A MICRON REFERENCE ARCHITECTURE

Micron® Accelerated All-Flash VMware vSAN™ 6.6 Solution

Reference Architecture

systems software storage memory

Contents

Executive Summary

Data-intensive businesses that thrive in today's environment move quickly, and data platforms must move quickly with them. Technologies such as SSDs and advanced DRAM in conjunction with standard servers, multicore processors and state-of-the-art virtualization like VMWare vSAN™ are chasing application lethargy out of the data center.

This reference architecture provides deployment and testing details for one of the most compelling configurations: The Micron Accelerated VMware vSAN all-flash, high-performance server solution.

Similar to the standard AF-6 all-flash VMware vSAN Ready Node™ definition, this design combines NVM Express® (NVMe™) (cache tier) and SATA (capacity tier) enterprise-grade SSDs with advanced Micron® DRAM and Supermicro® standard rack-mount servers with 10 GbE networking. This design leverages a mix of high-performance (but more costly) NVMe SSDs and lower cost SATA SSDs to provide the optimal balance of cost and performance with vSAN based on our testing.

Optimized and engineered for VMware vSAN 6.6, this reference design enables:

- Faster time to deployment: Lab-tested by Micron experts in vSAN and thoroughly documented so you can deploy more quickly with greater confidence
- Balanced design: The right combination of cache and capacity SSDs, DRAM, processors and networking
- Confident deployment: Micron's in-house vSAN and workload expertise means you can build and deploy the platform in this reference architecture with confidence

The configuration in this reference architecture ensures easy integration and operation with vSAN 6.6, offering predictably high performance that is easy to deploy and manage.

TESTED • INTEGRATED • OPTIMIZED
SOLID READY
DATA CENTER READY

Micron's Accelerated Solutions—A New Approach to Platform Designs

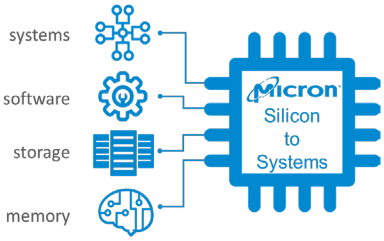
Micron's SOLID Ready family of accelerated solutions are optimized, pre-engineered, enterprise-leading platforms that are co-developed between Micron and industry leading hardware and software companies.

Designed and tested at Micron's Storage Solutions Center, these best-in-class solutions provide end users, channel builders, independent software vendors (ISVs) and OEMs with a broader choice in deploying next-generation solutions with reduced time investment and risk.

The Purpose of This Document

This document describes a reference architecture for deploying a performance-optimized all-flash vSAN-enabled VMware vSphere® cluster using a combination of Micron® SSDs (NVMe and SATA) in a Micron reference design, detailing the hardware and software building blocks and the measurement techniques to characterize the reference architecture's performance. This document covers the Micron reference design's composition including the vSphere configuration, network switch configurations, vSAN tuning parameters, configuration of the Micron reference nodes and Micron SSDs.

The purpose of this document is to provide a pragmatic blueprint for administrators, solution architects and IT planners who need to build and tailor a high-performance storage infrastructure that scales for I/O-intensive workloads.



Why Micron for this Solution

Storage (SSDs and DRAM) can represent up to 70% of the value of today's advanced server/storage solutions. Micron is a leading designer, manufacturer and supplier of advanced storage and memory technologies with extensive in-house software, application, workload and system design experience.

Micron's silicon-to-systems approach provides unique value in our reference architectures, ensuring these core elements are engineered to perform in highly demanding applications like vSAN and are holistically balanced at the platform level. This reference architecture solution leverages decades of technical expertise as well as direct, engineer-to-engineer collaboration between Micron and our partners.

Solution Overview

A vSAN storage cluster is frequently built from a number of vSAN-enabled vSphere nodes for scalability, fault-tolerance and performance. Each node is based on standard server hardware and utilizes VMware® ESXi™ hypervisor to:

- Store and retrieve data
- Replicate (and/or deduplicate) data
- Monitor and report on cluster health
- Redistribute data dynamically (rebalance)
- Ensure data integrity (scrubbing)
- Detect and recover from faults and failures

Enabling vSAN on a vSphere cluster creates a single vSAN datastore. When virtual machines (VMs) are created, virtual disks (VMDKs) can be allocated from the vSAN datastore. Upon creation of a VMDK, the host does not need to handle any kind of fault tolerance logic, as it is all handled by the vSAN storage policy applied to that object and vSAN's underlying algorithms. When a host writes to its VMDK, vSAN handles all necessary operations such as data de-duplication and compression, erasure coding, checksum calculation and placement based on the selected storage policy.

Storage policies can be applied to the entire datastore, a VM, a VMDK, or nearly any other individual object. Using storage policies enables the user to determine whether to add more performance, reliability, or capacity to an object. Multiple storage policies can be used on the same datastore, allowing a user to create high-performance VMDKs (for database log files, for example) and high-capacity/availability disk groups (for critical data files). Figure 1 shows the logical layers of the vSAN stack, from the hosts down to the vSAN datastore.

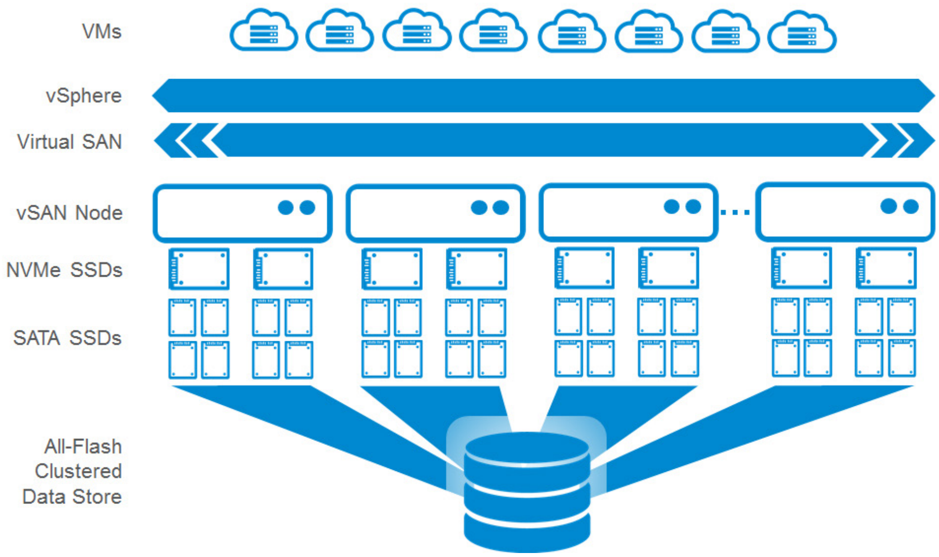


Figure 1: vSAN Architecture

Client VMs write to vSAN VMDKs while the vSAN algorithms determine how data is distributed across physical disks according to the storage policy for that VMDK. Below are some of the options that comprise a storage policy:

- **Primary levels of failures to tolerate (FTT):** Failures to tolerate specifies how many copies of data can be lost while still retaining full data integrity. By default, this value is set to 1, meaning there are two copies of every piece of data, as well as potentially a witness object to make quorum in the case of an evenly split cluster.
- **Failure tolerance method (FTM):** The method of fault tolerance. There are two choices: 1) RAID-1 (Mirroring) and 2) RAID-5/6 (Erasure coding). Choosing RAID-1 (Mirroring) will create duplicate copies of data in the amount of 1 + FTT. Choosing RAID-5/6 (Erasure coding) will stripe data over three or four blocks, as well as 1 or 2 parity blocks, for RAID-5 and RAID-6, respectively. Selecting FTT=1 means the object will behave similar to RAID-5, whereas FTT=2 will be similar to RAID6. The default is RAID-1 (Mirroring).
- **Object space reservation (OSR):** This value specifies the percentage of the object that will be reserved (thick provisioned) upon creation. The default value is 0%.
- **Disable object checksum:** If Yes is selected, the checksum operation is not performed. This reduces data integrity, but can increase performance, typically used when performance is more important than data integrity. The default value is No.
- **Number of disk stripes per object (DSPO):** This is the number of objects over which a single piece of data will be striped. This only applies to the capacity tier, not the cache tier. The default value is 1, and can be set as high as 12. Note that vSAN objects are automatically split into 255GB chunks, but are not guaranteed to reside on different physical disks. Increasing the number of disk stripes will guarantee that they reside on different disks on the host, if possible.

The Micron Accelerated vSAN Solution uses a combination of two different enterprise SSDs. Each node's cache tier is built with two NVMe SSDs. Each node's capacity tier is built with eight SATA SSDs combined with Micron high-speed memory and together with servers from Supermicro and networking components from Dell.

Software

VMware's vSAN is a hyper-converged infrastructure (HCI) solution, combining traditional virtualization with multi-host software-defined storage. vSAN is a technology that is part of VMware's vSphere environment, coupled with the ESXi type-1 hypervisor for which VMware is well known.

A vSAN cluster is created by installing ESXi on at least three nodes (four or more is recommended), and enabling vSAN via a license in the vSphere Client. vSAN uses a two-tier architecture, where all write operations are sent to the cache tier and are destaged to the capacity tier over time. Up to 600GB of cache can be utilized per host, split among 1-5 disk groups.

vSAN can operate in two modes:

- Hybrid: Utilizes SSDs for the cache tier and rotating media for the capacity tier
- All-flash: Utilizes SSDs for both cache and capacity tiers

With a hybrid configuration, the cache tier is used as both a read and write cache, keeping "hot" data in the cache for better performance. In this configuration, 70% of the cache capacity is dedicated to the read cache and the remaining 30% is dedicated for the write buffer.

In an all-flash configuration, 100% of the cache tier is used for the write buffer, with no read cache.

Micron SSD Components

This reference architecture uses two types of Micron enterprise SSDs. Each node's cache tier is built with a pair of NVMe SSDs—the Micron 9100 MAX (U.2, 1.2TB). The capacity tier is built with eight SATA SSDs—the Micron 5100 ECO (2.5-inch, 4TB).

Cache Tier (NVMe)		Capacity Tier (SATA)	
Model	9100 MAX	Model	5100 ECO
Form Factor	U.2	Form Factor	U.2
Interface	PCIe Gen3 x4	Interface	SATA 6 Gb/s
Capacity	1.2TB	Capacity	3.84TB
Sequential Read (128KB)	2.9 GB/s	Sequential Read (128KB)	540 MB/s
Sequential Write (128KB)	1.3 GB/s	Sequential Write (128KB)	520 MB/s
Random Read (4K)	700,000 IOPS	Random Read (4K)	93,000 IOPS
Random Write (4K)	210,000 IOPS	Random Write (4K)	18,000 IOPS
Endurance (Sequential)	4.8PB	Endurance	6.48PB
Endurance (Random)	3.5PB		

Table 1: vSAN SSD Specifications

NVMe SSDs NVM Express (NVMe) is an industry standard interface designed to accelerate the performance of nonvolatile SSDs. The 9100 family provides workload-focused endurance and capacities for both read-centric and mixed-use applications and environments.

Offered in half-height, half-length (HHHL) and U.2 industry standard form factors in capacities up to 3.2TB, the 9100 provides an exceptional balance of performance, endurance and price. For additional details, see the 9100 page on micron.com.

Micron is a Promoter Member of NVM Express, contributing technical expertise toward the development of the NVMe specification. Our 9100 MAX high-performance SSDs with NVMe are well suited for vSAN cache tier deployments:

- **Enhanced Performance:** The 9100 PCIe SSD accelerates data with read/write throughput of 2.9 GB/s and 1.3 GB/s and read/write IOPS of 700,000 and 210,000 (steady state) and 99.9% read/write latency of 230/70µs (QD=1). See the Micron 9100 datasheet for additional details.

- **Reliability and Quality:** Protect mission-critical data with power-loss protection and data path protection features (see the Micron 9100 datasheet for additional details).

- **Optimized Endurance:** Choose from endurance options matched to your workloads.

- **XPERT Firmware Features:** Rest easy with eXtended Performance and Enhanced Reliability Technology (XPERT) features such as power-loss protection, RAIN, data path protection, reduced command access latency, and adaptive read and thermal protection. (See the Micron 9100 benefits page for additional details.)

SATA SSDs Micron's 5100 ECO read-focused SATA SSD provides workload-focused endurance and capacities for read-centric, mixed-use and write centric applications and environments.

Offered in standard 2.5-inch (7mm high) form factors, the 5100 series easily integrate into standard server platforms, with the ECO version being well-suited for vSAN capacity tier deployments. For additional details, see the 5100 page on micron.com. Key features of the 5100 series include:

- **High Capacity:** Consolidate storage platforms and smooth migration from legacy storage. The 5100 offer up to 8TB of storage in a 2.5-inch form factor.

- **Consistent, High Performance:** Meet the demands of your data center. The 5100 comes in three models optimized for varying workloads with consistent, steady state IOPS and MB/s.

- **Flexibility:** Actively tune capacity to optimize drive performance and endurance with Micron's FlexPro™ firmware architecture.

- **Reliability:** Reduce downtime and latency with well-managed quality of service (QoS) that is unmatched compared to spinning media.

Micron DRAM This solution also utilizes Micron DRAM. For additional details, see the configuration information later in this document.

Server Platforms

This solution utilizes Supermicro SYS-2028U-TNRT+ servers (platform details available on Supermicro's website), which are 2U dual-socket servers based on the Intel® Grantley platform.

Each server is configured with two E5-2650v4 processors, each with 12 cores at 2.20GHz. These processors align with VMware's AF-6 requirements, which is their nomenclature for a medium-sized all-flash configuration.

Solution Network

Switch

vSAN utilizes commodity Ethernet networking hardware. This solution uses two Dell Networking N4032F switches for all cluster-related traffic. Both switches are connected with a single QSFP+ cable between them. Spanning Tree is enabled to avoid loops in the network. All ports are configured in general mode, with VLANs 100-114 allowed.

vSAN requires at least three separate logical networks, which are all segregated using different VLANs and subnets on the same switches. The three networks, and their respective VLANs, are as follows:

Role	VLAN ID	Network Subnet	Description
Management/ VM Network	100	172.16.17.x/??	This network will host all client-server traffic between the VMs and the data center as well as all VMware/vSAN management traffic for monitoring and management
vMotion	101	192.168.1.x/??	This network is used to support movement of VMs from one node to another within the cluster.
vSAN	102	192.168.2.x/??	This network is used to support transfer of all storage data associated with the vSAN shared storage pool between the various nodes of the cluster.

Table 2: Network Configuration

While using different subnets or VLANs alone would suffice, adding both ensures that each network has its own separate broadcast domain, even if an interface is configured with either the wrong VLAN or IP address. To ensure availability, one port from each server is connected to each of the two switches, and the interfaces are configured in an active/passive mode.

NIC

Each server has a single dual-port Intel® 82599EB 10 GbE NIC. As mentioned above, one port of each NIC is connected to one of each of the switches to ensure high availability in the case of losing one of the two switches. vSAN is active on one link and standby on the other, whereas management and vMotion are active on the opposite link. This ensures that vSAN gets full utilization of one of the links, and is not interrupted by any external traffic.

Solution Reference Design

Hardware

This section describes the configuration of each component shown below and how they are connected.

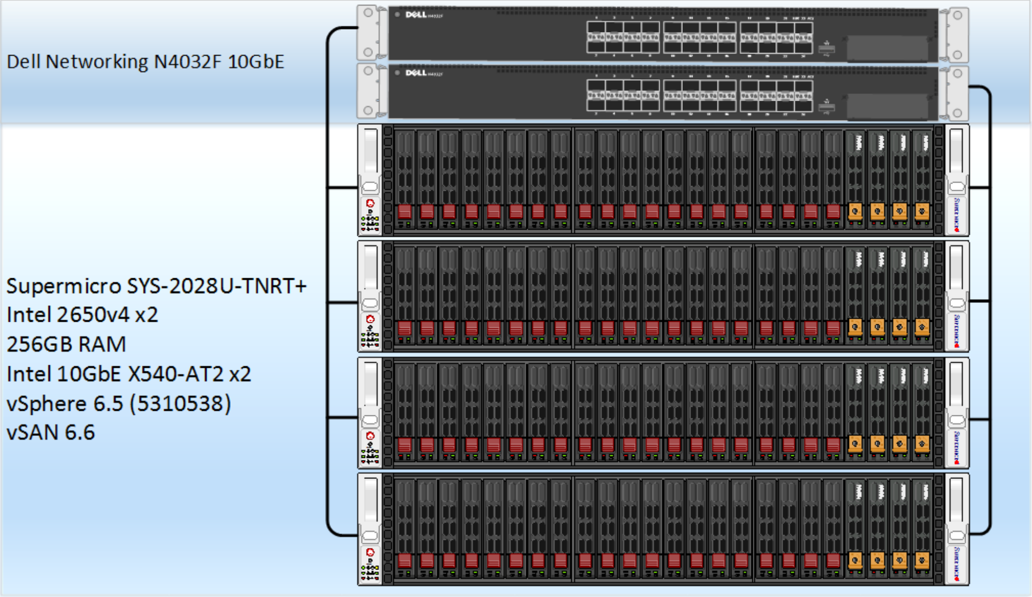


Figure 2: vSAN Reference Design Hardware

Node Components

- Supermicro Ultra Server SYS-2028U-TNRT+
- 2 Intel Xeon E5-2650v4 12-core CPUs
- Micron 256GB 2400 MHz RAM (16GB x 16)
- OS Drive: 16GB SATA DOM
- 2 Micron 9100 MAX NVMe SSDs 1.2TB
- 8 Micron 5100 ECO SATA SSDs 3840GB
- 1 Intel dual-port 10 GbE SFP+ NIC (82599EB)
- 2 LSI 9300-8i SAS/SATA HBA

Network Infrastructure

- 2 Dell Networking N4032F 10 GbE switches
- Dell SFP+ copper cables

Software

Software Components (per node)

- vCenter Server Appliance 6.5.0.10000 build 5973321
- ESXi build 5310538
- vSAN 6.6
- Disk format version 5.0
- HBA driver 6.611.07.00-1OEM.600.0.0.2768847
- HBA driver 6.611.07.00-1OEM.600.0.0.2768847
- HBA firmware 4.650.00-6223
- 9100 driver 1.2.0.32-2vmw.650.0.0.4564106
- 9100 firmware 00091634
- 5100 firmware D0MU402

Planning Considerations

Part of planning any configuration is determining what hardware to use. Configuring a system with the most expensive hardware might mean overspending, whereas selecting the cheapest hardware possible may not meet your performance requirements.

This study targeted a configuration close to VMware's AF-6 vSAN specifications, which aims to provide up to 50K IOPS per node. An AF-6 configuration typically calls for at least 8TB of raw storage capacity per node, dual processors with at least 12 cores per processor, 256GB of memory, at least two disk groups per node and 10GbE networking. For more information on AF-6 requirements, see https://www.vmware.com/resources/compatibility/vsan_profile.html

It is important to note that there are many ways in which performance can be increased, but often with added cost. Using a processor with a higher clock speed would potentially add performance, but could add thousands of dollars to the configuration. Adding more disk groups would also add significant performance, but again, it would add significant cost to the solution having to buy additional cache drives. Furthermore, adding faster networking—like 40 GbE, 100 GbE, or Infiniband—would yield better performance, but all the necessary hardware to do so would again add significant cost to the solution. The solution chosen for this study is moderately sized for good performance at a balanced price point.

Testing Methodology

Benchmarking virtualization can be a challenge because of the many different system components that can be tested. However, this study aims to analyze primarily vSAN—the storage component—and its ability to deliver a large number of transactions at a low latency. For this reason, this study focuses on using synthetic benchmarking to gauge storage performance.

The benchmark tool used for this study is HCIBench. HCIBench is primarily a wrapper around Oracle's Vdbench, with extended functionality to deploy and configure VMs, run vSAN Observer, and aggregate data, as well as provide an ergonomic web interface from which to run tests. HCIBench can be found at the following link: https://labs.vmware.com/flings/hcibench

HCIBench is deployed as a VM template. In this case, there is a separate vSAN cluster set up for all infrastructure services, such as for HCIBench, DNS, routing, etc. The HCIBench Open Virtualization Format (OVF) template was deployed to this cluster, and a VM was created from the template. An additional virtual network was created on a separate VLAN (114), and the HCIBench VM's virtual NIC was assigned to this network to ensure that it could not send unwarranted traffic. This ensures that HCIBench traffic cannot interfere with vSAN or other traffic.

As discussed above, vSAN offers multiple options to define your storage policy. To understand how each of these affect performance, three test configurations were chosen:

Configuration	FT Method	FTT	Checksum	Dedup+Compression
Baseline	RAID-1 (Mirroring)	1	No	No
Performance	RAID-1 (Mirroring)	1	Yes	No
Density	RAID-5/6 (Erasure Coding)	1	Yes	Yes

Table 3: Storage Policies

For each configuration, five different workload profiles were run, all generating 4K random read/write mixtures. Since read and write performance differ drastically, a sweep was run across different read%/write% mixtures of 0/100, 30/70, 50/50, 70/30, and 100/0. This allows inferring approximate performance based on the deployment's specific read/write mixture goals.

Furthermore, two dataset sizes were used to show the difference in performance when the working set fits 100% in the cache tier, and one when it is too large to fit fully in cache. In this document, we describe the tests where the working set fits in the cache tier as a cache test, and the tests where the working set is spread across both cache and capacity tiers as a capacity test.

To ensure that all storage is properly utilized, it is important to distribute worker threads evenly amongst all nodes and all drives. To do this, each test creates four VMs on each node. Each VM has eight VMDKs, each either 6GB or 128GB, depending on whether it is a cache or capacity test.

Upon deployment, each configuration is initialized with HCIBench using the RANDOM option. This ensures that the VMDKs actually have readable data, instead of simply all zeros. This is particularly important when it comes to checksumming to ensure that the checksum is always calculated on non-zero data. A checksum is meaningless when your data is all zeros. Additionally, OSR is set to 100% for all tests and stripe width is left at the default value of 1, as per the vSAN policy described in the earlier section. This ensures that data is spread physically across the entire disk, instead of potentially lying in only a subset of it in a thin provisioned manner.

The table below shows the HCIBench parameters used for all cache and capacity tests.

HCIBench Test Parameters	Value
Threads Per VMDK:	4
Test Duration:	30 minutes
Rampup Duration:	1 hour
% Read:	0/30/50/70/100%
% Random:	100%
Working Set Size:	100%
Disk Initialization:	RANDOM
Clear Cache Before Testing:	Yes

Table 4: HCIBench Test Parameters

Baseline Testing

To get a set of baseline performance data, a run was executed with a storage policy consisting of RAID-1, checksum disabled, and FTT of 1. This removes the overhead from CPU-intensive policies, such as RAID-5/6, checksum and dedupe+compression. This will be the test by which we gauge each policy's reduction in performance.

Note that this policy that would not be recommended for most customers, since because disabling checksum means there is a chance that the data you read back is not the data you wrote. However, this does allow us to see just how much performance is lost by enabling checksumming.

To reiterate, each test is run with OSR of 100% to ensure that we are writing to the total amount of disk that we intend. Furthermore, all tests start with an initialization of random data using HCIBench's "prepare virtual disk before testing" option with "RANDOM

Test Results and Analysis

Test Configurations

Depending on the storage policy chosen, vSAN duplicates blocks of data over multiple hosts in different ways. For RAID-1 (Mirroring), vSAN writes two copies of data to two different nodes, and a third block to another separate host as a witness to break quorum in the case of a split cluster. The traffic of the witness object is negligible, so we see roughly 2:1 writes at the vSAN level as compared to what the VMs think they are writing.

When moving to RAID-5/6 (Erasure coding) with FTT of 1, writes happen in a 3+1 format, meaning a single block of data is split into three chunks, each written to different hosts while the fourth host gets a parity value computed from the original block. The parity can help recreate a missing block of data in the case of a node failure. This means that vSAN will write four smaller blocks of data for every one block (striped across three smaller blocks) the VMs try to write.

This is important to consider when studying performance differences between different storage policies. RAID-5/6 will write less data to the physical devices, but because the CPU must work harder to perform the parity calculations, its performance is typically lower.

Configuration	FTM	FTT	Checksum	Dedupe+Compression
Baseline	RAID-1 (Mirroring)	1	No	No
Performance	RAID-1 (Mirroring)	1	Yes	No
Density	RAID-5/6 (Erasure Coding)	1	Yes	Yes

Table 5: Tested Configurations

As mentioned previously, there is a tradeoff for each FTM. The performance configuration offers better performance, but means you need twice the raw capacity of what you need for usable data. The density configuration improves upon this, and only requires an additional 33% more raw space than you need, but at a performance penalty.

The table below shows how much additional raw storage is needed for each option. Also note that when enabling deduplication and compression, capacity can be further extended, but it is highly dependent on how compressible your data is. The table below shows the capacity multiplier for each FTM and FTT.

FTM	FTT	Capacity Multiplier
RAID-1 (Mirroring)	1	2
RAID-1 (Mirroring)	2	3
RAID-5/6 (Erasure Coding)	1	1.33
RAID-5/6 (Erasure Coding)	2	1.5

Table 6: Additional Storage (by Option)

vSAN does deduplication and compression in what they call *near-line*, and is performed in one operation while destaging from cache to capacity. During destaging, each 4K block is hashed. If that hash matches another block's hash in the capacity tier, it will simply skip that write entirely, and just write a pointer to the previously written block. If the block's hash does not match, it will try to compress the block. If the block is compressible to less than 2K, it will be written as a compressed block. If not, it will simply be written as the original uncompressed raw 4K block. If your data is incompressible or minimally compressible, enabling Dedupe+Compression will likely not offer a significant capacity benefit, and may reduce your performance. The diagram below shows how vSAN's deduplication works.

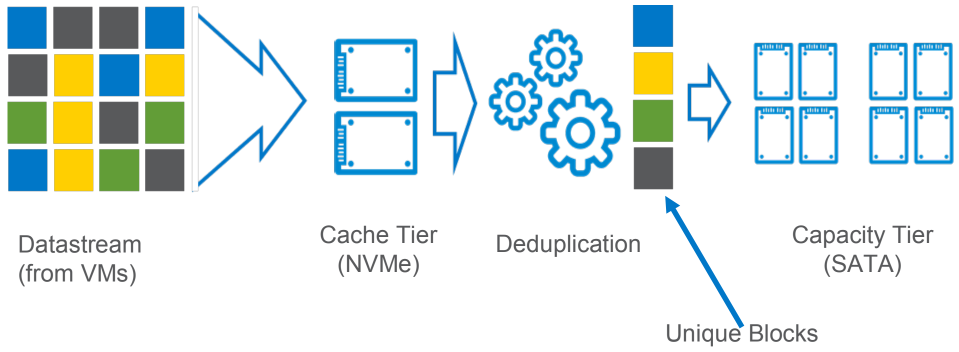
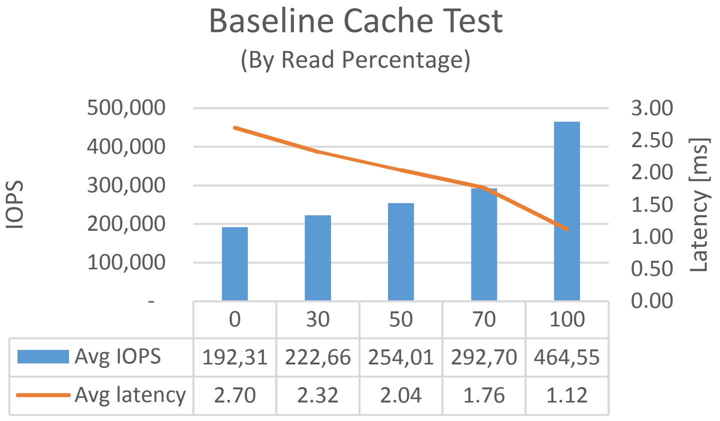


Figure 3: Data Deduplication

Performance Results: Baseline

To get a comparison point, we start with a baseline run. The following graphs show the average IOPS and latency this configuration can deliver with the baseline storage profile across each read/write mix.

Note that all test graphs show IOPS on the primary axis (left) and latency on the secondary axis (right), where the bars show IOPS, and the lines show latency.



	0	30	50	70	100
Avg IOPS	192,31	222,66	254,01	292,70	464,55
Avg latency	2.70	2.32	2.04	1.76	1.12

Tip: Leaving checksum enabled has a negligible effect on performance for cache tests. Enable it if most of your working set resides in cache.

Figure 4: Baseline Cache Test

Figure 4 shows the IOPS and latency for the baseline for each read percentage mixture. Doing a pure write test produces 192K IOPS at an average latency of 2.70ms. As more reads are added into the mix, the performance begins to increase, netting higher IOPS and lower latency. At 100% read, IOPS are up to over 464K at 1.12ms latency. This mean each node is able to deliver over 116K IOPS, which is more than double what vSAN claims an AF-6 configuration should consistently be able to serve, at 50K IOPS.

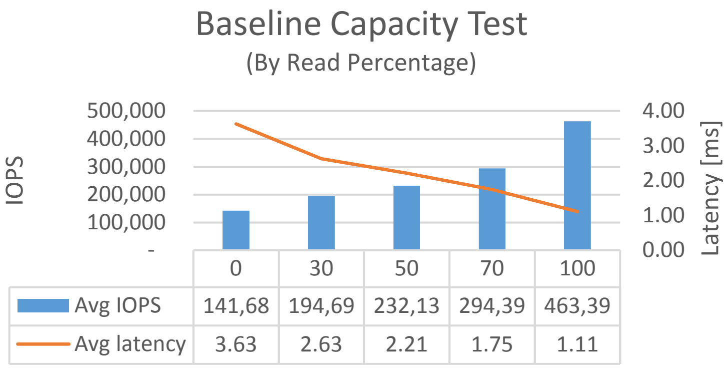


Figure 5: Baseline Capacity Test

Figure 5 shows the IOPS and latency for the baseline capacity test. We see the same trend followed as with the cache test, but with slightly lower performance, especially for the write workloads. As the reads get closer to 100% of the workload, the difference in performance becomes negligible, since all destaged reads come from the capacity tier in an all-flash configuration.

Tip: For working sets that reside in the capacity tier, checksum can be disabled if your workload has built-in checksum and error correction strategies.

Tip: Deduplication and compression have negligible performance effect if your working set resides mostly in the capacity tier. These settings should be enabled.

Above 70% reads, both tests offer practically identical performance. This means that below 30% writes, the capacity tier is able to keep up with the rate at which the cache tier tries to destage writes. Above 30%, the capacity drives begin to slow down, and can't keep up with the writes to the cache tier.

Performance Results: Cache Test

The first comparison is with a working set size that fits 100% in cache. This test eliminates any destaging actions, and increases performance for the mixed tests, since the cache tier is much more performant than the capacity tier.

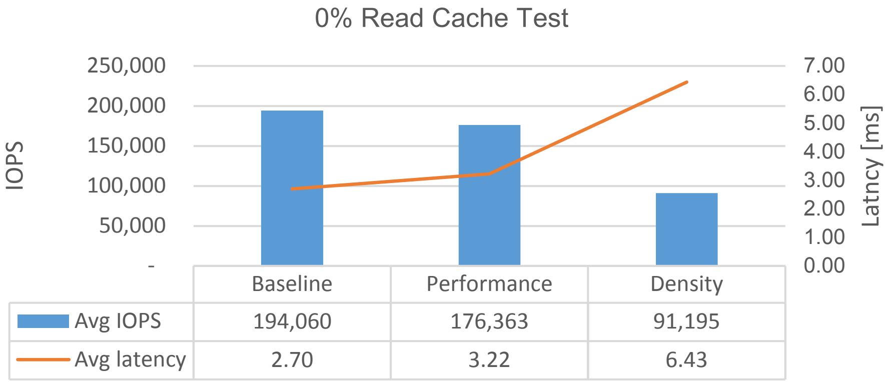


Figure 6: 0% Read Cache Test

Figure 6 shows how the performance changes for a pure write test on each storage profile. As expected, enabling checksum adds some overhead during write operations, since computing the checksum requires additional CPU cycles for each write operation. For this test, enabling checksum reduces IOPS by roughly 16% from the baseline, as well as adding approximately 19% latency. The density storage profile shows its worst-case performance here. We expect write performance to lag with RAID-5/6 and dedupe+compression. The performance of the density storage profile is about 55% lower than that of the baseline, with more than double the latency.

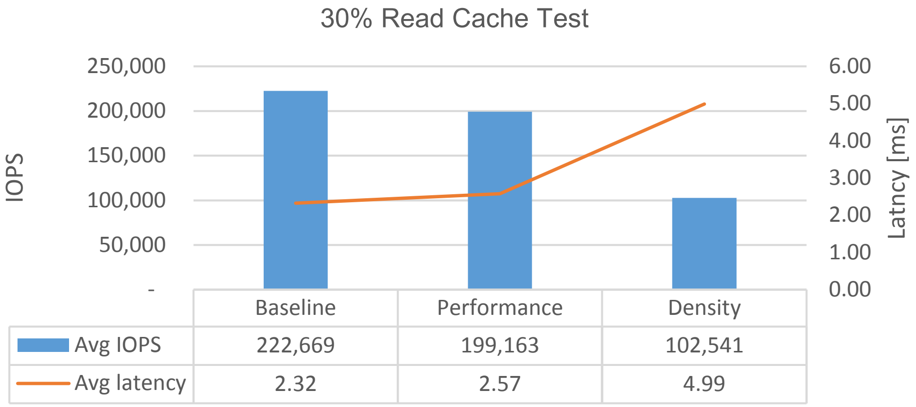


Figure 7: 30% Read Cache Test

Figure 7 shows the 30% read performance results. The trend is identical to that of the 0% read test, but with slightly higher performance. Again, density is less than half of the performance of the baseline, with more than double the latency.

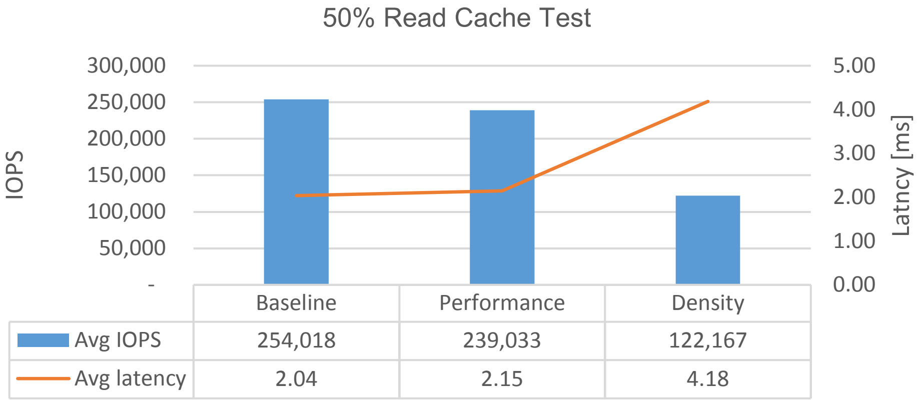
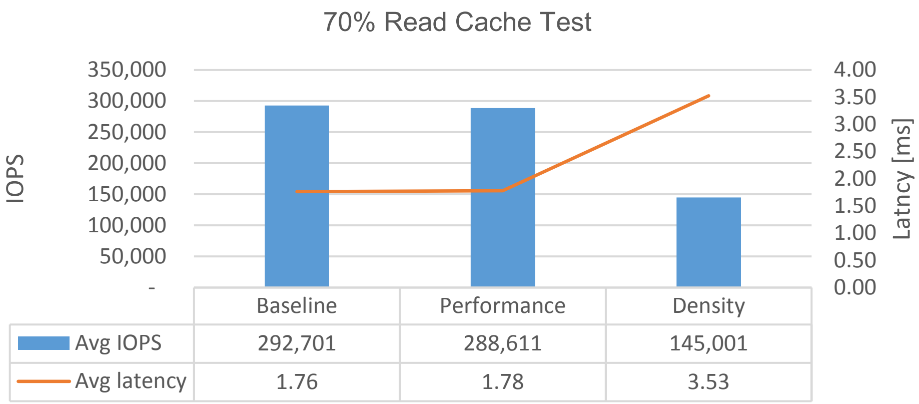


Figure 8: 50% Read Cache Test

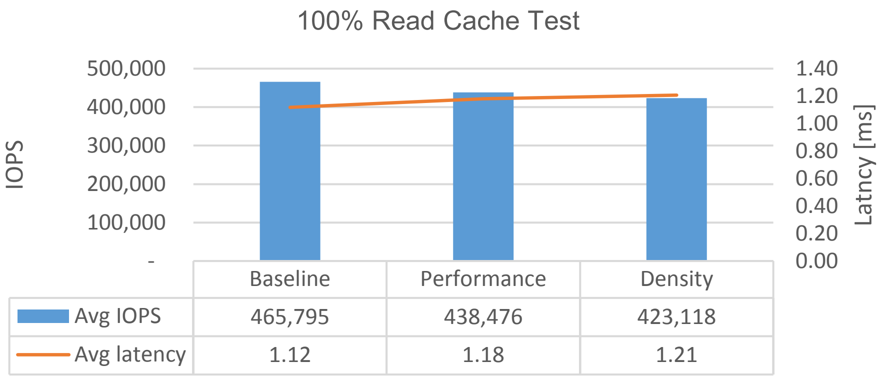
When reads view the 50% read results, we start to notice that the performance impact of enabling the checksum is diminishing. It is to be expected that enabling checksum offers less of a performance impact on read operations than on writes, especially when in cache. The density profile, however, still maintains less than half of the IOPS of the baseline, and more than double the latency. Again, this is because write operations are expensive when calculating parity. Since this is a cache test, having dedupe+compression enabled has virtually no effect on the results.



	Baseline	Performance	Density
Avg IOPS	292,701	288,611	145,001
Avg latency	1.76	1.78	3.53

Figure 9: 70% Read Cache Test

With 70% reads, it is even clearer that enabling checksum has a minimal impact on read operations, as there is now a very small difference in performance from the baseline. Yet again, density suffers at roughly half of the performance of the baseline.



	Baseline	Performance	Density
Avg IOPS	465,795	438,476	423,118
Avg latency	1.12	1.18	1.21

Figure 10: 100% Read Cache Test

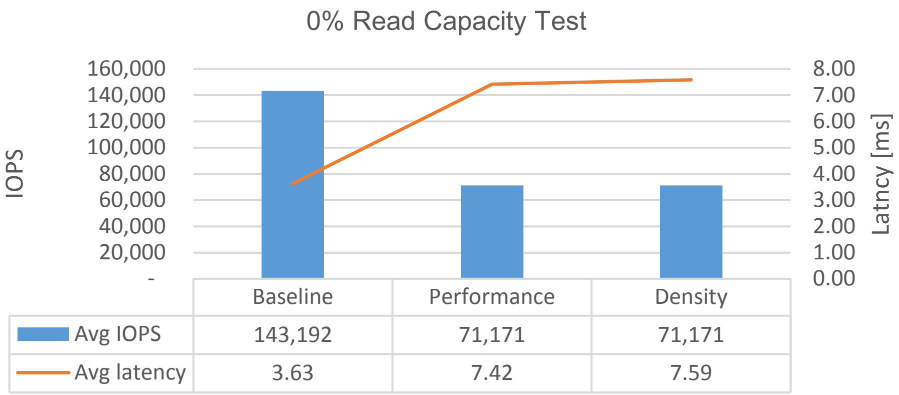
Tip: For read-heavy workloads, enabling RAID5/6 (erasure coding) reduces performance only slightly. Enable RAID5/6 when you have a read-heavy workload for additional useable storage capacity.

With 100% reads, the data tells a different story. Because checksum has a minimal impact, and there is no parity being calculated for RAID-5/6 writes, the performance between all three profiles is very similar. There is still a small difference between each, but the capacity profile now only has less than a 10% reduction in performance from the baseline.

This first study shows that if the working set fits primarily in cache, there isn't much of a downside to enabling checksum and RAID-5/6, especially if it is read-intensive. The more writes the workload requires, the less you may want to consider RAID-5/6 if performance is a requirement.

Performance Results: Capacity Test

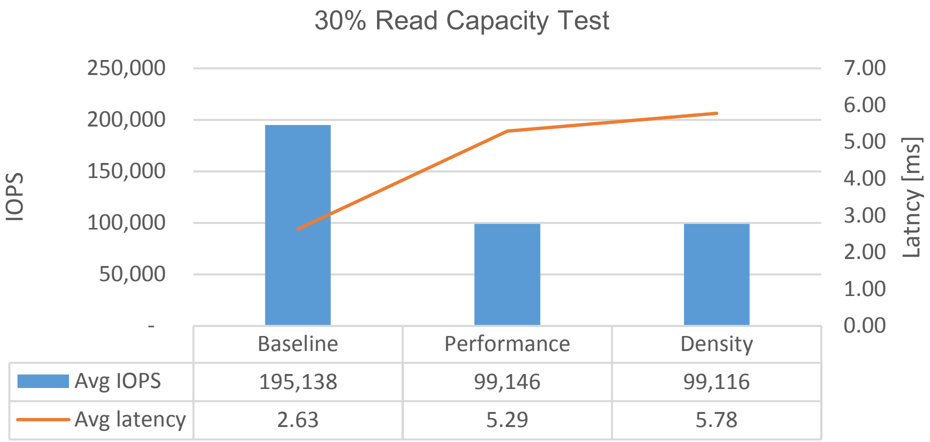
The second comparison looks at performance differences when the working set does not all fit into the cache tier. In this study, the total working set size per node is around 4TB, with each node only being able to use 600GB for cache. Therefore, about 15% of user data can reside in the cache tier, while the rest must be held in the capacity tier. Consequently, this means that there will be heavy destaging for write-intensive workloads, which will reduce performance.



	Baseline	Performance	Density
Avg IOPS	143,192	71,171	71,171
Avg latency	3.63	7.42	7.59

Figure 11: 0% Read Capacity Test

For the 100% write test, we see a drastic difference in performance between the baseline and the two other storage profiles. This indicates that checksumming is adding significant overhead. The difference between performance and density is negligible, suggesting that the overhead of the parity calculations is dwarfed by that of the checksumming. It is odd to note that checksum seems to have a larger impact on a large working set than it does on a small one. This is contrary to what is to be expected, since it is expected that the checksum happens while the data is first written to the cache layer. This behavior is, however, consistent with the vSAN 6.6 performance study released by VMware. Here, the performance for both the performance and density profiles are roughly half of the baseline performance.



	Baseline	Performance	Density
Avg IOPS	195,138	99,146	99,116
Avg latency	2.63	5.29	5.78

Figure 12: 30% Read Capacity Test

With 30% read transactions, we see the same trend as with 0% reads, but with higher numbers across the board. Performance and density profiles still show roughly half of the performance of the baseline.

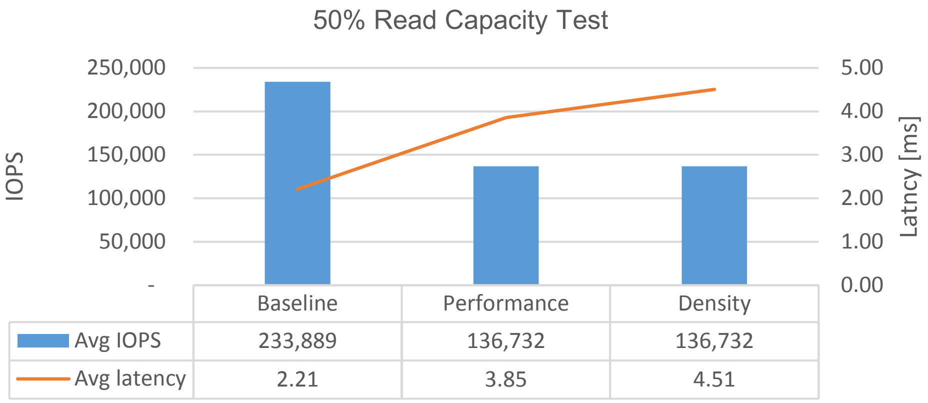


Figure 13: 50% Read Capacity Test

Once the read/write mix is at 50%, we notice that the reduction in performance from the performance and density profiles seems to be diminishing. Each are now roughly 60% of the baseline performance, instead of 50%, as seen by the previous two scenarios. We notice that the density has slightly higher latency than the performance profile, though the percentage difference is much smaller than the difference between the baseline and performance profile.

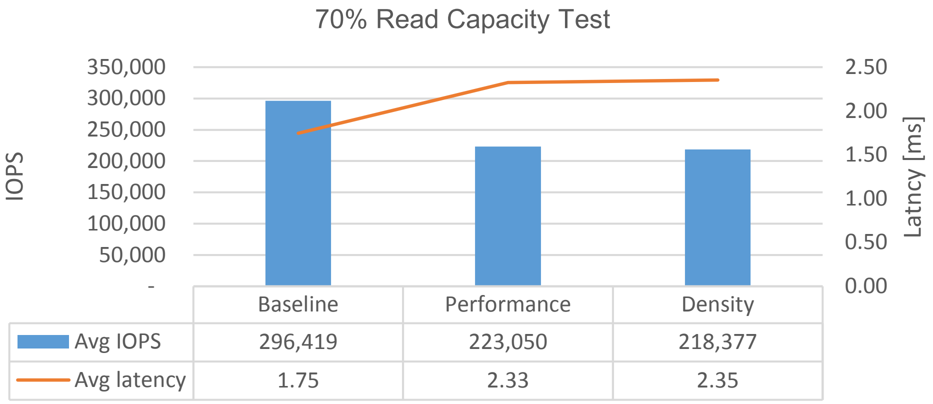
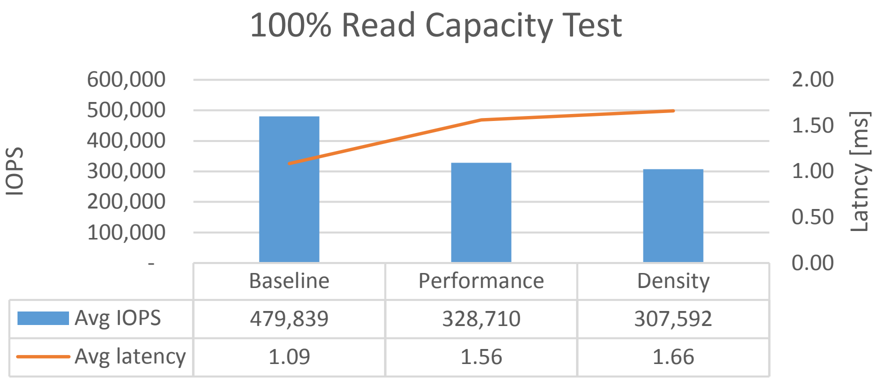


Figure 14: 70% Read Capacity Test

Here, we see that the difference in performance between each profile has reduced even further—now at roughly 75% of the baseline performance. We expect that as we get more and more reads, the performance implications of the writes get reduced, and performance gets closer to the baseline.



	Baseline	Performance	Density
Avg IOPS	479,839	328,710	307,592
Avg latency	1.09	1.56	1.66

Figure 15: 100% Read Capacity Test

Tip: For read-heavy workloads, enabling RAID5/6 (erasure coding) reduces performance only slightly. Enable RAID5/6 when you have a read-heavy workload for additional useable storage capacity.

With 100% reads, the baseline performance is nearly identical to that of the cache baseline test. This makes sense, considering that reads always come from the capacity tier after the blocks have been destaged. What is interesting is that we still see performance implications in pure reads with checksum enabled. This is because the checksum must be computed again when reading to ensure that the data matches the checksum.

Additional Information and Resources

Tuning Parameters

vSAN’s default tunings are setup to be safe for all users. When doing heavy write tests, a disk group can quickly run out of memory and run into memory congestion, causing a decrease in performance. To overcome this, we followed VMware’s performance document to alter three advanced configuration parameters. The table below shows the default value and the value this configuration used, as well as the documents referenced for the tunings.

Parameter	Default	Tuned
/LSOM/blPLOGCacheLines	128K	512K
/LSOM/blPLOGLsnCacheLines	4K	32K
/LSOM/blLLOGCacheLines	128	32K

Table 7: Tuning Parameters

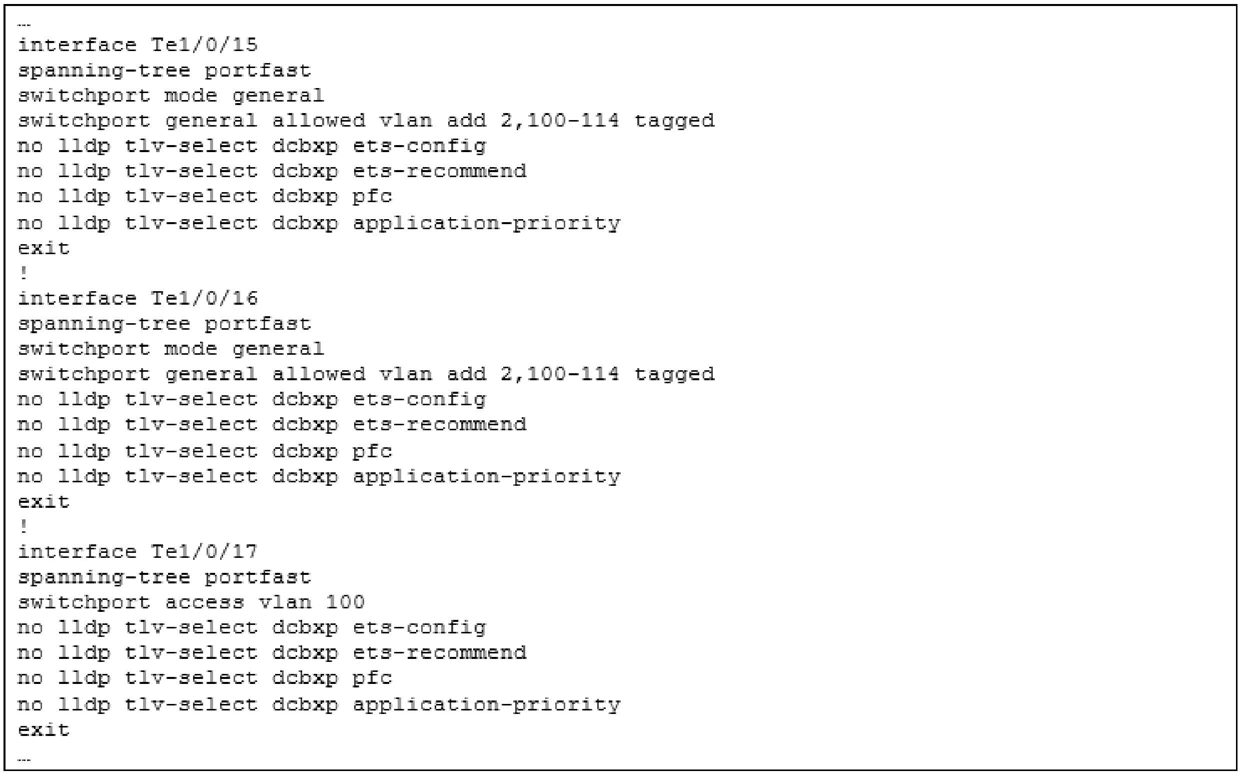
https://storagehub.vmware.com/#!/vmware-vsan/vsan-6-6-performance-improvements
https://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2150012

VDbench param file

Below is a sample VDbench parameter file for a 0% read test against 8 VMDKs with a run time of 30 minutes and a warmup(ramp) time of one hour.

```
*Auto Generated VDBench Parameter File
*8 raw disk, 100% random, 0% read
*SD:    Storage Definition
*WD:    Workload Definition
*RD:    Run Definition
debug=86
data_errors=10000
sd=sd1,lun=/dev/sda,openflags=o_direct,hitarea=0,range=(0,100),threads=4
sd=sd2,lun=/dev/sdb,openflags=o_direct,hitarea=0,range=(0,100),threads=4
sd=sd3,lun=/dev/sdc,openflags=o_direct,hitarea=0,range=(0,100),threads=4
sd=sd4,lun=/dev/sdd,openflags=o_direct,hitarea=0,range=(0,100),threads=4
sd=sd5,lun=/dev/sde,openflags=o_direct,hitarea=0,range=(0,100),threads=4
sd=sd6,lun=/dev/sdf,openflags=o_direct,hitarea=0,range=(0,100),threads=4
sd=sd7,lun=/dev/sdg,openflags=o_direct,hitarea=0,range=(0,100),threads=4
sd=sd8,lun=/dev/sdh,openflags=o_direct,hitarea=0,range=(0,100),threads=4
wd=wd1,sd=(sd1,sd2,sd3,sd4,sd5,sd6,sd7,sd8),xfersize=4k,rdpct=0,seekpct=100
rd=run1,wd=wd1,iorate=max,elapsed=1800,warmup=3600,interval=30
```

Switch Configuration (Sample Subset)



```
…
interface Te1/0/15
spanning-tree portfast
switchport mode general
switchport general allowed vlan add 2,100-114 tagged
no lldp tlv-select dcbxp ets-config
no lldp tlv-select dcbxp ets-recommend
no lldp tlv-select dcbxp pfc
no lldp tlv-select dcbxp application-priority
exit
!
interface Te1/0/16
spanning-tree portfast
switchport mode general
switchport general allowed vlan add 2,100-114 tagged
no lldp tlv-select dcbxp ets-config
no lldp tlv-select dcbxp ets-recommend
no lldp tlv-select dcbxp pfc
no lldp tlv-select dcbxp application-priority
exit
!
interface Te1/0/17
spanning-tree portfast
switchport access vlan 100
no lldp tlv-select dcbxp ets-config
no lldp tlv-select dcbxp ets-recommend
no lldp tlv-select dcbxp pfc
no lldp tlv-select dcbxp application-priority
exit
…
```

Monitoring and Performance Measurement Tools

HCIBench: HCIBench is developed by VMware, and is a wrapper around many individual tools, such as vSAN Observer, VDbench and Ruby vSphere Console (RVC). HCIBench allows you to create VMs, configure them, run VDbench files against each VM, run vSAN observer and aggregate the data at the end of the run into a single results file.

vSAN Observer: vSAN observer is built-in to the VMware vCenter Server® Appliance™ (VCSA), and can be enabled via the Ruby vSphere Console (RVC). HCIBench starts an observer instance with each test, and stores it alongside the test results files.

VDbench: VDbench is developed by Oracle, and is a synthetic benchmarking tool. It allows you to create workloads for a set of disks on a host, and specify parameters such as run time, warmup, read percentage, and random percentage.

Ruby vSphere Console (RVC): RVC is built-in to the vSphere Center Appliance as an administration tool. With RVC, you can complete many of the tasks that can be done through the web-GUI, and more, such as start a vSAN Observer run.

vSPhere Performance Monitoring: vSphere now has a great deal of performance metrics built right into the VCSA, including front-end and back-end IOPS and latency.

vSAN Node BOM

Component	Qty per Node	Part Number	Description
Server	1	SYS-2028U-TNRT+	Supermicro 2U Ultra Server
CPU	2	P4X-DPE52650V4-SR2N3(?)	E5-2650V4 12 core 2.2GHz
Memory	16	MEM-DR432L-CL02-ER24	Micron 16GB DDR4-2400MHz RDIMM ECC
Boot Drive	1	SSD-DM016-SMCMVN1	16GB SLC SATADOM
NVMe SSD	2	MEM-DR416L-CL03-ER24	Micron 9100 MAX NVMe 1200GB SSD
Networking (NIC)	1	AOC-MTGN-i2S	Intel 82599EB 10GbE SFP+ Dual port

Table 8: vSAN Nod BOM

About Component Suppliers

About Micron

Micron Technology (Nasdaq: MU) is a world leader in innovative memory solutions. Through our global brands — Micron, Crucial®, and Ballistix® — our broad portfolio of high-performance memory technologies, including DRAM, NAND, NOR Flash and 3D XPoint™ memory, is transforming how the world uses information. Backed by more than 35 years of technology leadership, Micron's memory solutions enable the world's most innovative computing, consumer, enterprise storage, data center, mobile, embedded, and automotive applications. To learn more about Micron Technology, Inc., visit micron.com.

About Supermicro

Supermicro (NASDAQ: SMCI), a leading innovator in high-performance, high-efficiency server technology is a premier provider of advanced server Building Block Solutions® for Data Center, Cloud Computing, Enterprise IT, Hadoop/Big Data, HPC and Embedded Systems worldwide. Supermicro is committed to protecting the environment through its “We Keep IT Green®” initiative and provides customers with the most energy-efficient, environmentally-friendly solutions available on the market.

About VMware

VMware (NYSE: VMW), a global leader in cloud infrastructure and business mobility, helps customers realize possibilities by accelerating their digital transformation journeys. With VMware solutions, organizations are improving business agility by modernizing data centers and integrating public clouds, driving innovation with modern apps, creating exceptional experiences by empowering the digital workspace, and safeguarding customer trust by transforming security. With 2016 revenue of $7.09 billion, VMware is headquartered in Palo Alto, CA and has over 500,000 customers and 75,000 partners worldwide. www.vmware.com

Benchmark software and workloads used in performance tests may have been optimized for performance on specified components and have been documented here where possible. Performance tests, such as HCIbench, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Rev. A 10/17 CCM004-676576390-10868