A **Micron** Reference Architecture

Micron® Accelerated All-Flash SATA vSAN™ 6.7 Solution

Reference Architecture

Collin Murphy, Storage Solutions Engineer
Doug Rollins, Principal Technical Marketing Engineer

Contents

Executive Summary

This reference architecture (RA) describes an example configuration of an all-flash VMware vSAN™ platform combining two classes of SATA SSDs (one in its cache tier and one in its capacity tier) into standard, x86 architecture rackmount servers with 10 Gb/E networking.

The combination of SATA SSDs with standard servers provides an optimal balance of performance and cost. Similar to an AF-6 configuration, this VMware vSAN 6.7 all-flash reference design enables:

- Fast deployment: The configuration has been pre-validated and is thoroughly documented to enable fast deployment.
- Balanced design: The right combination of cache and capacity SSDs, DRAM, processors and networking ensures subsystems are balanced and performance-matched.
- Broad deployment: Complete tuning and performance characterization across multiple IO profiles enables broad deployment across multiple workloads

This RA details the hardware and software building blocks and measurement techniques used to characterize the RA's performance as well as its composition, including the vSphere and network switch configurations, vSAN tuning parameters, Micron reference nodes and Micron SSD configurations.

The configuration in this RA ensures easy integration and operation with vSAN 6.7, offering predictably high performance that is easy to deploy and manage, providing a pragmatic blueprint for administrators, solution architects and IT planners to build and tailor a high-performance vSAN infrastructure that scales for I/O-intensive workloads.

This RA focuses on SATA SSDs. We offer other RAs optimized for performance, cost, and/or density at the Micron Accelerated Solutions site.

Note: The performance shown was measured using the components noted. Different component combinations may yield different results.

Micron's Reference Architectures

Micron Reference Architectures are optimized, pre-engineered, enterprise-leading platforms developed by Micron with industry leading hardware and software companies.

Designed and tested at Micron's Storage Solutions Center by our software and platform partners, these best-in-class solutions enable end users, channel participants, independent software vendors (ISVs), and OEMs to have a broader choice in deploying next-generation solutions with reduced time investment and risk.

Solution Overview

A vSAN storage cluster is built from a number of vSAN-enabled vSphere® nodes for scalability, fault-tolerance, and performance. Each node is based on commodity hardware and uses vSAN 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 stored within 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 VM writes to its VMDK, vSAN handles all necessary operations such as data duplication, erasure coding, checksum, and placement based on the selected storage policy.

Storage policies can be applied to the entire datastore, a VM, or a VMDK. Using storage policies enables a user to add more performance, capacity, or availability to an object. Numerous storage policies can be used on the same datastore, enabling creation of 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.

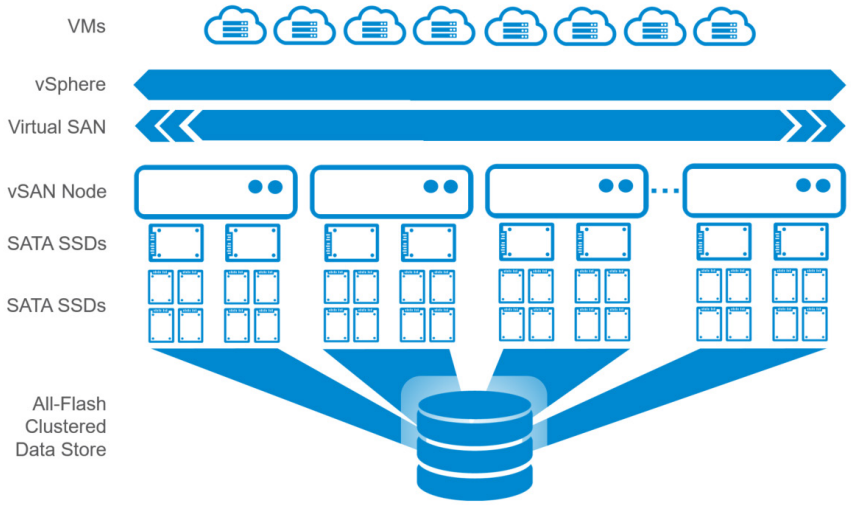


Figure 1: vSAN Architecture

Why Micron for this Solution

Storage (SSDs and DRAM) can represent up to 80% 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.

VMs write to vSAN VMDKs, while the vSAN algorithms determine how data is distributed across physical disks, depending on the storage policy for that VMDK. Below are some of the options that make up a storage policy.

- **Primary levels of failures to tolerate (FTT)**: 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: 1) RAID-1 (Mirroring), and 2) RAID-5/6 (Erasure coding). RAID-1 (Mirroring) creates duplicate copies of data in the amount of 1 + FTT. RAID-5/6 (Erasure coding) stripes 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)**: 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 (in the case where performance is more important than data integrity). The default value is No.
- **Number of disk stripes per object (DSPO)**: The number of objects over which a single piece of data is striped. This applies to the capacity tier only (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 guarantees they reside on different disks on the host, if possible.

Design Overview

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

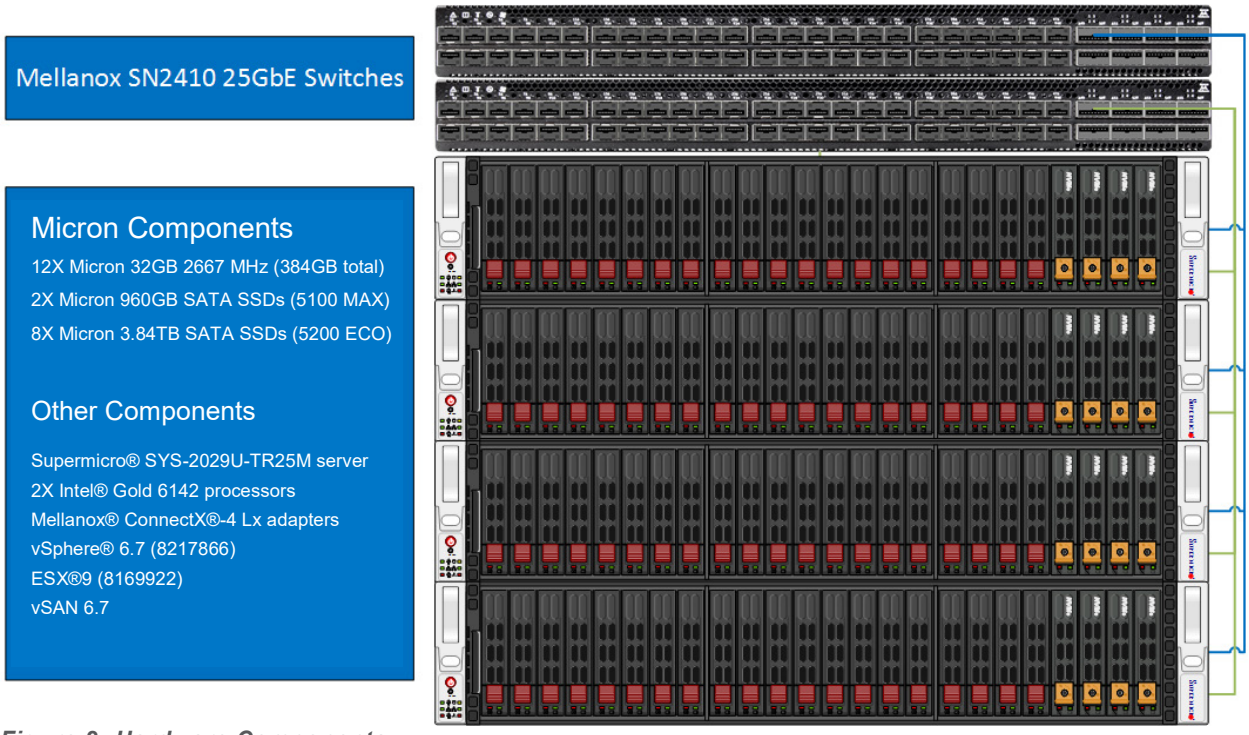


Figure 2: Hardware Components

Software

VMware's vSAN is an industry leading hyper-converged infrastructure (HCI) solution, combining traditional virtualization with multihost software-defined storage. vSAN is technology part of VMware's vSphere environment, coupled with the ESXi type-1 hypervisor.

We chose vSAN 6.7 for this solution because of its new capabilities and updates. According to VMware documentation, these include:

- **Unified management**: vSAN 6.7 uses a simplified HTML 5-based user interface for a consistent experience with other VMware products
- **FIPS 140-2:** vSAN offered the first native HCI encryption solution for data-at-rest, and now with vSAN 6.7, vSAN Encryption is the first FIPS 140-2 validated software solution
- **Proven data reduction:** Providing as much as 7X data reduction[1] (according to VMware's vSAN 6.7 datasheet)
- **Application resiliency:** Intelligent, self-healing capabilities that include adaptive resynchronization, fast failovers for air gapped networks, and replica consolidation
- **Support for business-critical applications**: vSAN now supports more mission-critical application deployments through Windows Server Failover Clusters (WSFC) support

[1] Assumes deployment enables 7X data reduction; actual data reduction is dependent on several external factors. See https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/products/vsan/vmware-vsan-datasheet.pdf for additional details on data reduction and https://www.vmware.com/products/vsan/whats-new.html for details on what's new in 6.7.

In accordance with VMware documentation, 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 storage architecture, where all write operations are sent to the cache tier and are subsequently de-staged to the capacity tier over time. Up to 600GB of cache tier storage can be utilized per disk group, with up to five disk groups per host.

vSAN can operate in two modes:

- **Hybrid**: SSDs for the caching tier and rotating media for the capacity tier
- **All-Flash**: 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 to improve hybrid design performance. In this configuration, 70% of the cache tier capacity is dedicated to the read cache and the remaining 30% is dedicated to the write buffer.

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

Micron Components

This RA employs Micron's 5100 MAX and 5200 ECO enterprise SATA SSDs (MAX for the cache tier and ECO for the capacity tier). This solution also utilizes Micron DRAM (the details of which are not discussed further in this document).

SSD	Use	Random Read	Random Write	Read Throughput	Endurance (TBW)
5100 MAX	Cache Tier	93,000 IOPS	74,000 IOPS	540 MB/s	8.8 PB
5200 ECO	Capacity Tier	95,000 IOPS	17,000 IOPS	540 MB/s	6.4 PB

Table 1: Micron SSDs

See www.micron.com for additional details, specifications and datasheets for these and other Micron SSD products.

Server Platforms

This RA utilizes standard rackmount 2U dual-socket Intel-based servers (Supermicro SYS-2029U-TR25M). Each server is configured with two Intel Gold 6142 processors, each with 16 cores at 2.60 GHz. These processors align with VMware's AF-6 minimum requirements (VMware's nomenclature for a medium-sized all-flash configuration).

Switches

vSAN utilizes commodity Ethernet networking hardware. This RA uses two Mellanox SN2410 switches for all cluster-related traffic. Both switches are interconnected by a single QSFP+ cable. Spanning Tree is enabled to avoid loops in the network. All ports are configured in general mode, with VLANs 100-115 allowed. Each server is connected via a QSFP+ quad-port breakout cable enabling 25 Gb/s throughput at each server network port.

In accordance with VMware's best practices, vSAN should have at least three separate logical networks that are all segregated using different VLANs and subnets on the same switches. The three networks in this RA, and their respective VLANs, are:

- Management/VM network: VLAN 100, subnet 172.16.17.X/16
- vMotion: VLAN 101, subnet 192.168.1.X/24
- vSAN: VLAN 102, subnet 192.168.2.X/24

While using different subnets or VLANs alone would suffice, adding both ensures 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.

Tip: Networking

Use different subnets and VLANs to ensure each network has its own separate broadcast domain (even if an interface is configured with an incorrect VLAN or IP address).

Connect each node to both switches to ensure availability.

Network Interface Cards

Each server has a single dual-port Mellanox MT27710 ConnectX-4 Lx EN 25 Gb/E NIC, with one port of each NIC 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.

The tables below summarize the hardware components used in this RA. If other components are substituted, results may vary from those described.

Hardware Components

Node Components	
2U, 2-socket standard rack mount server*	1X 480GB Micron Enterprise SATA SSD (OS boot drive)
2X Intel Xeon Gold 6142 16-core 2.60GHz CPUs	2X LSI 3108 SAS/SATA HBA
Micron 384GB 2666 MHz DRAM (32GB x 12)	1x Mellanox ConnectX-4 Dual-port 25GbE SFP28 NIC (MCX4121A-XCAT)
2X Micron 960GBSATA SSDs (5100 MAX)	
8X Micron 3.84TB SATA SSDs (5200 ECO)	

*Supermicro SYS-2029U-TR25M tested, other platforms may give different results)

Table 2: Components

Network Infrastructure

Network Components	
2X Mellanox SN2410 25GbE switches	2X Mellanox QSFP28 to 4x SFP28 Copper Breakout Cables

Table 3: Networking

Software Components

Software Components	
Server BIOS version 2.0b	Disk Format version 6
vCenter Server Appliance 6.7.0.10000 build 8217866	HBA driver 7.702.13.00-4vmw.670.0.0.8169922
ESXi build 8169922	HBA firmware 24.21.0-0015
vSAN 6.7	5100 firmware D0MU410
	5200 firmware D1MU404

Table 4: Software

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 RA targets a configuration based on VMware's AF-6 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, two disk groups per node with eight capacity drives per node, and 10 Gb/E networking minimum. For more information on AF-6 requirements, see VMware's vSAN Hardware Quick Reference Guide

This configuration utilizes two disk groups per node, with four capacity drives per disk group, resulting in two cache drives and eight capacity drives per node.

It is important to note there are many ways in which performance can be increased, but they all come 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 with having to buy additional cache drives. Furthermore, adding faster networking—like 40 Gb/E, 100 Gb/E or Infiniband—would potentially yield better performance, but all of the necessary hardware to do so would again add significant cost to the solution. The solution chosen for this RA is moderately sized for good performance at a balanced price point.

Measuring Performance

Test Methodology

Benchmarking virtualization can be a challenge because of the many different system components that can be tested. However, this RA focuses on vSAN's storage component and its ability to deliver a large number of transactions at a low latency. For this reason, this RA 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 is deployed as a VM template. In this case, there is a separate vSAN-enabled 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 (115), and the HCIBench VM's virtual NIC was assigned to this network to ensure it could not send unwarranted traffic.

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

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

Table 5: Storage Policies

For each configuration, five different workload profiles were run, all generating 4K random read/write mixtures. Since read and write performance differs 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 (or preconditioned) with HCIBench using a 128K sequential write test that is run sufficiently long to ensure the entire dataset is written over twice. This ensures the VMDKs 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 the data is all zeros. Additionally, OSR is set to 100% for all tests—except for the density profile—and stripe width is left at the default value of 1 as per the vSAN policy described earlier. This ensures that data is spread physically across the entire usable space of each disk, instead of potentially lying in only a subset of disks, in a thin provisioned manner.

When benchmarking storage utilities, it is important to ensure consistent and repeatable data. This means ensuring every test is run the same way, under the same conditions. Many things should be considered to ensure repeatable results. Each test must start in the same state, which is why we select the **clear read/write cache before testing** option in HCIBench. We also allow each test to get to steady state performance before we start our performance measurements. Steady state is found by running a test, monitoring performance, and seeing when it becomes stable. For all tests conducted in this paper, the time to reach steady state was approximately two hours—called ramp up time or duration. After ramp up, performance data is captured over a long enough time period to ensure that a good average is collected, while not collecting too long, since many runs need to be conducted. For our testing, the data capture period is one hour.

The table below shows the HCIBench parameters used for all tests. We also selected four threads per VMDK based on prior work, as four threads seemed to provide consistently high IOPS without excessive latency (see later figures for latency measurements).

The table below summarizes all run options used for testing.

HCIBench Test Parameters	Cache	Capacity
Threads Per VMDK:	4	
Test Duration:	1 hour	
Rampup Duration:	2 hours	
% Read:	0/30/50/70/100%	
% Random:	100%	
Working Set Size:	100%	
Disk Initialization:	128K sequential	
Clear Cache Before Testing:	Yes	

Table 6: HCIBench Test Parameters

Storage Policies

Depending on the storage policy chosen, vSAN duplicates blocks of data over multiple hosts differently. For RAID-1 (Mirroring), vSAN writes two copies of data to two different hosts, 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 using 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 attempt 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.

Deduplication and Compression Testing

vSAN performs deduplication and compression in what they call near-line, and it is performed in one operation while de-staging from cache to capacity. During de-staging, each 4K block is hashed. If that hash matches another block's hash in the capacity tier, it will skip that write entirely, and simply 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 deduplication and compression will likely not offer a significant capacity benefit, and may reduce performance. Figure 3 illustrates vSAN's deduplication.

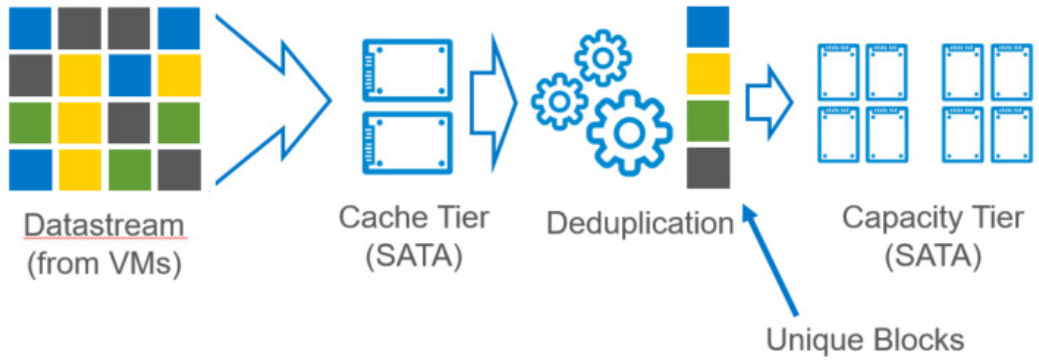


Figure 3. Data Deduplication

Testing deduplication and compression is slightly different from testing other profiles. Deduplication and compression offers no benefit if your data is not compressible. For this reason, the dataset must be compressible instead of purely random.

HCIBench utilizes Vdbench as its load-generating tool, which supports options for duplicable and compressible datasets. While HCIBench itself does not give options to configure deduplication and compression, it is easy to directly modify the Vdbench parameter files to do so. Appendix A details the modifications to the parameter files used in this RA. The settings used resulted in an approximately 8.21X combined deduplication and compression ratio for the capacity test and 2.56X for the cache test, since

de-staging the working set size for the cache test is relatively small compared to the size of the vSAN file system (see Appendix C for details on deduplication and compression ratio.) To get meaningful results, OSR was set to 0% for the density profile (otherwise, the deduplication and compression factor is not measurable by vSAN since it will reserve 100% of the raw capacity, regardless of how much of it gets utilized).

Baseline Testing

To get a set of baseline performance data, a test 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 deduplication and compression. This will be the test by which we gauge each policy's reduction in performance.

Note that this policy would not be recommended for most uses, since disabling checksum means there is a chance of getting a bit error and not being able to detect it. However, this does allow us to see just how much performance is lost by enabling the checksum and other additional features.

Each test—except for the density profile—is run with OSR of 100% to ensure we are writing to the total amount of disk that we intend. Furthermore, all tests start with an initialization of random data by running a 128K sequential write test.

Test Results and Analysis

Each FTM has tradeoffs. The performance configuration offers better performance, but requires twice the capacity the data set occupies. The density configuration improves upon this, requiring an additional 33% more space than the data set occupies, 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	Raid Level	Data Copies	Capacity Multiplier
RAID-1 (Mirroring)	1	RAID-1	2	2
RAID-1 (Mirroring)	2	RAID-1	3	3
RAID-5/6 (Erasure Coding)	1	RAID-5	3+1p	1.33
RAID-5/6 (Erasure Coding)	2	RAID-6	4+2p	1.5

Table 7: Additional Storage (By Option)

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 vertical axis (left), latency on the secondary vertical axis (right) and read percentage on the horizontal axis. The bars show IOPS; the lines show latency.

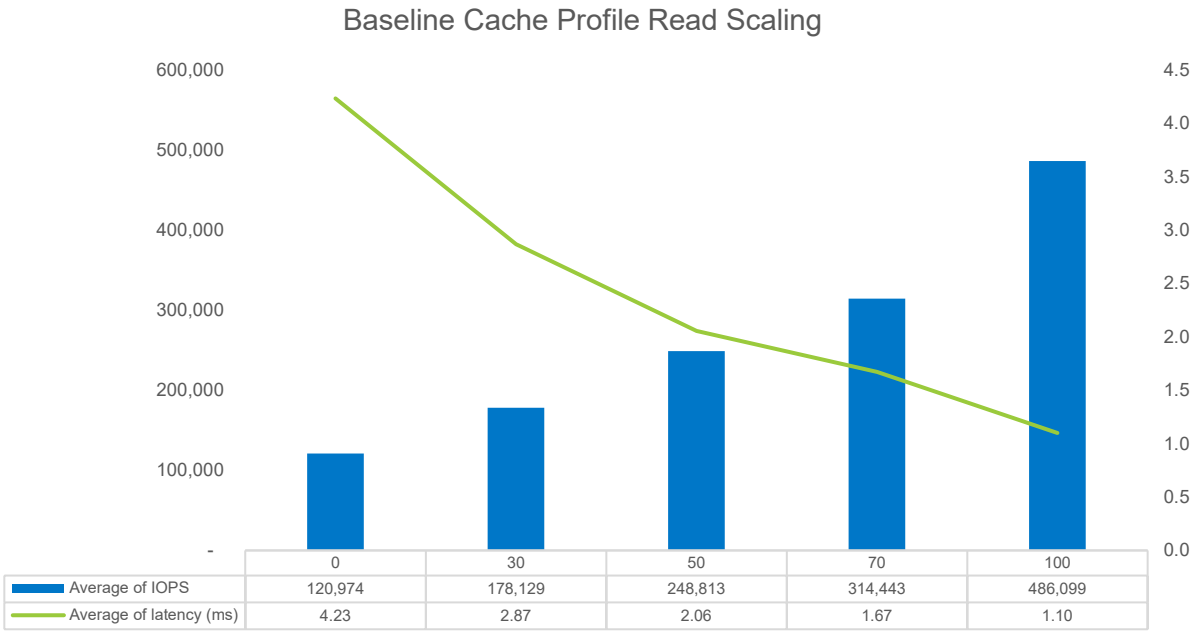


Figure 4: Baseline Cache Test

Figure 4 shows the IOPS and latency for the baseline for each read percentage mixture. Doing a pure write test produces 121K IOPS at an average latency of 4.23ms. 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 486K at 1.10ms latency. This mean each node can deliver over 120K IOPS, which is 143% more than what vSAN claims an AF-6 configuration should consistently be able to serve, at 50K IOPS.

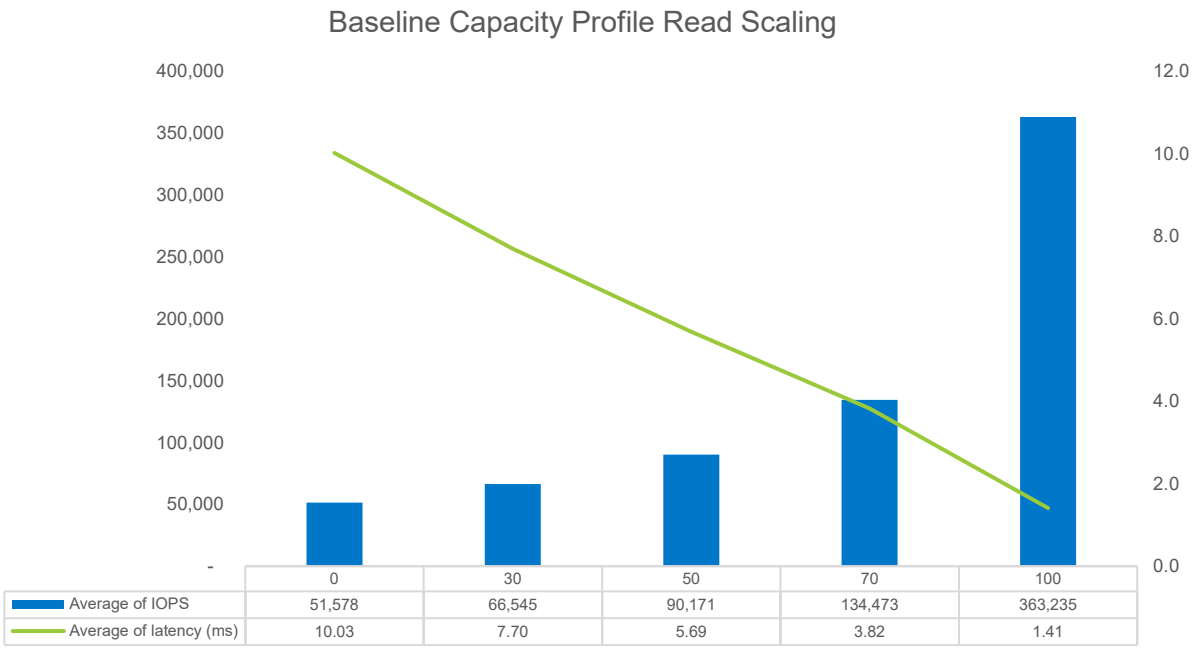


Figure 5: Baseline Capacity Test

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 smaller, since all de-staged reads come from the capacity tier in an all-flash configuration.

At 100% reads, the capacity test shows much less of a performance difference from the cache test. Read caching (in memory) is a large contributor to this observation, as vSAN dedicates a small amount of memory in each host for caching some data. The smaller working set size you use, the more apparent this feature will be, and the higher read performance will be realized.

Performance Results: Cache Test

The first comparison is with a working set size that fits 100% in cache (**cache test**). This test eliminates most de-staging actions, and increases performance for the mixed tests, since the cache tier is much more performant than the capacity tier.

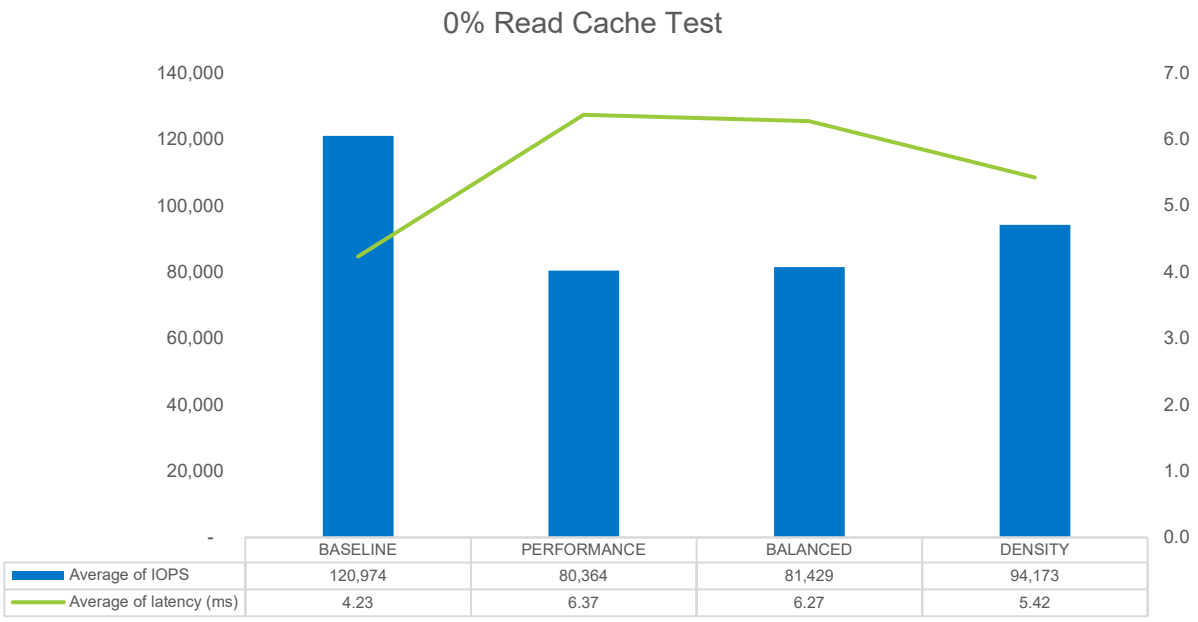


Figure 6: 0% Read Cache Test

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 23% from the baseline, as well as adding approximately 50% additional latency.

The balanced profile, which utilizes RAID-5/6, shows a negligible performance difference with 1% higher IOPS and 2% lower latency than the performance configuration. We typically expect write performance to suffer with enabling RAID-5/6, since parity calculations will be performed, but it appears to be negligible in this case. This means that the added latency associated with the calculation is very small compared to the disk write latency.

The density profile produces about 15% higher IOPS and 14% lower latency than the balanced profile. Typically, enabling deduplication and compression adds some CPU overhead. However, just as seen when enabling RAID-5/6, the CPU requirement has a small impact on the overall latency here, with the disk write latency being a larger contributor to the overall latency. Because the deduplication and compression ratio is so large, there is very little de-staging going on, and thus more writes are absorbed by the cache tier.

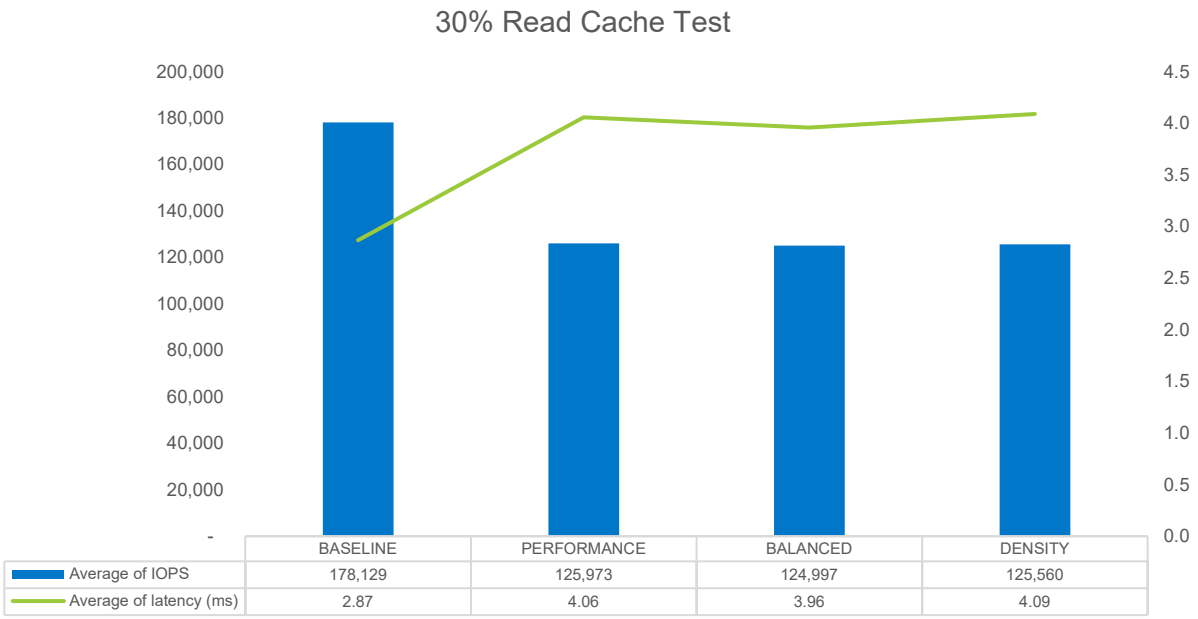


Figure 7: 30% Read Cache Test

Adding 30% reads into the mix shows an increase in performance on all profiles (as expected). The performance profile shows a 30% reduction in IOPS and a 41% increase in latency. Switching to RAID-5/6 in the balanced profile shows a negligible difference with less than 1% fewer IOPS and less than 3% increase in latency. Last, enabling deduplication and compression gives almost identical performance. The performance, balanced, and density profiles give essentially identical performance for this test, as the differences are statistically negligible.

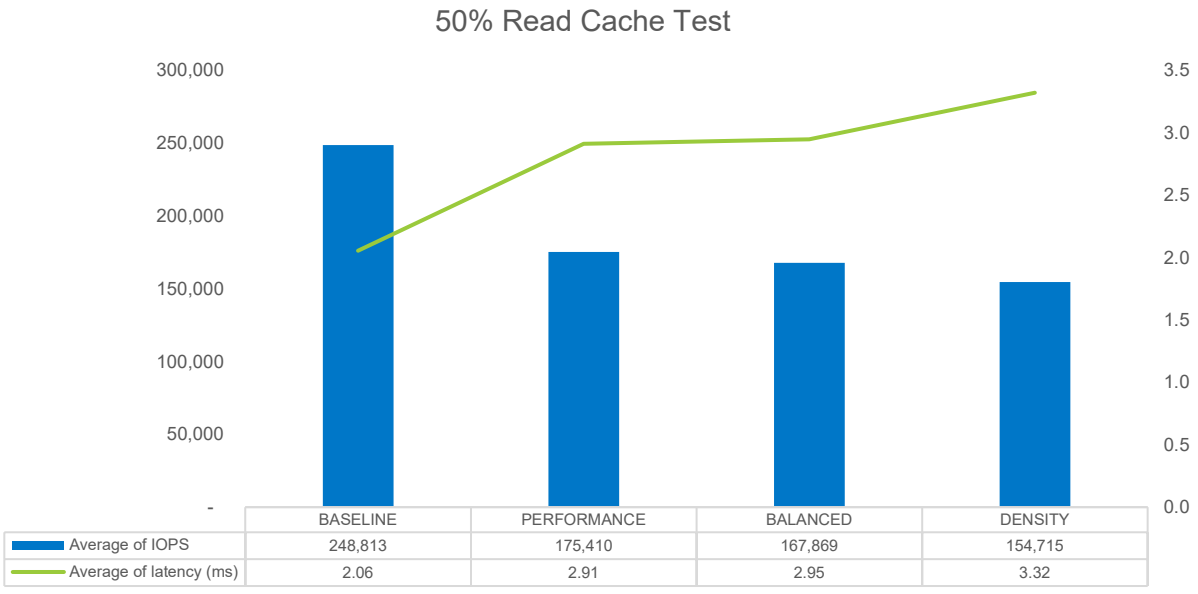
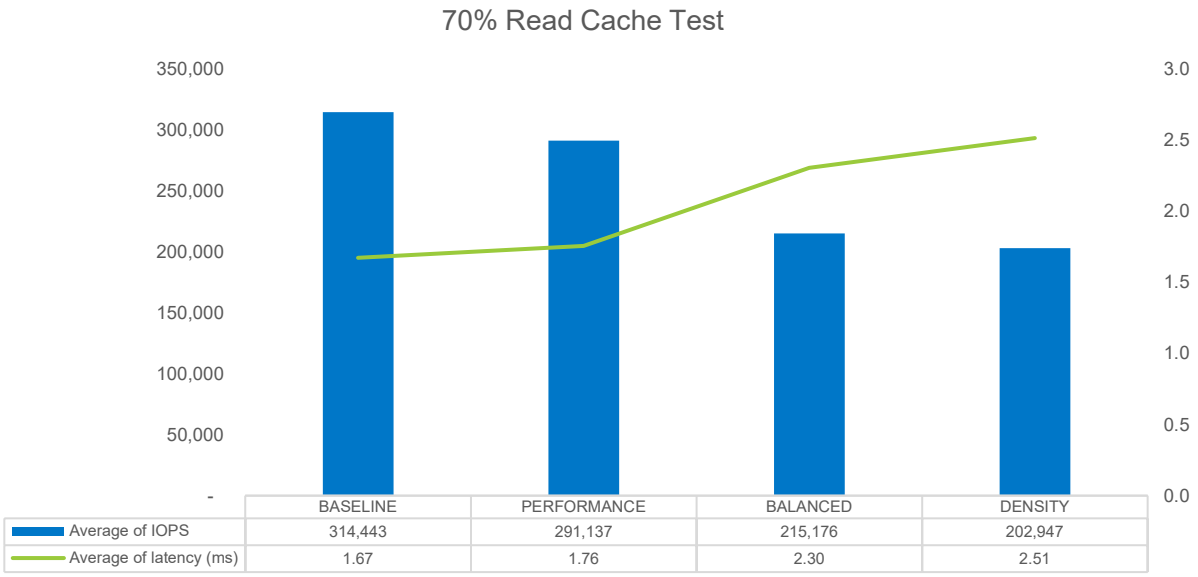


Figure 8: 50% Read Cache Test

At 50% writes, performance is again higher for all profiles, and the difference between each profile becomes more apparent. The performance profile shows 30% less IOPS and 42% higher latency than the baseline. The balanced profile shows an additional 4% reduction in IOPS and 1% increase in latency. The density profile further reduces IOPS 8% and increases latency 13%. At this point, RAID-5/6 is not reducing performance much, nor is deduplication and compression.



	BASELINE	PERFORMANCE	BALANCED	DENSITY
Average of IOPS	314,443	291,137	215,176	202,947
Average of latency (ms)	1.67	1.76	2.30	2.51

Figure 9: 70% Read Cache Test

At 70% reads, we continue to see IOPS increase and latency decrease. The performance profile IOPS are 8% lower than the baseline, with a 5% latency increase. Here, the reduction in performance from enabling RAID-5/6 becomes apparent. The balanced profile reduces IOPS by 26% and increases latency by 31%. This is because read latency is inherently much lower than write latency, and thus CPU overhead becomes a larger contributor to the overall write latency than the disk latency. This is further witnesses by enabling deduplication and compression, which further reduces IOPS by 6% and increases latency by 9%.

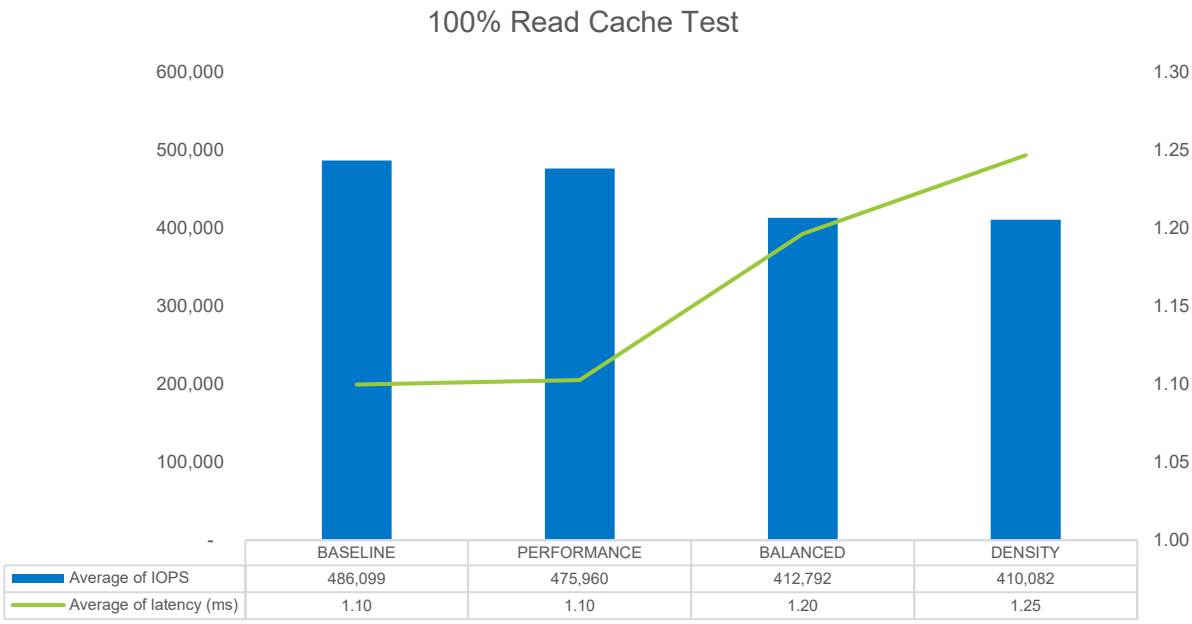


Figure 10: 100% Read Cache Test

At 100% reads, we see the highest performance for each profile. The baseline shows up to 486K IOPS at 1.10ms latency (about 2 GB/s throughput). Enabling checksum for the performance profile reduces IOPS by 2% and latency remains the same (suggesting that checksumming has a small effect on reads). Changing to RAID-5/6 from mirroring (balanced) further reduces IOPS by 13% and increases latency by 9%. Enabling deduplication and compression (density) has minimal impact, reducing IOPS by less than 1% and increasing latency by 4%.

This first test shows that if you have a relatively small working set size—meaning it fits entirely (or mostly) in the cache tier—there is very little downside to utilizing RAID-5/6 with deduplication and compression (especially if your workload is mostly writes). If increased usable capacity is your primary goal, the density profile may yield much more usable capacity than your raw capacity (due mostly to deduplication and compression). It is important to note that not all workloads are the same. Some are more compressible, and some are less compressible. If your workload is highly compressible, using deduplication and compression is strongly recommended. If your dataset is not very compressible, you may see better results without deduplication and compression, as you will take a small performance penalty with little or no capacity benefit.

Tip: Workload Compressibility

With relatively small working sets, there is very little downside to utilizing RAID-5/6 with deduplication and compression (especially if your workload is mostly writes). If increased usable capacity is your primary goal, the density profile may yield much more usable capacity than your raw capacity (due mostly to deduplication and compression).

It is important to note that workload compressibility may vary.

Performance Results: Capacity Test

The second comparison looks at performance differences when the working set does not 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 per disk group for cache. Therefore, only about 29% of user data can reside in the cache tier, while the rest must be held in the capacity tier. Consequently, this means there will be considerable destaging operations for write-intensive workloads, which will reduce performance.

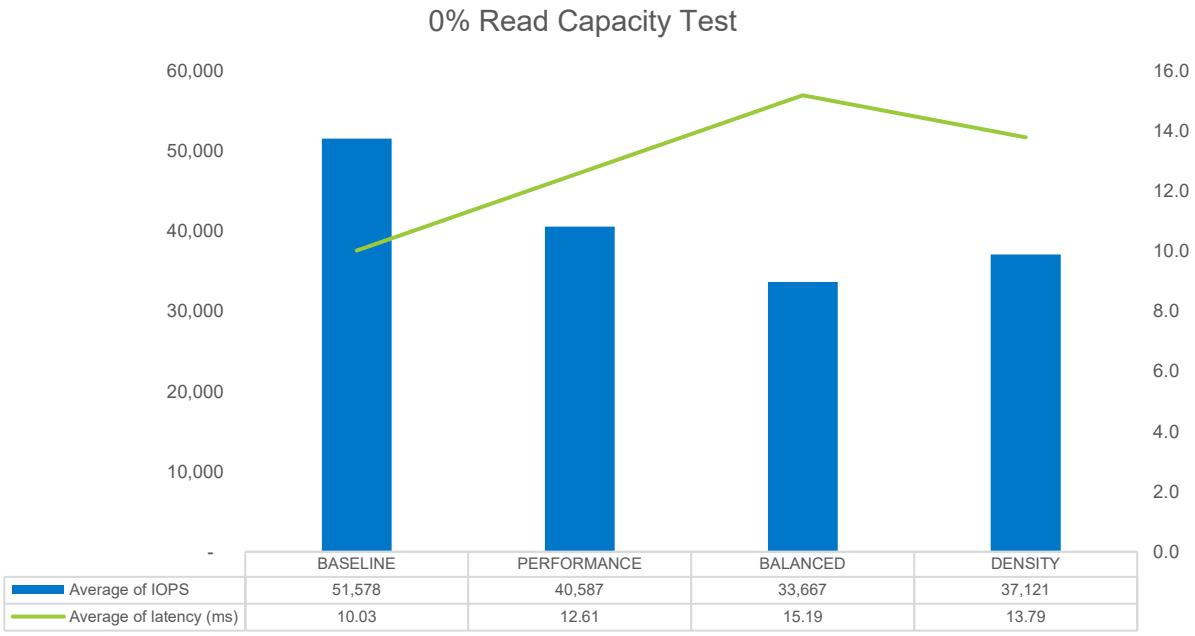


Figure 11: 0% Read Capacity Test

The capacity test shows performance is lower across the board. This is to be expected since significant de-staging occurs (and less of the working set is cached in memory). Enabling checksum drastically reduces performance, showing a 21% reduction in IOPS and 26% increase in latency. The balanced profile performance is also lower with 17% lower IOPS and 20% higher latency. As with the capacity test, we see that enabling deduplication and compression results in higher performance (since less data is being de-staged), resulting in a 10% increase in IOPS and 9% reduction in latency.

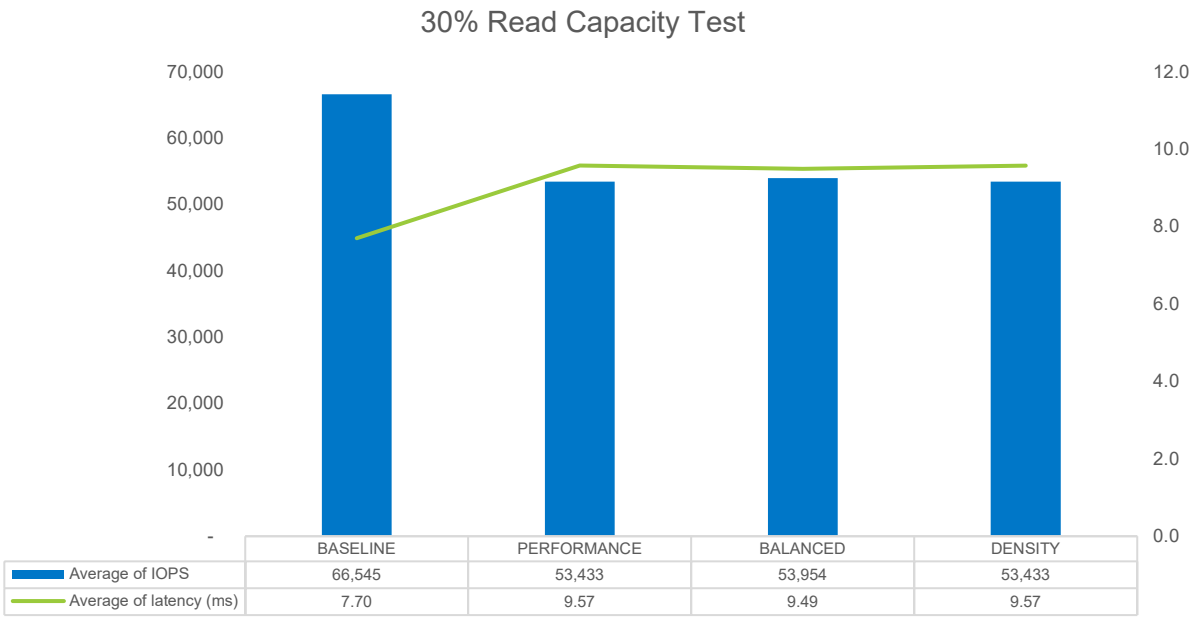


Figure 12: 30% Read Capacity Test

At 30% reads, we see a similar trend as seen at 0%, but with higher performance. Enabling checksum (performance) reduces IOPS by 20% and increases latency by 24%. Switching to RAID-5/6 (balanced) shows almost identical performance, with less than 1% difference in IOPS and latency. Enabling deduplication and compression also results in less than a 1% difference in performance from the performance and balanced profiles.

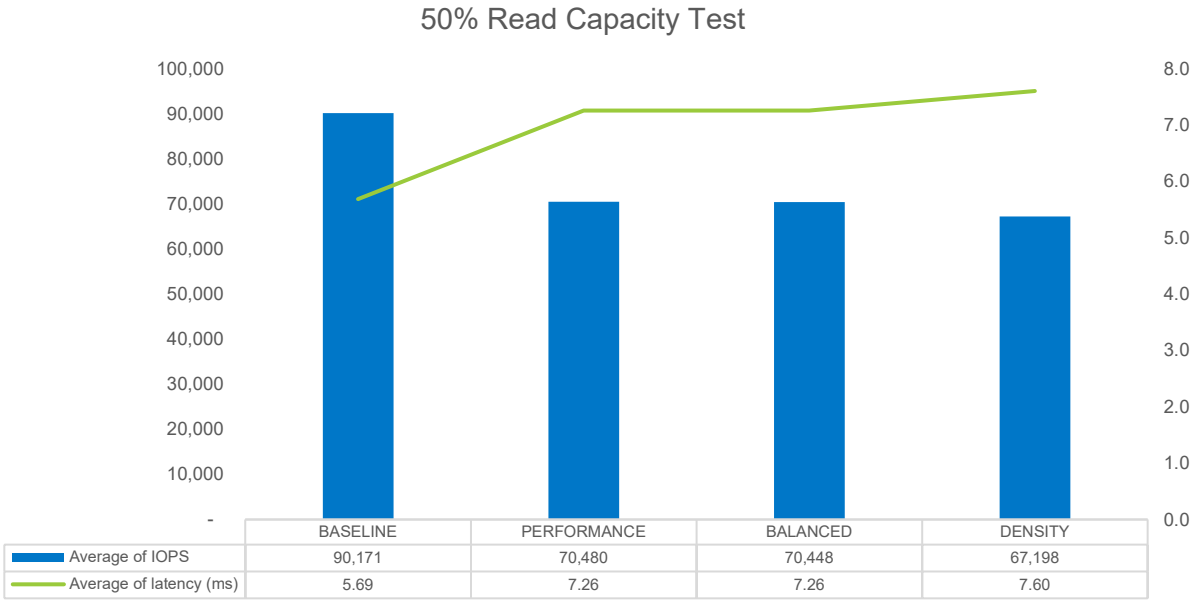


Figure 13: 50% Read Capacity Test

50% reads follows a similar trend, but with higher performance. The performance profile results in a 22% reduction in IOPS and 28% increase in latency. The balanced profile shows less than 1% difference from the performance profile. Compared to the balanced profile, density decreases performance slightly, with 5% fewer IOPS and 5% higher latency.

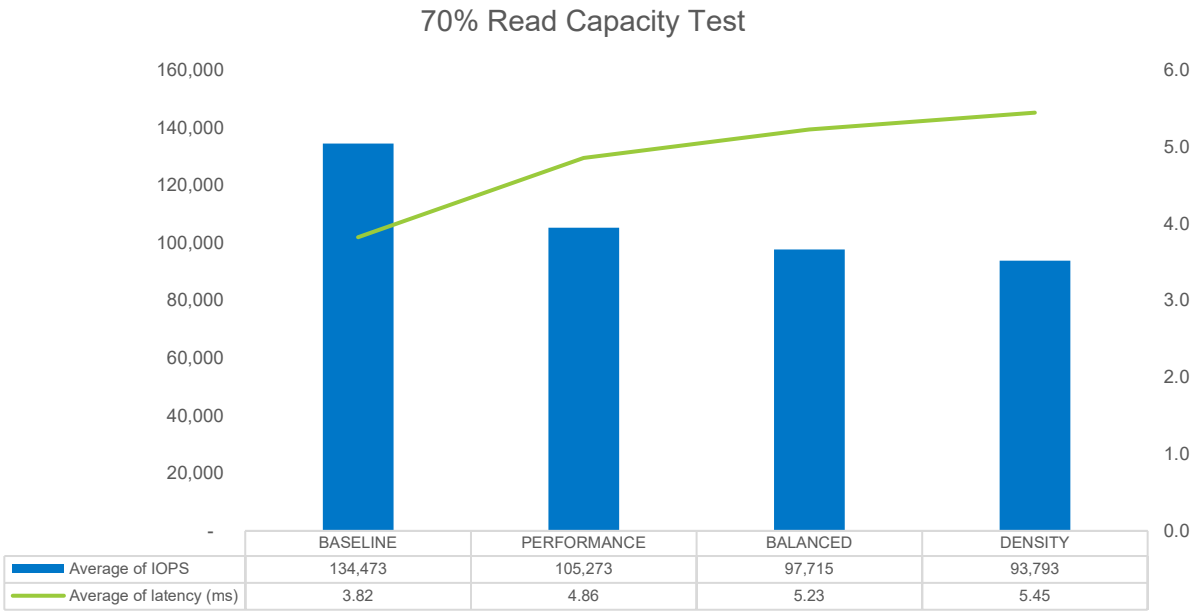


Figure 14: 70% Read Capacity Test

At 70% reads, the trend observed with each successive profile continues, showing lower IOPS and higher latency. Enabling checksum reduces IOPS by 12% and increases latency by 27%. Switching to RAID-5/6 further reduces IOPS by 7% and increases latency by 8%. Enabling deduplication and compression decreases IOPS by 4% and increases latency by 4%.

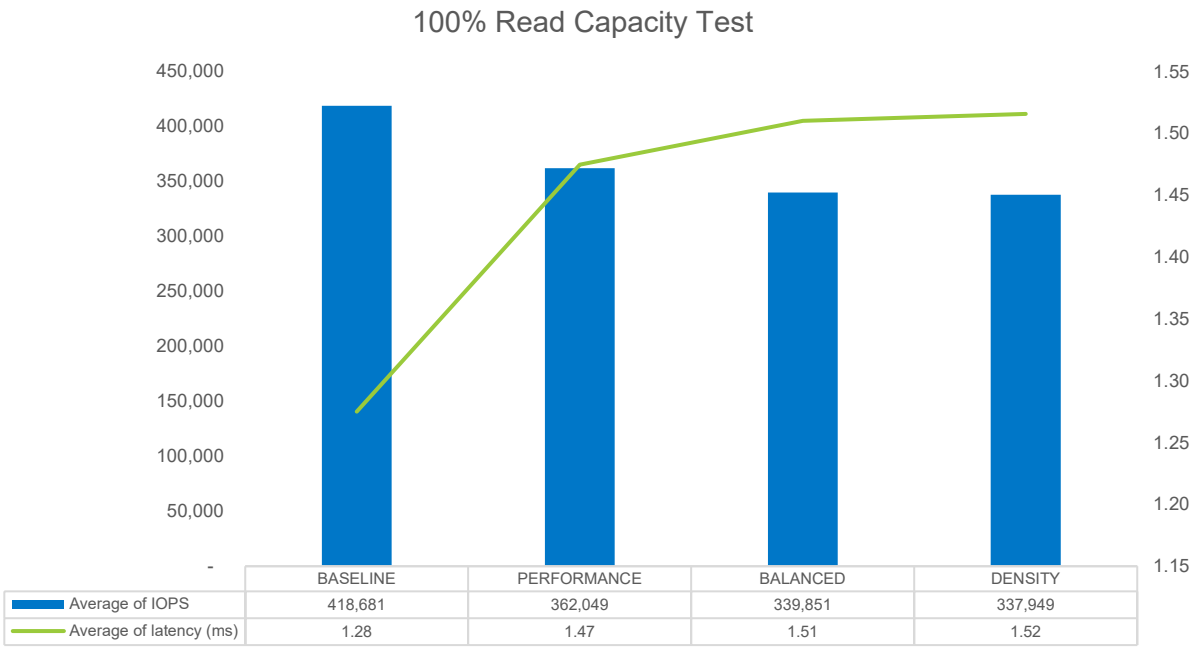


Figure 15: 100% Read Capacity Test

100% read shows a trend like the70% read profile, but with higher performance. Enabling checksum has the largest performance penalty, reducing IOPS by 13% and increasing latency by 15%. Switching to RAID-5/6 further reduces IOPS by 6% and increases latency by 3%. Enabling deduplication and compression has a negligible effect, showing less than 1% difference in IOPS and latency.

Summary

These tests showed the performance of a vSAN cluster strongly depends on the working set size. If the working set fits mostly into the cache tier, performance is higher than when a small portion of it fits. This is especially relevant if you plan to use deduplication and compression. Typically, choosing RAID-5/6 reduces performance significantly, making RAID-1 a common performance choice. However, with the tested SATA SSDs, a large percentage of writes on a working set size that does not fit in the cache tier results in a performance improvement (along with the obvious space savings benefit). If the working set size is large and consists of a large percentage of writes, and if performance is a main goal, enabling deduplication and compression may be beneficial.

Appendix A: vSAN Configuration Details

Tuning Parameters

vSAN's default tunings are set up 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 these three advanced configuration parameters. The table below shows the default value, the value this configuration used, as well as the documents referenced for the tunings.

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

Table 8: vSAN Tuning

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 Parameter File

Below is a sample Vdbench parameter file for a 0% read test against eight VMDKs with a run time of one hour and a warmup (ramp) time of two hours. This particular parameter file is the one used for testing deduplication and compression, using a deduplication ratio of 10 with 4K units, and a compression ratio of 10. This resulted in an initial compression ratio of 8.21X for the capacity test—which after accounting for using RAID-5/6, puts the total ratio of usable capacity to raw capacity at 6.17—and 2.56X for the cache test at a total usable-to-raw ratio of 1.92. The highlighted section denotes the modifications that were made to the Vdbench parameter file generated by HCIBench.

```
*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
dedupratio=10
dedupunit=4k
compratio=10
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=3600,warmup=7200,interval=30
```

Switch Configuration (Sample Subset)

Below is a collection of sample sections of one of the switch configuration files. The "…" denotes an irrelevant missing piece between sections of the configuration file.

```
…
##
## Interface Split configuration
##
   interface ethernet 1/49 module-type qsfp-split-4 force
   interface ethernet 1/51 module-type qsfp-split-4 force

##
## Interface Ethernet configuration
##
…
   interface ethernet 1/51/1 switchport mode trunk
   interface ethernet 1/51/2 switchport mode trunk
   interface ethernet 1/51/3 switchport mode trunk
   interface ethernet 1/51/4 switchport mode trunk
…
##
## VLAN configuration
##
   vlan 100-102
   vlan 110-115
   interface ethernet 1/49/1 switchport trunk allowed-vlan add 1
   interface ethernet 1/49/1 switchport trunk allowed-vlan add 100-102
   interface ethernet 1/49/1 switchport trunk allowed-vlan add 110-115
   interface ethernet 1/49/2 switchport trunk allowed-vlan add 1
   interface ethernet 1/49/2 switchport trunk allowed-vlan add 100-102
   interface ethernet 1/49/2 switchport trunk allowed-vlan add 1
   interface ethernet 1/49/2 switchport trunk allowed-vlan add 100-102
   interface ethernet 1/49/2 switchport trunk allowed-vlan add 110-115
   interface ethernet 1/49/3 switchport trunk allowed-vlan add 1
   interface ethernet 1/49/3 switchport trunk allowed-vlan add 100-102
   interface ethernet 1/49/3 switchport trunk allowed-vlan add 110-115
   interface ethernet 1/49/4 switchport trunk allowed-vlan add 1
   interface ethernet 1/49/4 switchport trunk allowed-vlan add 100-102
   interface ethernet 1/49/4 switchport trunk allowed-vlan add 110-115
```

Appendix B: Monitoring Performance and Measurement Tools

- **HCIBench**: Developed by VMware, HCIBench 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**: Built in to the vCenter Server Appliance (VCSA), vSAN Observer is 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**: A synthetic benchmarking tool developed by Oracle, Vdbench 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)**: Built in to the vSphere Center Appliance as an administration tool, RVC completes many of the tasks that can be done through the web GUI, such as start a vSAN Observer run.
- **vSphere Performance Monitoring**: vSphere has many performance metrics built into the VCSA, including front-end and back-end IOPS and latency.

Appendix C: Deduplication and Compression

With the deduplication and compression settings in the Vdbench parameter file, the theoretical upper limit of the deduplication and compression ratio is 100 (combining a 10X deduplication ratio with a 10X compression ratio). However, we do not see a ratio that is remotely close to the upper limit for reasons discussed below.

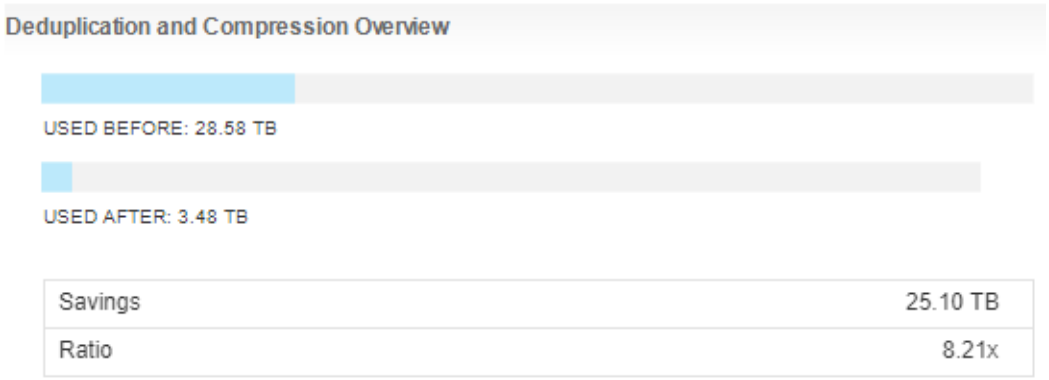


Figure 16: Capacity Test Deduplication and Compression Ratio

As shown in figure 17, these parameters wind up with a final deduplication and compression ratio of 8.21X. We wrote 28.58TB of data before the deduplication and compression, and afterwards, it was reduced to 3.48TB of unique data.

For the cache test, we used the exact same test parameters, just with smaller VMDKs.

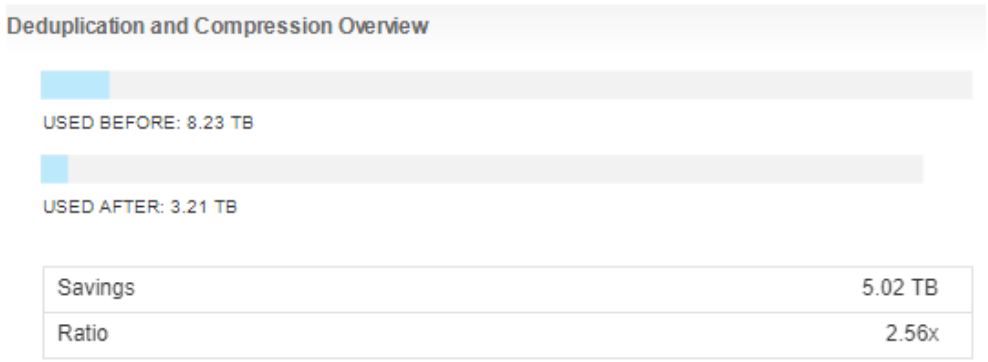


Figure 17: Test Deduplication and Compression Ratio

In figure 17, we see a deduplication and compression ratio of 2.56X, even though the deduplication and compression ratios specified in the Vdbench parameter file were exactly the same as for the capacity test. We notice that the "used before" value is much smaller than for the capacity test, but the "used after" value is very close to the capacity test. To see what is going on, we need to look at the capacity breakdown of the vSAN objects.

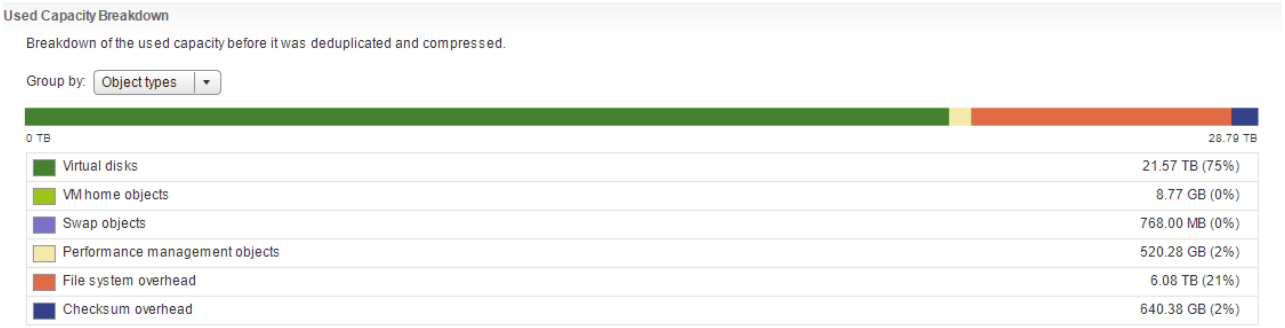


Figure 18: Capacity Test Object Type Breakdown

In figure 19, we can see that for the capacity test, the virtual disks make up the majority of the objects stored on disk, at 75% of the total objects. The remaining objects are overhead for the vSAN file system, checksum values, and other internal objects.

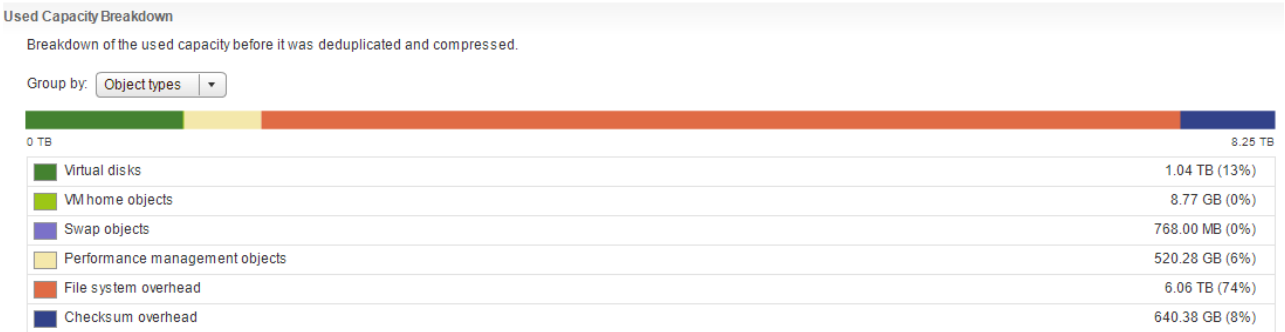


Figure 19: Cache Test Object Type Breakdown

Contrarily, for the cache test, the virtual disks only account for around 13% of the total storage consumed, and thus the ability for the virtual disks to be deduplicated and compressed has a small effect on the total compressed ratio. If the working set size were larger than the 16TB (logical) we studied in this RA, your deduplication and compression ratio should be much higher, as the virtual disks become a larger portion of the data written.

Appendix D: Bill of Materials

Component	Qty per Node	Part Number	Description
Server	1	SYS-2029U-TR25M	Supermicro 2U Ultra Server
CPU	2	BX806736142	6142 Gold 16 core 2.60GHz
Memory	12	MEM-DR432L-CL02-ER26	Micron 32GB DDR4-2666MHz RDIMM ECC
Boot Drive	1	MTFDDAK480TCB-1AR1ZABYY	Micron 5100 PRO SATA 480GB SSD
Cache SSD	2	MK000960GWEZK	Micron 5100 MAX SATA 960GB SSD
Capacity SSD	8	MTFDDAK3T8TDC-1AT1ZABYY	Micron 5200 ECO SATA 3840GB SSD
Networking (NIC)	1	AOC-S25G-m2S	Mellanox ConnectX-4 Lx EN

Table 9: Bill of Materials

Appendix E: About

Micron

Micron Technology is a world leader in innovative memory solutions that transform how the world uses Information. Through our global brands — Micron, Crucial and Ballistix — we offer the industry's broadest portfolio, and are the only company that manufactures today's major memory and storage technologies: NVMe™ and SATA SSDs, DRAM, NAND, NOR, and 3D XPoint™ memory.

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.

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.

©2018 Micron Technology, Inc. All rights reserved. All information herein is provided on an "AS IS" basis without warranties of any kind. Micron, the Micron logo, and all other Micron trademarks are the property of Micron Technology, Inc. All other trademarks are the property of their respective owners. Products are warranted only to meet Micron's production data sheet specifications. Products, programs and specifications are subject to change without notice. Dates are estimates only. Rev. A 7/18 CCM004-676576390-11117