid Virtual SAN Cluster

The 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
- One LSI / Symbios Logic MegaRAID SAS Fusion Controller with driver version: 6.603.55.00.1vmw, build: 4852043
- One 400GB Intel S3700 SSDs
- Four 900GB Western Digital WD9001BKHG-02D22 HDDs
- One dual-port Intel 10GbE NIC (82599EB, Fibre Optic connector)
- One quad-port Broadcom 1GbE NIC (BCM5720)

Appendix B. DVD Store Virtual Machine and Workload Configuration Detail

The DVD Store workload virtual machine was configured as follows:

- 64-bit Microsoft Windows Server 2008 R2
- VMXNET3 driver version 1.2.20.0, PVSCSI driver version 1.1.1.0
- 50GB disk for the operating system with the LSI Logic controller
- 200GB database disk and 10GB log disk on the PVSCSI controller
- Mircosoft SQL Server 2008

The DVD Store workload was configured with:

```
n_threads=8
ramp_rate=100
run_time=180
db_size=100GB
think_time=0.002
pct_newcustomers=20
n_searches=15
search_batch_size=15
n_line_items=15
virt_dir=ds2
page_type=php
windows_perf_host=
linux_perf_host=
detailed_view=n
```

Bibliography

[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. vSphere High Availability. http://www.vmware.com/products/vsphere/features/high-availability

[3] Amitabha Banerjee and Lenin Singaravelu. (2015, March) Virtual SAN 6.0 Performance: Scalability and Best Practices. https://www.vmware.com/resources/techresources/10459

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

[5] VMware, Inc. (2015) What's New: Virtual SAN 6.1. http://www.vmware.com/products/whats-new-virtual-san.html

[6] Linux Foundation. (2009, November) Netem. http://www.linuxfoundation.org/collaborate/workgroups/networking/netem

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

[8] Todd Muirhead and Dave Jaffe. DVD Store benchmark. http://en.community.dell.com/techcenter/extras/w/wiki/dvd-store

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

About the Authors

Amitabha Banerjee is a Staff Engineer in the I/O Performance Engineering group at VMware. Zach Shen is a Member of Technical Staff in the I/O Performance Engineering group at VMware. Their work strives to improve the performance of networking and storage products of VMware.

vmware®

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