Web 2.0 Applications on VMware Virtual SAN™

Performance Study

TECHNICAL WHITE PAPER

vmware®

Table of Contents

Executive Summary

Web applications are an integral part of Enterprise and small business IT offerings, and a robust storage backend for these applications is essential. VMware performance labs undertook testing of VMware Virtual SAN™—a distributed, virtualized storage product that leverages the speed of multiple solid-state drives (SSDs)—with the Cloudstone benchmark, which simulates typical Web 2.0 technology use in the workplace. Tests show that Virtual SAN performs exceptionally in a high concurrent user, work-heavy environment where applications maintain impressively low latencies even when handling increased IOPS under stressed configurations.

Introduction

Virtual SAN is a distributed storage product that leverages the combination of SSDs and magnetic disks to provide a hypervisor-integrated, scalable storage solution. While Virtual SAN achieves impressive I/O performance with micro-benchmarks [1], Virtual SAN is also capable of performing well with a variety of workloads. This paper presents the results of experiments with Cloudstone, a Web 2.0 benchmark in vSphere 5.5[1]. These experiments show that Virtual SAN is a good choice for such workloads.

Cloudstone is a Web 2.0 benchmark based on the Olio toolkit and Faban harness. The Olio toolkit implements a social-events application as a multi-tier service and a typical implementation consists of a database server, an application server, and a Web server. The workload is driven by the Faban test harness and specifies the number of users, think-time (the amount of time between user actions), and a transition matrix between various types of operations (see "Appendix C"). The Olio toolkit and Faban harness have been used for various studies including Web server benchmarking, caching layer effectiveness, and hardware benchmarking [2]. Olio and Faban are also a part of the VMmark benchmark.

The test system was installed with a Cloudstone implementation of a MySQL database, NGINX Web server with PHP scripts, and a Tomcat application server provided by the Faban harness. Additional caching layers were not employed. The Web server and database were preconfigured for 500 users, for a starting size of 1.1GB for the database and 44GB for the Web server files. The default configuration was used for the workload generator and the number of users was varied. All components of the application were collocated in a single virtual machine to create a test scenario that focused on storage performance rather than the interactions among application components. Only the client ran in a separate virtual machine on separate hosts. One client-server pair provided a single independent Olio instance for up to 500 users. Scaling was extremely easy by running additional instances with more virtual machines.

Configuration

Hardware Configuration

"Appendix A" provides detailed hardware configuration. 3 hosts for server virtual machines and 3 hosts for client virtual machines were used. All 6 hosts had identical hardware configurations. Each host had two Intel Xeon Processors E5-2650 v2 (dual sockets, totaling 16 cores, 32 threads, @2.6GHz), 128GB memory, an LSI 9207-8i controller hosting one 400GB Intel S3700 SSD, and seven 1.1TB, 10K RPM SAS drives. The power management option was set to the "Performance" profile from BIOS of each host. Olio application traffic was configured to use a 1Gb network (over a 1Gb switch), and Virtual SAN traffic was configured to use a dedicated 10Gb network (over a 10Gb switch). Figure 1 illustrates the network connectivity of the hosts.

Virtual machines running Ubuntu 10.04 were configured according to the instructions listed in the Cloudstone benchmark [3]. Caching was disabled to stress the system. Virtual machine configuration details are listed in "Appendix B."

[1] vSphere 5.5 U1 was used for all experiments.

ESXi Configuration

The three server hosts were configured to form a server cluster and Virtual SAN was configured in this cluster. The remaining three hosts were set to be the client cluster. The client cluster was used to host virtual machines that initiate Web requests in the Olio workload and collect response time samples. Hence Virtual SAN was not configured in the client cluster, although it had the capability. The server virtual machines were equally distributed among the three server hosts. The client virtual machines were similarly distributed among the three client hosts. In this way, server virtual machines on one ESXi host only talked to client virtual machines on one client host. This was to eliminate 1Gb network bottlenecks for the Olio application traffic.

All storage components of the server virtual machines (configuration, disk, and log files) were stored on Virtual SAN, and they were configured with the default policy of FailuresToTolerate=1 and StripeWidth=1. Figure 1 shows the diagram of the host connectivity and its components relationship.

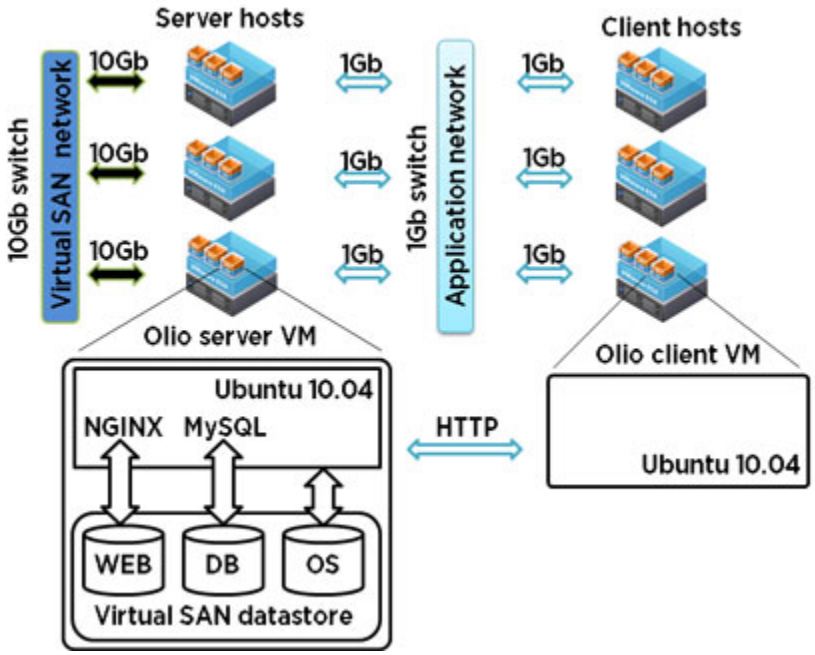

Figure 1. Host connectivity and component configuration

Performance

Evaluation Methodology

All experiments consisted of 6-hour runs and workload performance was measured with two sets of data points: average and 95th-percentile latency across the whole 6-hour period. Since Virtual SAN is designed as a storage caching layer, standard deviation of average latency for a steady state duration is also reported. Additional metrics[2], such as CPU and network utilization, and virtual disk and physical disk activity, were collected for further workload characterization.

Olio consists of seven types of operations: HomePage, Login, TagSearch, EventDetail, PersonDetail, AddEvent, and AddPerson. The proportion and frequency of the operations can be controlled and the default transition matrix

[2] The Virtual SAN cluster or the server cluster was the system under test and all the storage metrics were collected and reported against it.

provided by the Faban driver for Olio (described in "Appendix C") was used. The resulting frequency (operations per minute, per user) and the passing criteria are shown in Table 1.

The driver reported latency for each of these operations, and the average and 95th-percentile latencies were computed across all the virtual machines. These metrics were compared against a fixed criteria. Studies indicate that users are less likely to visit a Web site if the response time is greater than 250 milliseconds [4]. This number was used as an upper bound for average latency for frequent operations (HomePage, TagSearch, EventDetail, and Login). For the less frequent operations (AddEvent, AddPerson, and PersonDetail), a slightly more generous threshold of 500 milliseconds was used. For the 95th-percentile, the thresholds were doubled and used 500 milliseconds for the frequent operations and 1000 milliseconds for the less frequent operations.

	AddEvent	AddPerson	EventDetail	HomePage	Login	PersonDetail	TagSearch
Freq. Ops/Min	0.248	0.102	2.980	3.159	1.234	0.314	4.038
Avg RTT PASS (ms)	500	500	250	250	250	500	250
95th% RTT PASS (ms)	1000	1000	500	500	500	1000	500

Table 1. Operation frequencies and pass criteria

Experiment Configuration

The first experiment measured the activities of 500 users per virtual machine. Each server virtual machine was configured with 4 vCPUs and 4GB of memory to provide enough computational and caching capability. Client virtual machines were configured the same way. The experiments were conducted under a realistically busy load where more than 50% of the CPU resources were consumed, but the system was not saturated yet. To achieve that level, 15 virtual machines[3] were used, totaling 7500 Olio users. A 3 virtual machines[4] case was run (1500 users) for reference. The configuration is summarized in Table 2.

Name	Olio users	Per VM resources	
		vCPU	Memory (GB)
3 VMs	1500	4	4
15 VMs	7500	4	4

Table 2. Experiment configuration

The 3 virtual machines case uses only 17% of system CPU, while the 15 virtual machines case consumes 70% of CPU resources[5]. Figure 2 shows the Olio operation latencies results for both cases. The average and 95th-percentile latencies are well below the expected latencies and easily meet the passing criteria described in Table 1 of the previous section. For example, for the 15 virtual machines configuration, average latencies for operations of AddEvent, AddPerson, and PersonDetail top at 160 milliseconds. This is well below their passing threshold at 500 milliseconds. Other frequent operations have average latencies of less than 80 milliseconds and hence easily satisfy their passing criteria of being below 250 milliseconds. The average response time also shows low variation—the standard deviation of average response time samples[6] is in the range of 2% - 7% as shown by the

[3] 15 pairs of server-client virtual machines

[4] 3 pairs of server-client virtual machines

error bars in the figure. This indicates that when the system is moderately busy, Virtual SAN achieves good application latency performance.

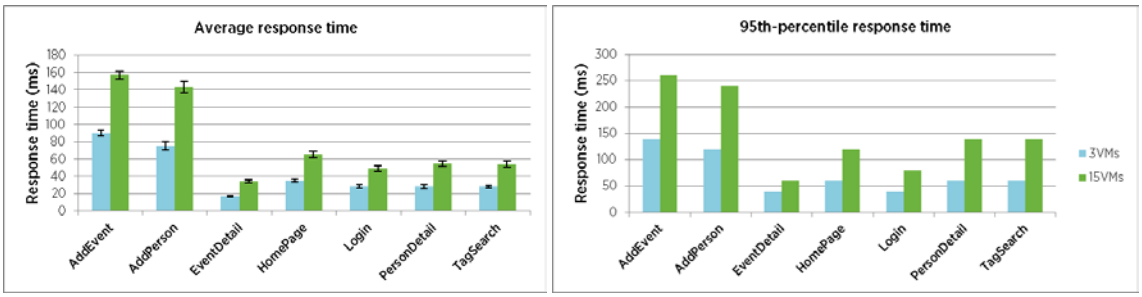

Figure 2. Round Trip Times (RTTs) for various operations with 3 and 15 virtual machines

Workload Characterization

In this 15 virtual machines configuration, the system is moderately busy. As mentioned before, the CPU utilization of the 3 hosts in the server cluster averaged 70%. This configuration also exercised a non-trivial-sized working set throughout the test. Table 3 shows the disk utilization at test start and ending time for the 3 dedicated disks in a single server virtual machine. The majority of disk space is consumed by Web files, and the file size also increased considerably throughout the test. After initialization, pre-copied Web files take 44GB in one virtual machine. At the end of a 6-hour test, the Web files size is doubled to 89GB. In the entire Virtual SAN cluster, the total working set size increases from 750GB to 1.4TB during the whole test duration.

	Start of test (GB)	End of test (GB)
Web	44.0	89.0
DB	1.1	1.3
OS	5.0	5.1
Total	50.1	95.4

Table 3. Disk utilization per virtual machine

Table 4 shows the aggregated IOPS, throughput and average I/O latency for the Web, DB, and OS disks in the 3-host Virtual SAN cluster. These storage metrics are all measured and collected at the vSCSI layer. The workload is write-intensive: there are about 10 times write operations than read operations. There is only light read IOPS, which can be attributed to efficient caching at the application and the OS layers. The disk write average latency is also much higher than read for each type of disk. For network traffic, Virtual SAN generates around 170Mbps in the dedicated Virtual SAN network while there is roughly 1Gbps non-Virtual SAN traffic in the application network ("Appendix D").

[5] The CPU usage number is the median CPU utilization of the hosts in the server cluster. The client hosts are generally at low CPU utilization for all the test cases in this paper and hence are not reported.

[6] The standard deviation is calculated based on a serial of average latency sub-samples. Each sub-sample is collected over a 30-second window. These sub-samples are collected consecutively in the steady state duration, which is the whole 6 hours test run excluding the first 30 and last 30 minutes.

	Read			Write		
	IOPs	Throughput (Mbyte/second)	Latency (ms)	IOPS	Throughput (Mbyte/second)	Latency (ms)
Web	56	0.46	1.41	562	13.93	9.19
DB	80	0.66	1.52	833	19.99	4.06
OS	42	0.35	1.52	433	10.51	7.53
Total / Average	179	1.47	1.48	1828	44.42	6.92

Table 4. Aggregated IOPS/throughput and average I/O latency for disks in the Virtual SAN cluster

Stressing Storage with the Olio Workload

From the IOPS numbers in the previous section, it is clear that even though the workload disk size is high, the number of operations sent to the physical disk is low due to efficient caching. To further stress the system, the same workload was run on two stressed configurations. In the first set, the users were kept steady at 500 per virtual machine but the memory size was reduced to 1GB. This reduces the amount of data that can be cached and therefore increases the number of IOPS reaching the Virtual SAN layer. One might note that it is an unrealistic configuration: although the workload is expected to consume similar CPU resources as in the previous set of experiments, it is configured to use only 4% of memory resources (15GB out of 384GB in the server cluster).

Name	Olio users	Per VM resources		Cluster resources used	
		vCPU	Memory (GB)	vCPU	Memory (GB)
15VMs - 4GB	7500	4	4	60	60
15VMs - 1GB	7500	4	1	60	15
70VMs - 1GB	7000	1	1	70	70

Table 5. Experiment configuration for stressing storage

For a more realistic scenario, the second stressed configuration was also used. 1-vCPU virtual machines were configured, each capable of handling 100 Olio users, but ran a total of 70 virtual machines[7] on the 3-node server cluster. This is similar to a Web server hosting scenario and boosts system memory utilization to 18% (70GB out of 384GB) while CPU utilization remained around 70%. The configurations are summarized in Table 5, and the previous un-stressed configuration is also shown as "15VMs-4GB" for reference. The experiments for the 2 stressed configurations were also run for 6 hours. The actual CPU utilization came out to be similar to the previous un-stressed case: 15VMs-1GB at 80%[8] and 70VMs-1GB at 72%. Figure 3 shows that aggregate vSCSI IOPS and throughput in the cluster have greatly increased, especially for the read path.

[7] 70 pairs of client-server virtual machines - the 70 server virtual machines are distributed randomly among 3 server hosts and leads to a roughly even distribution. So are the client virtual machines. The network connection between client and server virtual machines is also randomly chosen without a pre-defined rule.

[8] The intuitive reason for higher CPU utilization is from extra work on memory swapping due to less caching.

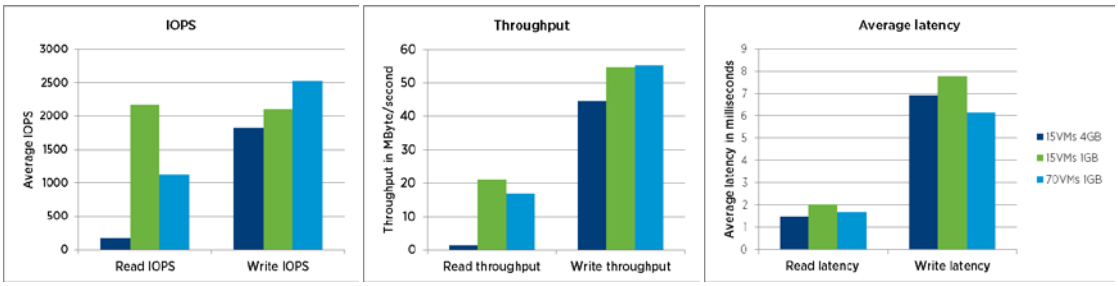

Figure 3. Aggregated storage metrics in the Virtual SAN cluster for stressed configurations

Correspondingly, an increase in response time at the application layer is seen in Figure 4. However, it is also clear that the passing criteria for all test cases are still met. This demonstrates Virtual SAN's capability in handling increased IOPS while retaining good application performance.

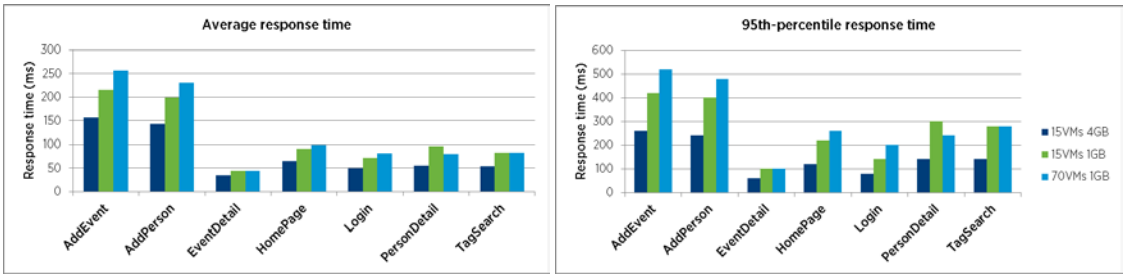

Figure 4. Operations RTT comparison for stressed configurations

Conclusion

Web 2.0 workloads perform well when using Virtual SAN as the storage layer. The Cloudstone benchmark is conducted where a workload of 7500 Olio users are run across 15 virtual machines in a 3-node Virtual SAN cluster. The systems are moderately busy at 70% average CPU utilization. The total working set size in the entire cluster almost doubled from 750GB to 1.4TB in the 6-hour test duration. The result shows the Olio application in the benchmark achieves exceptionally low latency with small variation over time. All Olio operations obtain average latencies that are well below 250 milliseconds and with only 2%-7% variation over time. Virtual SAN also retains its performance while handling increased storage stress. When the memory used for caching is reduced under two different configurations, the application latencies are still below the passing criteria, although the IOPS are dramatically higher in both cases.

Appendix A. Hardware Configuration

6 Dell R720 servers were used, and each server had the following configuration:

- Two Intel Xeon Processor E5-2650 v2 @ 2.60GHz
  - Dual sockets, totaling 16 cores/32 threads; Hyper-Threading is on
- 128GB RAM DDR3 @1866MHz
- 1 400GB Intel S3700 SSD (INTEL SSDSC2BA40, Firmware: DL04)
- 4 Hitachi 1.1TB, 10KRPM SAS drives (HUC101212CSS600, Firmware: U5E0)
- 3 Seagate 1.1TB, 10KRPM SAS drives (ST1200MM0007, Firmware: IS04)
- LSI Logic 9207-8i controller (LSI2308_2 chipset)
- 1 quad-port Broadcom 1Gb NIC (BCM5720)
- 1 dual-port Intel 10Gb NIC (82599EB, fibre optic connector)

One 1Gb port on each machine was connected to an Extreme X350 switch, and one 10Gb port of each machine was connected to an Arista 7050 switch. The rest of the NICs were not used in our experiments.

Appendix B. Virtual Machine Configuration

Every virtual machine in the experiment (both client and server) had the following configuration:

- 64-bit Ubuntu 10.04, Kernel 2.6.32-38-server
- VMXNET3 driver version 1.1.30.0, PVSCSI driver version 1.1.1.0
- 16GB disk with operating system on LSI Logic controller
- 16GB database disk and 100GB Web disk, both on PVSCSI controller
- NGINX 1.0.11, MySQL 5.5.20, PHP 5.3.9, Olio v0.2, Faban v1.0, Tomcat v6.0
- Java 1.6.0_32 is used for Faban harness on client virtual machines. Java heap size is configured to be 768MB

Appendix C. Faban Workload Driver Transition Matrix

The following example shows how to read the matrix shown in Table 6.

Example: Row EventDetail indicates that 72% of the time, the next operation is HomePage, 21% for Login, 6% for PersonDetail, and 1% for AddPerson.

	HomePage	Login	TagSearch	EventDetail	PersonDetail	AddPersion	AddEvent
HomePage	0	11	52	36	0	1	0
Login	0	0	60	20	0	0	20
TagSearch	21	6	41	31	0	1	0
EventDetail	72	21	0	0	6	1	0
PersonDetail	52	6	0	31	11	0	0
AddPerson	0	0	0	0	100	0	0
AddEvent	0	0	0	100	0	0	0

Table 6. Faban workload driver transition matrix

Appendix D. Network Utilization in the 15VMs-4GB Configuration

Table 7 shows per-host network traffic statistics for hosts in the Virtual SAN cluster.

	Virtual SAN traffic (10Gb PNIC stats)				Non-Virtual SAN - Olio traffic (1Gb PNIC)			
	Tx pps	Rx pps	Tx Mbps	Rx Mbps	Tx pps	Rx pps	Tx Mbps	Rx Mbps
15VMs–4GB	4963	16372	144.5	150.7	72240	31684	715.3	123.0

Table 7. Network utilization in 15 virtual machines 4GB memory case

References

[1] Wade Holmes. (2014, March) How to Supercharge Your Virtual SAN Cluster (2 Million IOPS!). http://blogs.vmware.com/vsphere/2014/03/supercharge-virtual-san-cluster-2-million-iops.html

[2] Will Sobel, Shanti Subramanyam, Akara Sucharitakul, Jimmy Nguyen, Hubert Wong, Sheetal Patil, Armando Fox, and David Pattersonl, "Cloudstone: Multi-Platform, Multi-Language Benchmark and Measurement Tools for Web 2.0," in *CCA*, 2008.

[3] EPFL PARSA. (2013) Cloudstone Software Download. http://parsa.epfl.ch/cloudsuite/web.html

[4] Steve Lohr. (2012, February) For Impatient Web Users, an Eye Blink Is Just Too Long to Wait. http://www.nytimes.com/2012/03/01/technology/impatient-web-users-flee-slow-loading-sites.html

About the Authors

Lenin Singaravelu is a staff engineer in the Performance Engineering team at VMware. His work focuses on improving storage and network performance of VMware's virtualization products. He has a PhD in Computer Science from the Georgia Institute of Technology.

Zach Shen is a member of technical staff in the Performance Engineering team. He also works on storage and networking performance. He has a PhD in Electrical Engineering from the University of Cambridge.

Acknowledgements

The authors would like to thank Shilpi Agarwal, Radu Berinde, Julie Brodeur, Priti Mishra, Harold Rosenberg, and Reza Taheri (in last names' alphabetical order) for reviews and contributions to the paper.

vmware

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