StarWind Virtual SAN®

Best Practices

FEBRUARY,2019

BEST PRACTICES

Trademarks

Changes

The material in this document is for information only and is subject to change without notice. While reasonable efforts have been made in the preparation of this document to assure its accuracy, StarWind Software assumes no liability resulting from errors or omissions in this document, or from the use of the information contained herein. StarWind Software reserves the right to make changes in the product design without reservation and without notification to its users.

Technical Support and Services

If you have questions about installing or using this software, check this and other documents first - you will find answers to most of your questions on the Technical Papers webpage or in StarWind Forum. If you need further assistance, please contact us .

About StarWind

StarWind is a pioneer in virtualization and a company that participated in the development of this technology from its earliest days. Now the company is among the leading vendors of software and hardware hyper-converged solutions. The company’s core product is the years-proven StarWind Virtual SAN, which allows SMB and ROBO to benefit from cost-efficient hyperconverged IT infrastructure. Having earned a reputation of reliability, StarWind created a hardware product line and is actively tapping into hyperconverged and storage appliances market. In 2016, Gartner named StarWind “Cool Vendor for Compute Platforms” following the success and popularity of StarWind HyperConverged Appliance. StarWind partners with world-known companies: Microsoft, VMware, Veeam, Intel, Dell, Mellanox, Citrix, Western Digital, etc.

Copyright ©2009-2018 StarWind Software Inc.

INTRODUCTION

The basic principle of building any highly available environment is eliminating any single points of failure in the hardware and software configurations. Since a single hardware failure can lead to downtime of the whole system, it is vital to achieve the redundancy of all the elements in the system in order to eliminate or minimize downtime caused by failures.

StarWind Virtual SAN makes it possible for customers to minimize or avoid the downtime associated with storage or host failures. StarWind also enables virtualized environments maintenance with zero downtime. This is achieved by clustering multiple StarWind servers into a fault tolerant storage cluster to guarantee seamless storage failover in the event of hardware/software failures and power outages.

StarWind HA relies on redundant network links between the StarWind hosts to ensure storage resilience. This allows StarWind to maintain a fully fault tolerant storage cluster with just two hosts. On the contrary, the major part of storage solutions available on the market requires some sort of a 3rd entity in the storage cluster to maintain resilience and arbitrate the storage cluster in the event of hardware failure.

Using the internal and native OS tools, StarWind HA constantly monitors the state of all the network links between the servers in the HA cluster. Should any of the cluster nodes fail or stop processing requests properly, the failover is instantly initiated from the client OS/Hypervisor side. This guarantees the correct failover procedure and makes StarWind HA automatically compatible with various initiators. StarWind also provides an internal heartbeat mechanism, which ensures proper storage path isolation in the event of synchronization network failures and prevents so-called storage “split-brain”. Another mechanism to avoid “split-brain” is node majority failover strategy employed when Heartbeat is not available.

StarWind HA can be easily combined with native Windows file storage features and can also leverage Microsoft SoFS acting as a continuously available file share service for multiple non-clustered client computers.

StarWind HA has multiple advantages over the traditional storage failover solutions:

Hardware agnostic – no proprietary hardware necessary, commodity x86 servers supported.

Reduced TCO – doesn’t require dedicated storage hardware, can be installed directly into the hypervisor.

Minimal setup time – installation and full HA storage configuration take under 30 minutes.

Ease of use and management – Native Windows application with a user-friendly centralized management console.

Instant failover and failback – StarWind HA leverages MPIO driver on the initiator host for a fully transparent failover procedure.

A full set of the up-to-date technical documentation can always be found here, or by pressing the **Help** button in the StarWind Management Console.

For any technical inquiries, please, visit our online community, Frequently Asked Questions page, or use the support form to contact our technical support department.

Design Principles

Host considerations

Proper equipment selection is a very important step in the architecture planning. Always choose the appropriate equipment for the tasks the highly available environment is to cope with. The used equipment should support redundant hot-swappable parts, especially if a minimalistic cluster (less than 4 nodes) is planned. Note that overestimating the storage requirements is not always a good practice. It can turn out that the equipment purchased according to these estimations will never be used effectively resulting in low ROI. Ensure to always plan for the scalability of the purchased servers. Keep in mind that it is possible to not only scale-out to more nodes but also scale-up the existing hosts as compute and storage requirements grow.

Hardware differences

StarWind HA performs in active-active mode, thus it is a best practice to use identical hardware for all the nodes participating in an HA storage cluster. Rarely, one HA node can have faster storage to improve read performance. In this case, ALUA (Asymmetric Logical Unit Access) is configured to achieve the optimal performance. ALUA marks a certain storage path as optimal or non-optimal for write IO. The initiator server (if supported) uses these marks to optimize LU access. Please refer to the Asymmetric configurations chapter of the document for more information.

*Implementation of the ALUA mechanisms on the initiator side does not always result in the optimal performance even though all pre-requisites are fulfilled from the storage side.

OS differences

When configuring StarWind Virtual SAN, it is a best practice to use an identical Windows Server OS version installed on all StarWind hosts. An edition difference is possible though: e.g., host 1 runs Windows 2016 Standard, and host 2 runs Windows 2016 Datacenter.

Please note that certain Windows editions are not supported. Please refer to the System requirements page for the list of supported operating systems: http://www.starwindsoftware.com/system-requirements

StarWind VSAN Versions

It is mandatory to have the same version and build of StarWind Virtual SAN installed on all nodes of the HA storage cluster. Always update all StarWind Virtual SAN servers in the environment to avoid version and build mismatch due to differences in the HA device compatibility, performance, and operational features. It is strongly recommended to keep all StarWind servers in the environment up-to-date and install StarWind updates as soon as they are available. By default, email notifications about the available updates are sent to the address specified while registering on the StarWind website. Please make sure that all @starwind.com and @starwindsoftware.com whitelisted in the mailbox in order not to miss any important announcements.

OS configuration

HA environment has special requirements for uptime and availability. In order to fulfill these requirements, the user needs to adjust certain settings in the operating system.

Updates

Since any clustered environment has strict requirements for maintenance downtime, Administrators are required to control all the update processes on the server. All the automatic Windows updates should be either disabled or configured to prevent the cluster interruption. Never apply updates to more than one node of the HA cluster at a time. After applying updates to a node, verify that the functionality is intact, and all iSCSI devices are resynchronized and are reconnected to the initiator hosts. Only after verifying everything described above, it initiates the update processes on the next HA node.

Firewalls

StarWind operates through ports 3260 and 3261. 3260 is used for iSCSI traffic, and 3261

serves for StarWind Management Console connections. StarWind installer automatically opens these ports in the Windows Firewall during the initial installation. If a third-party firewall is used, ports 3260 and 3261 have to be opened manually.

Additional software

It is not recommended to install any kind of third-party applications on the server running StarWind Virtual SAN. Exceptions here are benchmarking tools, remote access utilities, and hardware management utilities such as Network card managers or RAID controller management packs. In order to clarify some issues about the software that is to be installed on the server running StarWind Virtual SAN, please consult StarWind Software support.

Backup Recommendations

Configuring backups is vital for any environment including highly-available ones.

When using StarWind with synchronous replication feature inside of a virtual machine, it is recommended not to make backups and snapshots of the virtual machine with StarWind service which could pause the StarWind virtual machine. Pausing the virtual machines while StarWind service is under load may lead to split-brain issues in devices with synchronous replication and data corruption.

It is recommended to backup data and virtual machines which are located on StarWind HA storage instead of StarWind virtual machine backup.

Asymmetric configurations

For certain tasks, e.g., fulfilling an IO pattern like 90% Read with a high degree of random IO, StarWind HA cluster can be configured asymmetrically: e.g., all HA nodes have identical network performance, but one node uses flash storage to reach the high rate of random read IO. The asymmetric configuration allows users to increase the read performance of the HA SAN while keeping the TCO lower. In this case, ALUA is configured for the HA device that is required to serve this IO pattern. With ALUA, all the network paths to the storage array remain active, but the initiator only writes data over those paths, which are marked as optimal in the ALUA configuration. Writes are routed to slower storage in order to avoid storage bottlenecks.

Networking

The network is one of the most important parts of the Virtual SAN environment. Determining the right network bandwidth for the SAN is the #1 task along with finding

the approximate IOPS number the storage has to deliver in order to fulfill the requirements of applications it serves.

Networking speed considerations

Once finished with the IOPS calculations, pick the networking equipment that won't cause bottlenecks on the interconnect level. E.g., if the calculations say that the cluster requires 63,000 IOPS (or demands ~250mb/s streaming speed capabilities), 1GbE network will not be enough for the setup. Networking throughput demand grows along with IOPS demand, so after 250,000 IOPS a single 10 GbE card becomes a bottleneck. Below, there is a table showing the recommended network equipment throughput depending on the IOPS.

IOPS (4K)	Network
26,000–48,000	2x 1 Gigabit (MPIO)
48,000–250,000	10 Gigabit
250,000–480,000	2x 10 Gigabit (MPIO)
480,000 and higher	40/56 Gigabit
1–26,000	1 Gigabit

Networking layout recommendations

The main goal of a highly available storage solution is 24/7 uptime with zero downtime in the event of most failures, or during maintenance and upgrades. Thus, it is very important to understand that High Availability is not achieved just by clustering the servers. It always is a combination of redundant hardware, special software, and a set of configurations that make the system truly highly available. Below, find the reference architecture diagrams showing the recommended redundant networking for HA. These network layouts are considered the best practice of StarWind Virtual SAN design.

It is recommended to use a direct connection for Synchronization and iSCSI channels, but network switches can also be used. All switches used in the StarWind Virtual SAN deployment have to be redundant. This applies to all iSCSI traffic switches and, if used, synchronization channel switches.

The connections pictured on the diagrams are dedicated to StarWind traffic. In hyperconverged scenarios (when StarWind VSAN is running on the hypervisor host), it is possible to share the heartbeat with other networks, e.g., vMotion/Live Migration/iSCSI, except for synchronization network.

Diagrams provided below are based on most popular configuration cases and assume

using SSD drives as the storage workload equal to almost 250,000 IOPS.

NOTE: The diagrams only show the SAN connections. LAN connections, internal cluster communication, and any auxiliary connections have to be configured using the separate network equipment or be separated from the Synchronization/iSCSI traffic. Networking inside the cluster should be configured according to the hypervisor vendor recommendations and best practices.

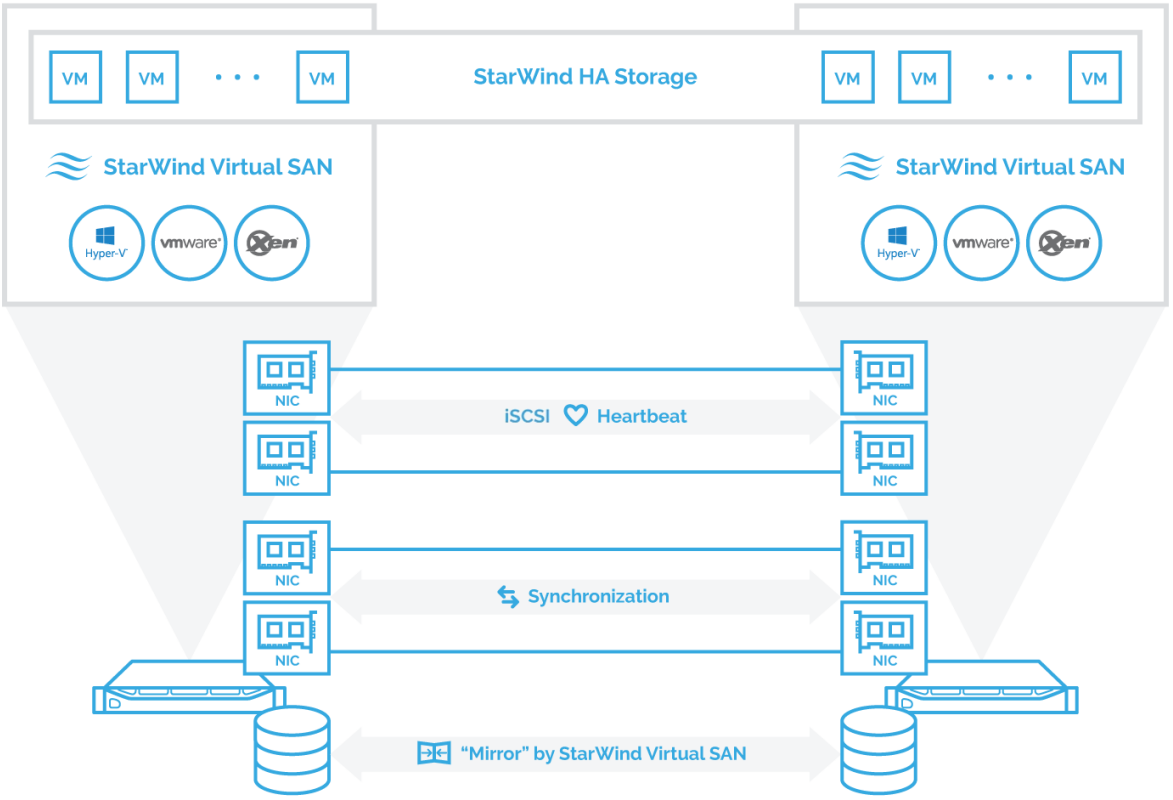

Fig. 1: Hyperconverged setup. 2-node cluster with StarWind Virtual SAN. Direct redundant physical connections are used for Synchronization and iSCSI/Heartbeat channels.

NOTE: Do not use ISCSI/Heartbeat and Synchronization channels for the same physical link.

Synchronization and iSCSI/Heartbeat links and can be connected either to the redundant switches or directly between the nodes.

The setup is stable if non-redundant physical connections for Synchronization and iSCSI/Heartbeat channels are used since one node can take the workload of the whole HA system thus providing the node-level redundancy.

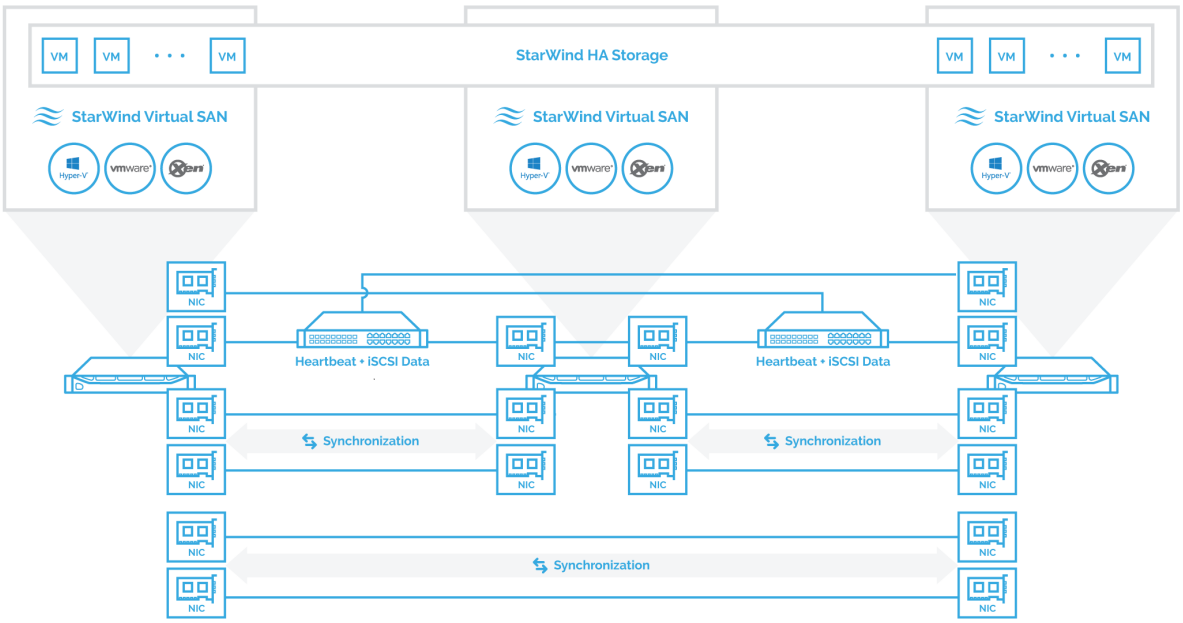

Fig. 2: Hyperconverged setup. 3-node cluster with StarWind Virtual SAN. Direct redundant connections are used for Synchronization channel while iSCSI/Heartbeat channels are connected via redundant switches.

NOTE: Do not use ISCSI/Heartbeat and Synchronization channels for the same physical link. Synchronization and iSCSI/Heartbeat links and can be connected either to the redundant switches or connected directly between the nodes.

The setup is stable if non-redundant physical connections for Synchronization are used since one node can take the workload of the whole HA system thus providing the node-level redundancy.

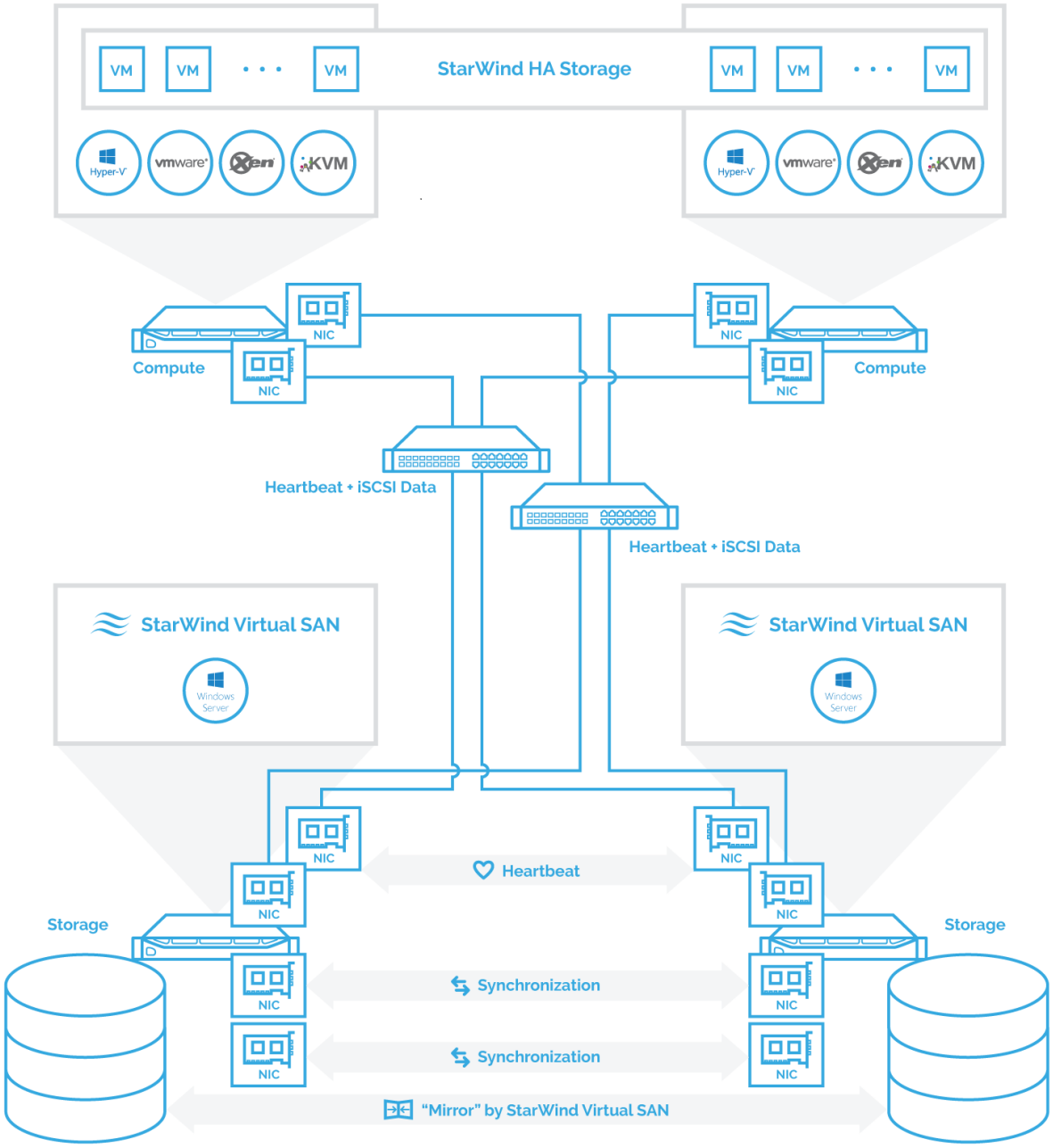

Fig. 3: Compute and storage separated configuration: 2- node cluster connected to a 2 – node StarWind Virtual SAN, redundant direct connections are used for synchronization channels.

NOTE: Synchronization and iSCSI/Heartbeat links can be connected either to the redundant switches or directly between the nodes (recommended).

The number of cluster nodes is limited by the hypervisor vendor recommendations

Cabling considerations

Shielded cabling (e.g., Cat 6a or higher) has to be used for all network links intended for StarWind Virtual SAN traffic. Cat. 5e cables are not recommended. StarWind Virtual SAN does not have specific requirements for 10/40/56/100 GbE cabling. In order to clarify which cabling type to use, please, contact the networking equipment vendor for recommendations.

Teaming and Multipathing best practices

StarWind Virtual SAN does not support any form of NIC teaming for resiliency or throughput aggregation.

All configurations on the diagrams shown in the “Networking Layouts Recommendations” section are configured with MPIO. In Compute & Storage Separated configurations, the recommended MPIO mode is Round-Robin. In HyperConverged configurations, the recommended MPIO mode is Failover Only or Fixed path**.** The Least Queue Depth MPIO policy can also be used in Microsoft environments to get better performance.

NOTE: Multipathing is not designed to show linear performance growth along with increasing the number of network links between the servers.

To aggregate synchronization throughput and achieve network redundancy, StarWind Virtual SAN can use multiple non-bonded network interfaces.

Synchronization channel recommendations

The synchronization channel is a critical part of the HA configuration. The synchronization channel is used to mirror every write operation addressing the HA storage. It is mandatory to have synchronization link throughput equal or higher than the total throughput of all links between the client servers and the Virtual SAN Cluster.

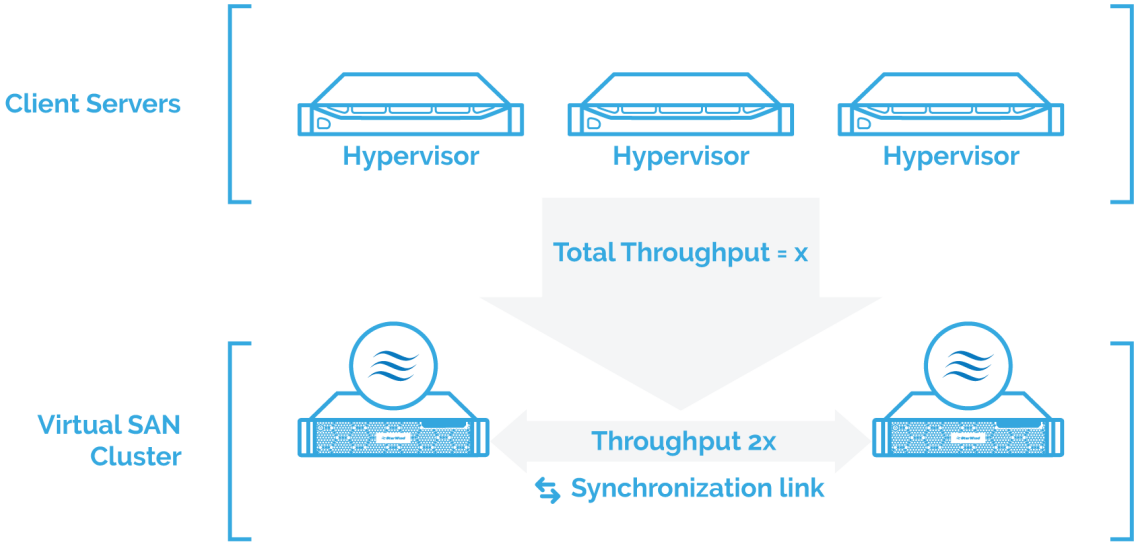

Fig. 4: 2-way mirror synchronization link throughput

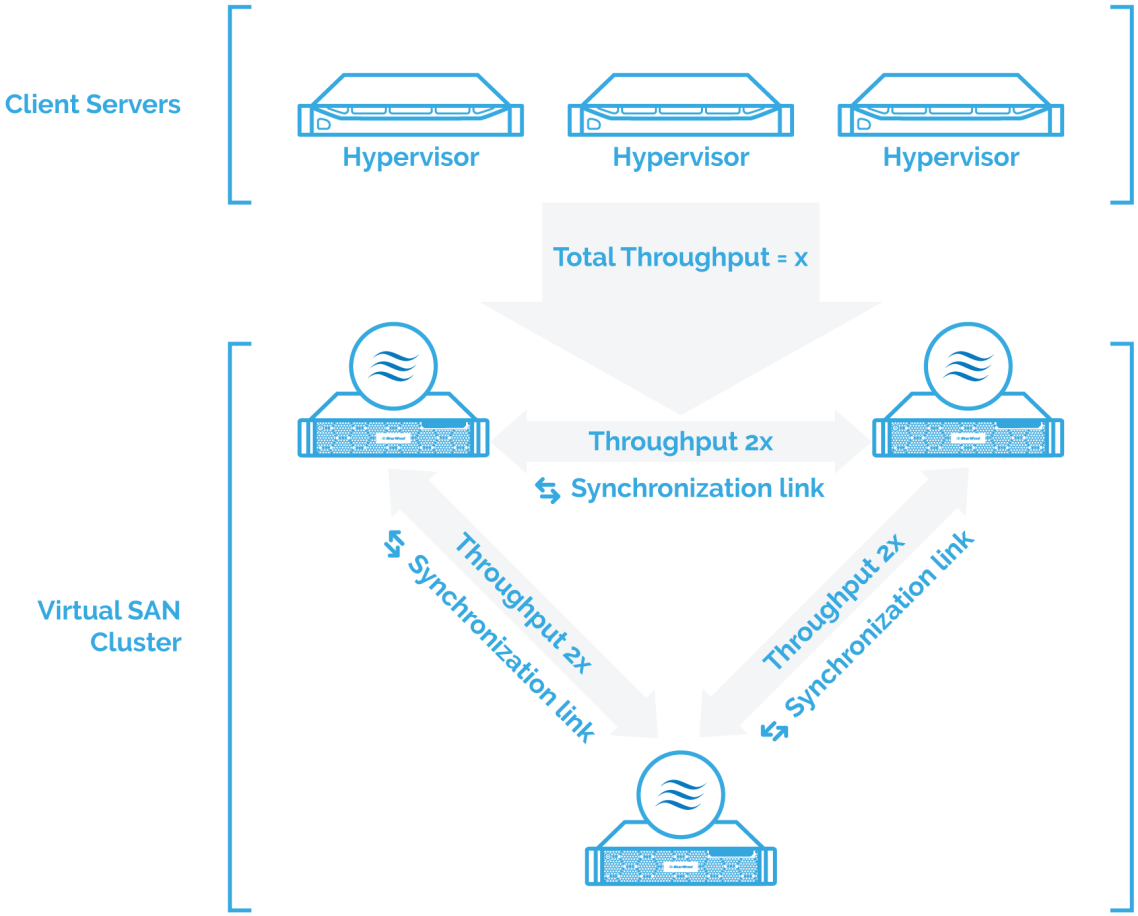

Fig. 5: 3-way mirror synchronization link throughput

For the HA storage, the maximum performance is limited by 3 factors:

1. Performance of the storage arrays used for HA storage.

2. The total performance of iSCSI multipath links from the client servers.

3. Synchronization channel performance.

Synchronization channel performance should exceed or equal the total performance of iSCSI multipath links from the client servers. In the real-world scenarios (1) may be slower or faster than (2) or (3), but at this point, the user has to understand that the HA device performance will be limited by the smallest value from the three mentioned above.

The HA device IO response time directly depends on the synchronization link latency. Certain hypervisor vendors have very strict requirements for storage response time. Exceeding the recommended response time limits can lead to the various application or virtual machine issues. In addition, certain features (e.g., Microsoft Hyper-V Live Migration or VMware vSphere HA) may fail or work incorrectly if the storage response time is out of the recommended limits.

The maximum Synchronization network latency should not exceed 5 ms.

Heartbeat channel recommendations

Heartbeat is a technology that allows avoiding so-called “split-brain”, situations when the HA cluster nodes are unable to synchronize but continue to accept write commands from the initiators. With StarWind Heartbeat technology, if the synchronization channel fails, StarWind attempts to ping the partner nodes using the provided heartbeat links. If the partner nodes do not respond, StarWind assumes that they are offline. In this case, StarWind marks the other nodes as not synchronized, and all HA devices on the node flush the write cache to the disk to preserve data integrity in case the node goes out of service unexpectedly.

If the heartbeat ping is successful, StarWind blocks the nodes with the lower priority until the synchronization channels are re-established. This is accomplished by designating node priorities. These priorities are used only in case of a synchronization channel failure and are configured automatically during the HA device creation. Please note that these settings have no effect on the multipath IO distribution between the HA nodes.

In order to minimize the number of network links in the Virtual SAN cluster, the heartbeat can be configured on the same network cards with iSCSI traffic. Heartbeat is only activated if the synchronization channel has failed and, therefore, it cannot affect the performance.

It is recommended to enable additional heartbeat connections for other physically separated links and network adapters between the HA SANs, except the ones used for HA device synchronization.

Storage Considerations

It is critical to identify the storage requirements properly. Doing this includes two factors: the real storage capacity and storage performance. The number one objective for the administrator deploying Virtual SAN is performance and capacity planning.

Performance: Calculate the approximate IOPS number the system needs to sustain. In addition, it is never a bad idea to add a power reserve with plans for future growth in mind.

Capacity: Estimate how many terabytes of data is to be stored.

StarWind recommends using the identical the storage configuration to the one used in the HA SAN cluster. This includes RAID controllers, volumes, and settings, as well as OS-level partitioning.

This section covers 2 configuration approaches:

- FLAT image file architecture – traditional data layout on the disks. The storage used for the general purpose.
- Log-Structured File System (LSFS) architecture – log-structured-based data layout. Recommended for intense write workloads. VM-centric file systems don`t support certain workload types.

Each approach assumes different storage architecture.

RAID controllers

There is no preferred vendor for RAID controllers. Therefore, StarWind recommends using RAID controllers from any industry-standard vendor. There are two basic requirements for the RAID controller with Highly Available SAN. They are discussed one-by-one below.

FLAT image file configuration:

- Write-Back caching with BBU depending on disk type
- RAID 0, 1, and 10 are supported for a spindle and all-flash arrays
- RAID 5, 50, 6, and 60 are only supported for all-flash arrays

LSFS configuration:

- Write-back caching with BBU
- RAID 0, 1, 10, 5, 50, 6, 60 support

Software RAID controllers are not supported for use with StarWind Virtual SAN.

StarWind Virtual SAN supports Microsoft Storage Spaces.

More information about RAID configuration is available in KB article: Recommended RAID settings for HDD and SSD disks

Stripe Size

iSCSI uses 64K blocks for network transfers. It is recommended to align the stripe size with this value. Modern RAID controllers often show similar performance independent of the used stripe size. However, it is still recommended to keep the stripe size aligned with 64K to change a full stripe with each writes operation. This increases the lifecycle of flash arrays and ensures the optimal performance of spindle-disk arrays.

Volumes

There are two data separation approaches in StarWind Virtual SAN. One is to keep both OS and StarWind device images on one physical volume, but segregate those using partitions. The second option is to segregate the OS from the dedicated disk/RAID array. Both options are supported. This applies to both FLAT image file and LSFS configurations.

Partitioning

It is recommended to use a GUID Partitioning Table (GPT) when initializing the disks used for StarWind devices. This allows creating volumes bigger than 2 TB and makes it possible to expand the partitions without taking the volume offline. This recommendation applies to both FLAT image file and LSFS configurations.

Filesystem Considerations

The general recommendation for Windows Server 2012 and above, is to use NTFS filesystem for the logical drive that contains Virtual SAN devices.

For hybrid deployments that use Windows Storage Spaces “Automated Tiering” on Windows Server 2019 it is recommended to use ReFS filesystem for the logical drive that contains Virtual SAN devices.

Virtual disk device sector size

StarWind supports underlying storage with both 512B and 4KB physical sector size. However, to optimize interoperability and performance, it is necessary to select the sector size of the underlying storage when creating virtual disks in the StarWind Management Console. The value can be picked when choosing the location to store the virtual disk file(s).

When creating a virtual disk device on storage spaces, please, use the 4K sector size option.

Benchmarking tools

It is critical to benchmark the disks and RAID arrays installed in the StarWind servers to avoid possible performance problems after deploying StarWind Virtual SAN into production. Make sure the array performance does not have abnormal drops on mixed read and write operations and random write tests. The local array benchmark should be used as a reference point for judging the performance of an HA device.

Please note that file copy is not a valid way to benchmark the performance of neither local nor iSCSI attached storage. Always use disk benchmarking tools like Diskspd, IOmeter, FIO, or VDbench to get the relevant information about the storage performance.

A network performance issue discovered after the HA SAN deployment often becomes a stopper. It often is nearly impossible to diagnose and fix the problem without putting the whole SAN infrastructure offline. Therefore, every network link and every disk in a Virtual SAN environment has to be checked to operate at peak performance before the system is deployed in a production environment. Detailed guidelines for Virtual SAN benchmarking can be found in the StarWind Virtual SAN benchmarking guide.

HA Device Considerations

Size Provisioning considerations

There is no strict requirement for the size of the HA devices created with StarWind Virtual SAN. Creating one big HA device that consumes all the available space on the SAN can cause management inconvenience. It is not an issue for devices up to 5-6 TB in size. Bigger devices can cause the inconvenience mentioned above due to increased full synchronization times. Using bigger devices also makes granular VM/application restore after outages or major failures more difficult. Allocating the mission-critical VMs or applications on the separate HA devices can make the management easier.

Since the HA caching is provisioned per device, segregating the devices according to the application load profiles also allows getting better utilization of the memory allocated for HA device caching.

For Hyper-V environments, it is a best practice to create at least two HA device per Hyper-V/Microsoft Failover Cluster for optimum performance.

FLAT image file configuration: Device size limited by the underlying FS maximum file size.

RAM consumption and Caching

StarWind HA is designed to show peak performance with Write-Back caching. Each written block is first cached on the local SAN node and then is synchronized with the second node cache. Once these two operations are done, StarWind considers the block as written. Along with providing great performance improvements, write-back caching also introduces specific requirements for underlying hardware stability. UPS units have to be installed to ensure the correct shutdown of at least one StarWind Virtual SAN node in case of a power outage to make sure that cached data will not be lost.

With Write-Through caching, write operations can become significantly slower and fully depend on the underlying array performance. This happens because a write operation is only confirmed when the block is written to the disk itself. Although write-through caching gives no boost to the write performance, it does not depend on power stability (compared to write-back), and still maintains the read cache, which balances the read/write access distribution to the underlying disk array.

The cache effectiveness depends on the cache size to working set size ratio. It is recommended to configure the cache size that is appropriate to the working set size.

The minimum recommended RAM cache size is 128 MB.

LSFS devices consume additional 7.6 GB of RAM per terabyte of stored data. This value is hard-coded and cannot be changed from StarWind Management Console.

Please, keep in mind that write cache size affects the time server needs for a graceful shutdown.

During the shutdown, the server needs to flush the RAM cache to the disk. Shutdown time can be calculated as the total amount of RAM provisioned as write-back cache divided on the performance of the disk array under 100% random 100% 4K write load.

More information about StarWind cache operational principles is available on this link.

CONCLUSION

Following StarWind Virtual SAN best practices allows configuring highly-available storage for the virtualized production environments and minimize or avoid downtime associated with storage or host failures. Also, it allows getting the HA device performance similar to the underlying storage performance that is vital for cluster environments which run intense production and need storage failover in the event of hardware/software failures.

Contacts

US Headquarters	EMEA and APAC
1-617-449-77 17 1-617-507-58 45 1-866-790-26 46	+44 203 769 18 57 (UK) +34 629 03 07 17 (Spain and Portugal)

Customer Support Portal: https://www.starwind.com/support
Support Forum: https://www.starwind.com/forums
Sales: sales@starwind.com
General Information: info@starwind.com

StarWind Software, Inc. 100 Cummings Center Suite 224-C Beverly MA 01915, USA

www.starwind.com