DOCUMENT – AUGUST 2020

PRINTED 20 OCTOBER 2021

VSAN PERFORMANCE EVALUATION CHECKLIST

vmware®

Table of Contents

vSAN Performance Evaluation Checklist

Checklist

The following is a performance checklist to guide you through some best practices related to getting the best possible results from a performance proof-of-concept on vSAN

Before you Start

The following is a performance checklist to guide you through some best practices related to getting the best possible results from a performance proof-of-concept on vSAN. You should, first of all, determine the desired outcome.

Does the customer wish to see the maximum IOPS, the minimum latency, the maximum throughput or even if vSAN can achieve a higher VM consolidation ratio?

You need to document the success criteria for the benchmark test. Get agreement on this matter before proceeding.

"Before you Start" Tasks	Due Date	Done	Initials
Read the **VMware vSAN Design and Sizing Guide** for information on supported hardware configurations, and consideration when deploying vSAN.			
Read the **VMware vSAN Network Design Guide** for information on supported network topologies, configurations and considerations when deploying vSAN networking.			
Read the **vSphere 6.5 Performance Best Practices Guide** for information on ESXi and VM performance considerations.			
Read the Performance Testing section of the **VMware vSAN Proof of Concept Guide** for performance considerations. This contains useful information about many aspects of performance bench marking which should be well understood before continuing.			
VMware's vSAN benchmark tool of choice is HCIbench. Familiarize yourself with HCIbench by visiting the HCIBench fling site , and downloading the User Guide (found under the instructions tab).			
Ensure that the vSphere software versions are supported for vSAN. Ensure that the vCenter server version and ESXi version match for a specific version of vSAN. Latest version is always preferable as it will have the latest fixes and enhancements.			
Verify that you have a uniform cluster - host model, CPU model, number of CPUs, memory size, controller type, cache device(s), capacity device(s)			
Verify network requirements. 1Gbsec for small hybrid vSAN deployments; 10Gbsec minimum for larger hybrid vSAN deployments and all-flash vSAN deployments.			
Verify that the storage controller model is supported and appears on the vSAN VCG. Driver and firmware versions can be confirmed later via health when vSAN has been deployed.			
Verify that the devices used for cache and capacity are on the vSAN VCG. Driver and firmware versions can be confirmed later via health when vSAN has been deployed.			
Verify that the cache and capacity devices have been configured for pass-through mode in the controller BIOS. This is the preferred mode. If this is not possible, verify via the VCG that the device is supported in RAID-0 mode, then configure the device in RAID-0 mode, one device per RAID-0 volume.			
If using RAID-0 mode and the storage controller supports caching, disable the cache. If disabling the cache is not possible, set the storage controller cache to 100% read.			
Disabled vendor specific controller features. Some of these features, e.g. HP SSD Smart Path, have had a negative impact on vSAN – see KB 2092190			
Make a note of the device types and model number. Are they SAS or SATA, are the device Magnetic Disk, SSD or NVMe? All of this may be useful later for evaluating if the best performance has been achieved. This information can usually be found during boot of the ESXi host.			
Consider the number of disk groups per host. Most vSAN Ready Nodes recommend 2. More disk groups can lead to more performance. Click here for an example of how performance can be boosted with an additional disk group.			
Ensure that the cache to capacity ratio adheres to the latest guidelines. The latest caching guidelines can be found on the virtual blocks blog here.			
When using 10Gbsec or perhaps 40/100Gbsec networking, ensure that the cards are placed in the appropriate PCI slot on the host. Different PCI slots can have different specifications. 10Gbsec cards should be placed in 2X factor slots, 40/100Gbsec cards should be placed in 8X factor slots.			
While vSAN works perfectly fine with an MTU of 1500 on the vSAN network, and MTU of 9000 (known as jumbo frames) can increase performance for certain workloads, and can be less intensive on the CPU, possibly leading to higher throughput.			
If planning to test vSAN Encryption, ensure that the hosts support AES-NI (Intel's Advanced Encryption Standard New Instruction Set), and that it is enabled in the BIOS of the host.			
Ensure network partitioning (commonly known as NPAR) features offered by some host NICs are disabled. NPAR will restrict the maximum bandwidth available to NICs used for uplinks, and can lead to reduced levels of performance. For more information, see page 32 of the Troubleshooting vSAN Performance document on StorageHub			
Understand what the customer goals of the performance test are. Is it maximum IOPS, maximum throughput, minimum latency or a combination of each. Please read this performance guidelines blog which contains some very relevant information about the trade offs that are needed for performance testing.			
Have you informed the vSAN POC (proof-of-concept) team about your benchmarking test? PLEASE DO THIS! SE/SDS teams should contact the vSAN POC team before starting any actual setup of hardware / PoC equipment or any benchmark testing. Look for the list of POC Architects under the Product Enablement section. The vSAN POC team spends a lot of time on vSAN performance testing and tuning, and you can leverage their knowledge for your testing. And this team would rather be engaged sooner rather than later.			

This completes the 'Before you start tasks' section.

Host Based Tasks after vSAN is Deployed

Some additional tuning might be required on the hosts. Here are some guidelines:

Host Based Tasks after vSAN is Deployed	Due Date	Done	Initials
Download the latest version of the HCL DB file in the vSAN Health Checks.			
Verify that ALL vSAN health checks are green . Any health check warnings must be addressed before proceeding with a performance evaluation. KB 2114803 describes the various vSAN health checks, and can give you guidance to specific articles on failed checks. The vSphere UI should also have links to Ask-VMware KB articles directly from the check. Particular attention should be paid to storage controller model, driver and firmware.			
Set the Host Power Management to 'OS Controlled' in the Server BIOS for the duration of the performance test. Check out the steps in the Performance Best Practices Guide for vSphere 6.5 . Verify that the setting has taken effect by checking the Power Management of the host in the vSphere client. Technology should show APCI P-States and C-states, and the active policy should show 'High performance'.			
Interrupt Remapping allows all CPUs to handle interrupt requests and should be enabled for performance testing. If it is disabled (which it is by default on ESXi 6), it forces the first CPU to handle all interrupt requests. See KB 1030265 for reasons on why it is disabled, and steps to enable it.			
Make a note of the device queue depths. Note that LSOM (low-level disk layer of vSAN, short for Log Structured Object Manager) calculates queue depth at 90% of the device queue depth. Thus, if device queue depth is 32, LSOM will calculate this as 28. This is done at boot time. Use *zcat /var/log/boot.gz | grep "Queue Depth"* on an ESXi shell to verify.			
If vSAN does not have its own dedicated physical network, then consider utilizing NIOC to ensure fairness between network users. NIOC is covered in detail in the VMware vSAN Network Design Guide .			
Verify that the vSAN network is optimal. Use pktcap-uw on the ESXi hosts to capture inbound and outbound traffic. Check for the presence of Keep-Alive/TCP Dup Ack packets which could be indicative of issues. KB 2051814 has further details on how to run pktcap-uw. Wireshark is a useful tool for scanning the resulting packet trace.			

Choosing An Appropriate Policy to Test

You may want to test performance with different policies. Here are some guidelines on policies:

Choosing an Appropriate Policy to Test	Due Date	Done	Initials
Choosing a policy is usually related to choosing between availability and performance. RAID-1, which is used for optimum performance, can tolerate up to 3 failures, but creates 4 copies of the data. RAID-5 and RAID-6 can tolerate 1 or 2 failures respectively, consumes less space than RAID-1, but does not perform as well. Also, be aware that the hybrid version of vSAN does not support RAID-5 or RAID-6. Consider these points when choosing an appropriate policy to test.			
Create a policy of "Number of Failures to Tolerate (FTT) = 0". This instantiated RAID-0 objects which should exist on a single host. One can then use vMotion to place the VM's compute and it's attached VMDK on the same host, which means that the network can be excluded from any tests. However there is no way to automatically place a RAID-0 VMDK on the same host as its compute in the current release of vSAN.			
Create a policy of FTT=1, and a Stripe Width. This will stripe the VMDK across multiple capacity devices, as well as mirroring it with a RAID-1, reducing hot-spotting. Additionally, performance may be boosted if the striped components are placed on different hosts and/or different disk groups. However, there is no guarantee of an additional performance increase if the striped components are placed on the same host or even on the same disk group. Increase the stripe width from 2 to 3 or more to improve performance, if the available resources allow.			
Create a policy of FTT=1. This creates RAID-1 objects, and will place components on two different hosts, and will evenly distribute the read requests across both components. This object will always incur network overhead on write, as writes need to go to both sides of the RAID-1.			
Create a RAID-5 policy if the customer plans to use this policy in production. This policy is only available on vSAN All-Flash configurations. It is not available on hybrid vSAN. Note that performance will not be as good as RAID-1, due to overheads such as parity calculations and Read-Modify-Write operations for partial writes.			
Create a RAID-6 policy if the customer plans to use this policy in production. This policy is only available on vSAN All-Flash configurations. It is not available on hybrid vSAN. Note that performance will not be as good as RAID-1, due to overheads such as parity calculations and Read-Modify-Write operations for partial writes.			

This completes the policy selection section.

Choosing Data Services

You may want to test performance with different data services.

Here are some options on data services:

Choosing Data Services	Due Date	Done	Initials
Be aware that All-Flash vSAN supports more data services than the hybrid version of vSAN.			
Checksum On/Off – Policy Driven Recommendation: **Leave Checksum enabled.** Checksum has a performance impact for write workloads. This is due to the overhead of checksum calculations and extra checksum IO to disk.			
Deduplication/Compression On/Off – Cluster wide change Enabling Deduplication and Compression will cause additional IOPS and latency overhead, especially on write heavy workload. This is mainly due to metadata IO overhead.			
Encryption On/Off – Cluster wide change Recommendation: **Only enable encryption if there is hardware assisted encryption support such as Intel's AES-NI.** Enabling Encryption increases the CPU cost per IO increases because of data encryption overhead.			

This completes the data services section

Prepping for the HCIBench Benchmark

Prepping for the HCIBench Benchmark	Due Date	Done	Initials
Download the HCIBench OVA As of March 2018 the latest version is v1.6.6 which this guide is based on			
Download the HCIBench User Guide			
Decide which Data Services to Enable (e.g. Deduplication) first			
The Test VMs require a working network, typically DHCP should be used.This is the simplest way to deploy the benchmark Verify DHCP is available on VM network. Refer to alternative methods documented in the User Guide if DHCP is not available see doc https://download3.vmware.com/software/vmw-tools/hcibench/HCIBench_User_Guide_1.6.5.pdf			
The benchmark uses the vSAN Default Storage Policy. the vSAN default policy used FTT=1. If you want the benchmark to use an alternate policy, change the vSAN Default Storage Policy to meet your requirements before you start your tests			
Ensure that DRS is enabled, but only in 'Partially Automated Mode'. This ensures that the VMs are deployed evenly, but also avoids vMotion operations occurring during testing.			
Run HCIBench with following vdbench parameters			
• Decide on number of VMs per host. Initial recommendation is to deploy 2 VMs per diskgroup.			
• Decide on number of VMDKs per VM (e.g. 8 which is default)			
• Decide on size of VMDK (e.g. 10GB, which is default)			
• Decide on Outstanding IO (OIO) per VM (e.g. 2 to 4). VMware recommends 4 OIO per VMDK. If resulting latency is too high, OIO can be lowered.			
• Decide on Block Size (e.g. 4K) Smaller blocks sizes give better IOPS results on vSAN, but larger block sizes can give better throughput.			
• Decide on Read/Write Ratio (e.g. 70/30)			
• Decide on Random or Sequential IO. Random IOs give better performance on vSAN.			
Make a note of your Oracle credentials as you will need these to download vdbench. from If you do not have these to hand, do not worry. Steps are provided to help you create a new account.			

This completes the benchmark prep section.

Initial Functional Test -HCIBench Easy Run

Initial Functional Test - HCIBench EASY RUN	Due Date	Done	Initials
Check the EASY RUN checkbox. This automatically defines the number of VMs, VMDKs and Outstanding IO. It creates 2 VMs per disk group, 8 VMDKs per VM and sets the VMDK size based on size of cache tier. It also sets the appropriate preparation mode to be either Zero or Random by looking at the vSAN configuration. **Note: The prepare step will take a considerable amount of time as each VM disk will have data written to it in a sequential fashion to ensure we do not hit a first write penalty.** The workload is set to 70% Read, 100% Random and Outstanding IO is set to 4 threads per VMDK. There is a 30 minute warmup period, following by 60 minutes tests (results are only based on the 60 minutes testing).			
Click on the "Download the Vdbench" button. This will open a new browser tab which will direct you to Vdbench downloads. Click to accept the license agreement, and then click to download the latest zip. You will now need to login to an Oracle account to complete the operation. If you do not have an Oracle account, you will need to create one. Save the Vdbench zip file.			
Click on the " **Save Configuration** " button. If you have missed any fields, these will be reported here. If you see a Progress Finished message pop-up, the configuration has been populated correctly. Close the pop-up by clicking X.			
Click the 'Test' button. This will start the deployment of test VMs and run the Vdbench tests. The tasks can be monitored from the vSphere client. The test VMs are named vdbench-<datastore>-X-Y. After the VMs are successfully deployed, I/O Tests are started.			
Deploy the HCIBench OVA in your environment, and login to portal (http://vdbench-ip:8080) following instructions outlined in the User Guide. Any issues with this process should be directed to vsanperformance@vmware.com			
For vSAN Observer UI display of performance, navigate to the IO Profile folder, then the iotest-vdbench folder, and select the stats.html file			
In the HCIBench Configuration Page, add your vSphere environmental details. This includes vCenter, Datacenter, Cluster, Network, and of course Datastore.			
Leave the 'Deploy on Hosts' button unchecked. This will deploy test VMs to ALL hosts in the cluster rather than specific hosts.			
Next, click on the " **Browse...** " button and select the Vdbench zip file that you have just downloaded from Oracle. Once it has been selected, click on the Upload Vdbench button. When the "Upload finished" message pops up, click OK.			
Next, click on the "Validate Configuration" button. This can take a few moments to complete. Once complete, a report is generated. You should see the final message state "All the config has been validated, please go ahead and kick off testing". If not, you need to address any outstanding issues before proceeding. Any issues with this process should be directed to vsanperformance@vmware.com . Click X to close the information report.			
Populate the ESXi user and password field. This is a requirement to drop vSAN cache before each test. The ESXi username and password must be uniform across all hosts.			
The next option is the 'Clear Read/Write Cache Before Each Testing' which, when checked, will drop the vSAN cache contents. There are two considerations with this setting: 1. Clear the cache: the reason for clearing the caches is to start from a blank slate with each test – removing the effect of a prior test (except for the actual data stored on the capacity disks) and therefore increasing the repeat-ability of test results. This is a best practice for comparing performance between different test configurations, e.g. RAID-1 vs RAID-5, dedupe enabled vs dedup disabled, etc. Though we recommend clearing caches for repeat-ability and isolation between tests, this will increase the amount of soak-time needed to achieve run a benchmark test due to the fact that cache needs to be repopulate to achieve optimal performance conditions. 2. Don't clear the cache: Clearing the cache on hybrid vSAN causes a drop in read performance until the read cache is re-warmed. Running a test for a long duration will effectively push out the old contents of both the read and write cache, accomplishing the same goal flushing the caches and then re-filled them with the new contents. Keeping the caches intact will show more realistic and consistent performance throughout testing. Thus if the goal is to gather "steady state results" or benchmark against a competing product, the recommendation is not to clear the cache to achieve optimal performance conditions more quickly. If you wish to clear the cache, this feature requires that SSH is enabled on all hosts, and that the ESXi username and password fields are populated.			
When the test is finished, examine the results by clicking on the 'Result' button. The results are saved in a folder using the name of the test, i.e. MyFirstTest. Here you will find results based on the IO Profile used to make the vdbench parameter file. An XLS spreadsheet has all of the captured metrics. The <IO-Profile>.txt file a summary of the results.			

EASY RUN might be just what you need for your performance benchmark. However, you can also reuse the EASY RUN results to

fine-tune your next benchmark run.

Success Criteria

- What do you want to achieve from this benchmark?
- What is the customer's success criteria?

The success criteria are based on a number of things – achieving

1. Max IOPS,
2. Max Throughput,
3. Minimum Latency,
4. a mixture of IOPS, TPUT and Latency or
5. VM Consolidation Ratio.

Depending on your priority on achieving 1, 2, 3, 4 or 5, the configuration may be different.

> For example, VDI desktop VMs may only have a single VMDK per VM, and since these generally do not generate many IOPS, you should be able to deploy many of these VMs and still achieve minimum latency. OLTP may require many VMDKs per VM, so you might only need to deploy a few of these VMs to achieve maximum IOPS. More IOPS and Throughput can be achieved with more VMs and more VMDKs. The trade-off is always IOPS and Throughput versus Latency – the more IO you wish to drive to a datastore, the higher the latency can become. Outstanding IO (called 'Number of Threads per Disk' in HCIBench) is also an incredibly important factor when it comes to performance benchmarks. It can help to generate more IOPS and Throughput by making sure that the IO queue is always filled, but the downside is that the more IO that is allowed to queue up, the higher the latency will be.

Note down the success criteria once agreed with the customer.

All of this is a balance, as you try to figure out how much you can push the system. Please read this performance guidelines blog which contains some very relevant information.

HCI Bench - Further Tuning

For a more advanced performance benchmark, the following steps can be considered.

HCIBench - Further Tuning - VDBench Guest VM Specification	Due Date	Due	Initials
Number of VMs: Variable • Increment as you test, until you find your sweet spot between IOPS, throughput and latency. Alternatively, use 'Number of Threads Per Disk' to achieve the goal. • Rinse and repeat			
Number of VMDKs: Static • Avoid tuning the number of VMDK as you will need to initialize the disks with each new test. Outstanding IO can be tuned via Number of Threads per Disk.			
Number of Threads Per Disk: **Variable** • Increment as you test, until you find your sweet spot between IOPS, throughput and latency. Alternatively, use 'Number of VMs' to achieve the goal. • Start with a small value of 1 or 2, and gradually increment. This will give you a balance for IOPS, Throughput and Latency that you can fine tune. • Rinse and repeat			
Block Size: Variable • Pick a few common block sizes, for example 4KB, 16K and 64K. Note that very large blocks are chunked by vSAN to 64KB, resulting in IO amplification and triggering congestion. Smaller blocks should be used for benchmarking. • Rinse and repeat			
Re-use The Existing VMs If Possible Recommendation: Check this box • This will avoid having to reinitialize the data in the VMDKs with each run. • This will also ensure consistency across tests as you are using the same data with each iteration.			
Clean up VMs after testing Recommendation: **Uncheck this box if not changing policy and/or enabling data services between tests** • This will avoid having to reinitialize the data in the VMDKs with each run. • This will also ensure consistency across tests as you are using the same data with each iteration.			
Prepare Virtual Disk Before Testing Recommendation: **Zero for non-dedupe, Random for dedupe** • This may take a long time to complete. DO NOT SKIP THIS TEST, or your performance results will be sub optimal • If you decide to enable/disable dedupe during your testing, you will need to choose a different prepare option • If you are re-using the VMs, you do not need to repeat this step			
Testing Duration: **2 hours for a typical run** Length depends on cache size, incoming rate of writes, and write cache drain rate to capacity tier (which can vary with disk group configuration and features) • Objective is to capture performance while cache is being utilized and destaging from cache tier to capacity tier is occurring. • While the test workload is running, look at the graph for "Write Buffer Free Percentage". If the percentage of write buffer free space is decreasing, then vSAN is taking in writes faster than it is processing them to their final home in the Capacity tier. You should continue to run the workload until the Write Buffer Free Percent stays the same or increases for a 30 minute period. • Write Buffer Free and Cache Disk Destage rate can be viewed in the vSAN Performance Graphs via Hosts > Monitor > Performance > vSAN Disk group			
Overheads with policies and data services • Deduplication and Compress will add significant overheads with large block, random workloads • Checksum will add significant overheads with large blocks (>64KB) • Erasure Coding Policies (R5/R6) will introduce IO amplification on ALL writes, but partial writes (which involve a read-modify-write operation) will introduce considerably more IO amplification. • Recommendation: **Enable new data services or introduce policy changes one at a time.** Don't change lots of things at once. Make sure that the current set of test VMs are removed before making any policy and/or data service changes, as this will add unnecessary time to the performance test			
Performance Diagnostic Guidance for tuning benchmarks Performance Diagnostics is a built-in utility which can provide guidance towards achieving better benchmark results. This requires CEIP (Customer Experience Improvement Program) to be enabled. This is integrated with HCIBench User Guide and the vSAN Support Insight documentation https://core.vmware.com/resource/vsan-support-insight			

This complete the approach to performance testing of vSAN with HCIBench.

Need Help?

Where to get Help?

In the ' **Before you start tasks** ' section, we mentioned that you should have informed the vSAN POC team before attempting any sort of vSAN benchmark. Normally this engagement is via your vSAN Specialist SE, who can seek SABU help if necessary. This team can give you guidance based on the many benchmarking efforts that they have already carried out. They should always be consulted first for advice if the benchmark is not performing as expected.

For issues with HCIbench, reach out to vsanperformance@vmware.com .

For other issues encountered during the POC, such as device or controller issues, it is recommended that a ticket is raised with GSS. Remember to capture the appropriate logs, etc, before opening a ticket so you can get a speedy resolution.

Authors:

Cormac Hogan - Director and Chief Technologist, Storage & Availability Business Unit

Paudie O'Riordan - Staff Engineer, Storage & Availability Business Unit

Andreas Scherr - Sr. Solutions Architect, Storage & Availability Business Unit

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