StarWind High Availability Best Practices

Max Craft
Solution Engineer, StarWind Software, Inc.

Bohdan Naumets
Head of Q&A, StarWind Software, Inc.

CONTENTS

INTRODUCTION 2

DESIGN PRINCIPLES 3

HOST CONSIDERATIONS 3

HARDWARE DIFFERENCES 3

OS DIFFERENCES 3

STARWIND VERSIONS 3

OS CONFIGURATION 4

NETWORKING 4

1/10/40 GBIT CONSIDERATIONS 5

NETWORKING LAYOUT RECOMMENDATIONS 5

SYNCHRONIZATION CHANNEL RECOMMENDATIONS 8

HEARTBEAT CHANNEL RECOMMENDATIONS 9

BENCHMARKING TOOLS 9

STORAGE CONSIDERATIONS 9

RAID CONTROLLERS 10

PARTITIONING 10

BENCHMARKING TOOLS 10

HA DEVICE CONSIDERATIONS 11

SIZE AND PROVISIONING CONSIDERATIONS 11

CACHING 11

INITIALIZATION OPTIONS 11

CONTACTS 12

INTRODUCTION

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

StarWind High Availability solution makes it possible for customers to minimize or avoid downtime of the storage area network. StarWind HA also enables maintenance without storage downtime. This is achieved by putting together multiple StarWind servers into a fault-tolerant storage cluster and by providing instant recovery from hardware/ software failures and power outages.

Using internal and native OS tools StarWind HA constantly monitors the state of all the network links between the servers in an HA cluster. Should an HA cluster node fail, the failover is initiated from the client OS/Hypervisor side. This guarantees correct failover time and allows customers to create a centralized HA storage resource capable of proper failover without any additional configuration for individual client server hosts. StarWind also provides an internal heartbeat mechanism, which ensures proper HA cluster node isolation in case of synchronization network failures.

High Availability easily combines with native Windows File Share services or Scale-Out File Server (SoFS) and can act as an uninterrupted file share service for multiple non-clustered client computers.

StarWind HA has multiple advantages over traditional storage failover solutions:

- Hardware independence
- Reduced TCO
- Minimal setup time
- Ease of use and management
- Instant failover and failback

DESIGN PRINCIPLES

Host considerations

Proper equipment selection is a very important step of the SAN architecture planning. Always choose the appropriate equipment for the tasks you put on the HA SAN. Note that overestimating the SAN requirements is not always a good practice. It can turn out that the equipment purchased according to these estimations will never be used effectively. Ensure to always plan for the scalability of the SAN servers that you are purchasing. Keep in mind all possible storage extension and network upgrades in the future.

Hardware differences

StarWind HA performs in active-active mode and, therefore, it is a best practice to use identical hardware for all the nodes participating in an HA storage cluster. In some cases, the HA node has disks or network interfaces slower than the rest of the HA nodes in the cluster. This may be caused by using high-speed drives on one server in order to improve the read performance of the HA array. In this case, the customer needs to use Asymmetric Logical Unit Assignment (ALUA) to achieve the optimal performance of the asymmetric configuration. ALUA marks certain HA nodes as an optimal or non-optimal path. The client server then uses these marks to optimize disk access performance. Please refer to the Asymmetric configurations chapter of the document for more information.

OS differences

When configuring a StarWind iSCSI SAN HA cluster it is a best practice to install the same edition of the latest operating system version on all the servers participating in the HA cluster. An edition difference is possible: e.g. SAN 1 running Windows 2008 R2 Standard and SAN 2 running Windows 2008 R2 Enterprise. Please note that certain Windows editions are not supported. Please refer to the following System requirements page for the list of supported operating systems: http://www.starwindsoftware.com/system-requirements.

StarWind versions

It is a best practice to have the same version of StarWind iSCSI SAN installed on all nodes in the HA storage cluster. Always update all StarWind iSCSI SAN servers in the environment to avoid version and build mismatch. Version mismatch should be avoided due to the differences in the HA device compatibility, performance, and operational features.

OS configuration

Every SAN has special requirements for uptime and availability. In order to fulfill these requirements, the user needs to modify certain settings of the operating system and its embedded tools and utilities.

Updates

Since a SAN has strict requirements for maintenance downtime, Windows users are required to control all the update processes on the server. All the automatic updates of the OS should be either disabled or set to "Check for updates but let me choose whether to download and install them". Never apply OS updates to more than one node of the HA cluster at a time. After applying OS updates to a node, verify that the functionality is intact and that all iSCSI devices are resynchronized and successfully reconnected to the client servers. After completing the verification, you can start the update process on the next HA node.

Firewalls

By default, StarWind operates through TCP ports 3260 and 3261. Port 3260 is used for iSCSI traffic and port 3261 is used for the StarWind management console connection to the server. The StarWind installer opens these ports during the initial installation. If third party firewall software is used, the user may need to open these ports manually. Later on, these ports can be changed through the StarWind management console. Please make sure that appropriate Firewall rule modifications are performed if either iSCSI traffic or Management port is changed.

Additional software

It is not recommended to install any kind of third party applications on the server running StarWind iSCSI SAN.

Exceptions here are benchmarking tools, remote access utilities, and hardware management utilities such as Network card managers or RAID controller management packs. If you have any doubts about the software that you want to install on the SAN, please contact the StarWind support department.

Asymmetric configurations

For certain tasks, a StarWind HA cluster can be configured in an asymmetric way. E.g.: all HA nodes have identical network performance, but one node has a faster disk subsystem. If used asymmetrically, it allows users to increase the read performance of the HA SAN, while keeping the TCO lower. In this case, ALUA has to be configured when creating the HA devices on this HA SAN. With ALUA, all the network paths to the storage array remain active, but the client only writes data over those paths, which are marked as optimal in the ALUA configuration. This eliminates writes to the SAN node with a slower disk subsystem and thereby increases the overall performance of the SAN.

Networking

The network is one of the most important parts of the SAN. Determining a correct network bandwidth for your SAN is the #1 task along with finding the right amount of IOPs that your storage needs to produce in order to fulfill the requirements of the environment where it will be deployed.

1/10/40 Gbit considerations

The calculations below use IOPS relevant for 4KB block size. If the client server uses block size of a different value, you can calculate the final IOPS using the following formulas:

> IOPS = (MBps Throughput / KB per IO) x 1024
>
> MBps = (IOPS x KB per IO) / 1024

According to the amount of IOPs or MB/s you have calculated for your future SAN, you need to choose the equipment that will not cause bottlenecks on the network level. E.g., if your cluster consumes 64K IOPS (which is approx. 250 MB/s with 4K block), then a 1Gb Ethernet is not an option anymore. Networking demand grows along with IOPS demand, and after ~333,000 IOPS (1,300 MB/s with 4K block), a single 10GbE card becomes a bottleneck. Below is a table showing the recommended network equipment throughput depending on the IOPS requirements of the client applications.

IOPS (4K)	Network
1–26,000	1 Gigabit
26,000–52,000	2 x 1 Gigabit (MPIO)
52,000–282,000	10 Gigabit
282,000–564,000	2 x 10 Gigabit (MPIO)
564,000 and higher	40 Gigabit

Networking layout recommendations

The general purpose of a highly available solution is 24/7 uptime with zero downtime for maintenance and upgrades. Thus, it is very important to understand that high availability is not achieved with just clustering the servers. It is always a combination of redundant hardware, special software, and a set of configurations that makes the solution truly HA. Below you can find reference architecture diagrams showing the network redundancy configuration for HA. These network layout configurations are considered the best practice of HA SAN design.

Fig. 1: Two-node hypervisor cluster connected to a two-node StarWind HA cluster; direct connection is used for synchronization channels

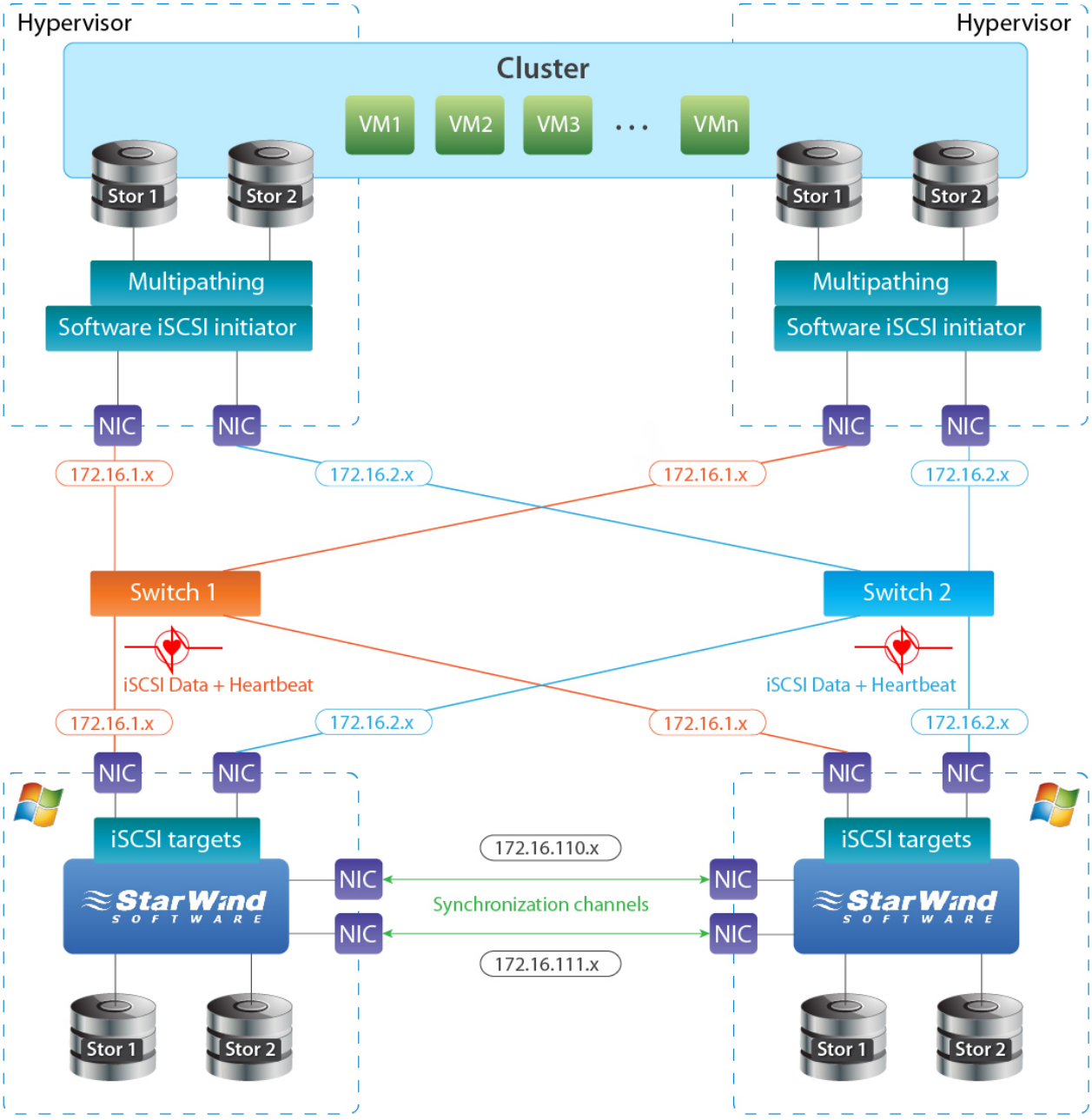
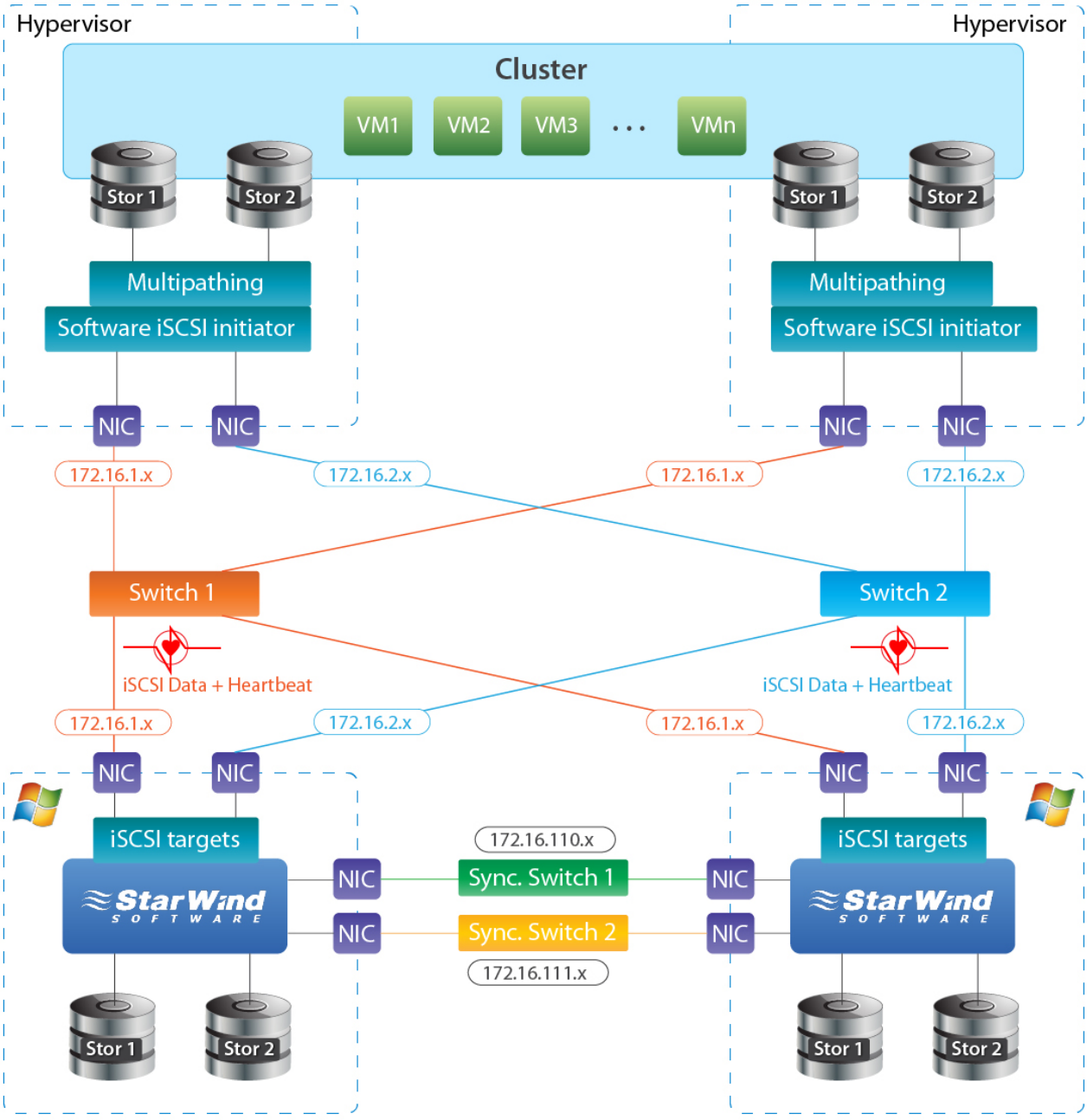
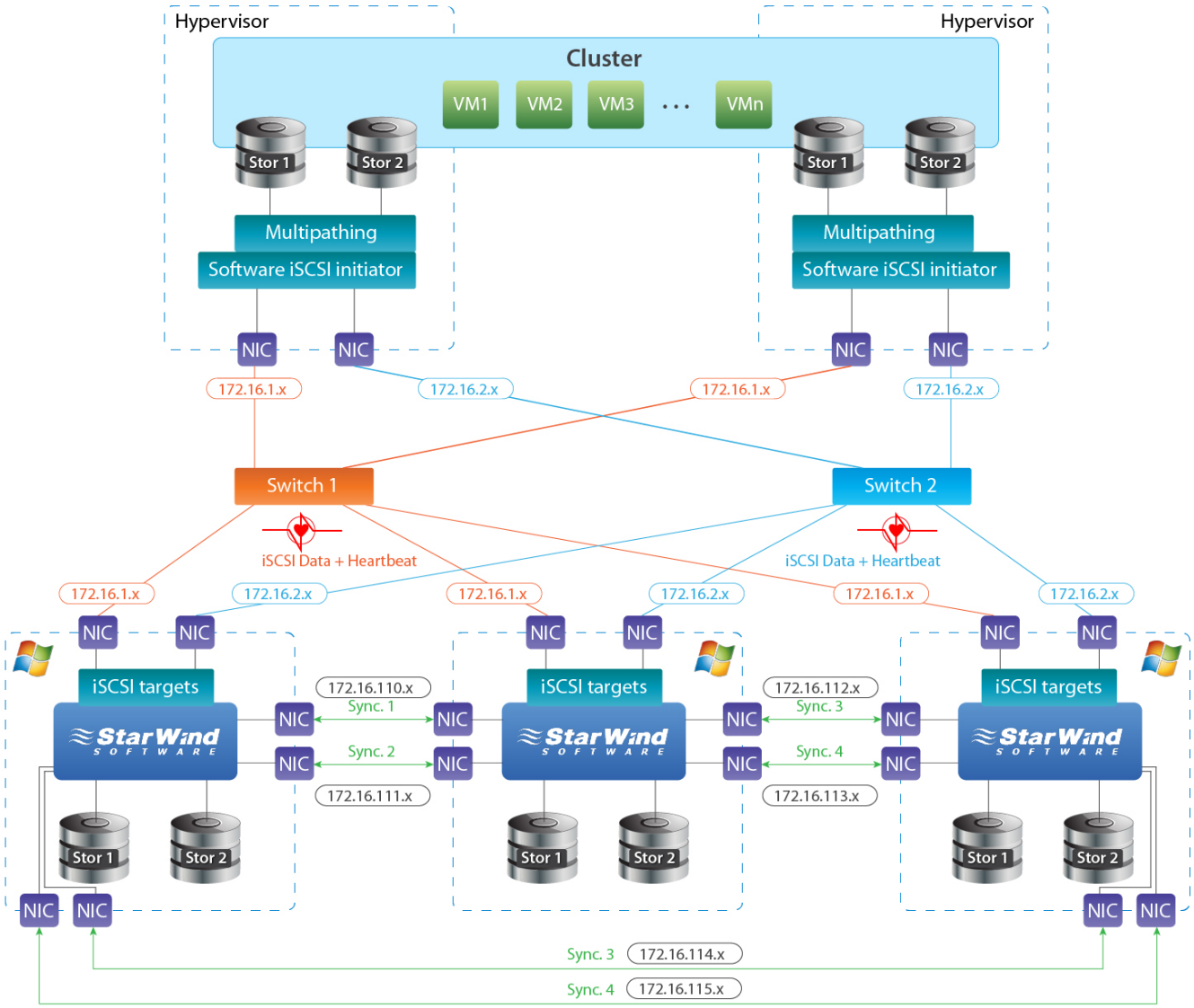


Fig. 2: Two-node hypervisor cluster connected to a two-node StarWind HA cluster; switch and link are used for fault tolerance



Note: *Switch redundancy recommendations also apply to the synchronization channels. If a switch is used for the synchronization channel, it has to be redundant. The configuration illustrated by the diagram above is often used in situations where the pair "first SAN node - cluster node" is geographically separated from the pair "second SAN node - cluster node".*

Fig. 3: Two-node hypervisor cluster connected to a three-node StarWind HA cluster



***Note**: All of the diagrams above show only the SAN connections. LAN connections, internal cluster communication, and auxiliary connections have to be configured using separate network equipment. Best practices for the networking configuration inside the cluster should be configured according to your hypervisor guidelines and best practices.*

Cabling considerations

Shielded cabling (e.g. Cat 6a or higher) has to be used for all the network links within the storage area network. Cat 5e cables can expose the network links to external electromagnetic interference and, therefore, it is not recommended for use in a SAN.

Teaming and Multipathing best practices

Multiple online resources as well as vendors state that the use of NIC teaming in iSCSI networks does not justify itself. Moreover, NIC teaming appeared to increase the link response time, limit throughput, and cause a connection loss in some cases. That's why it is not recommended to use teaming on any link within your SAN.

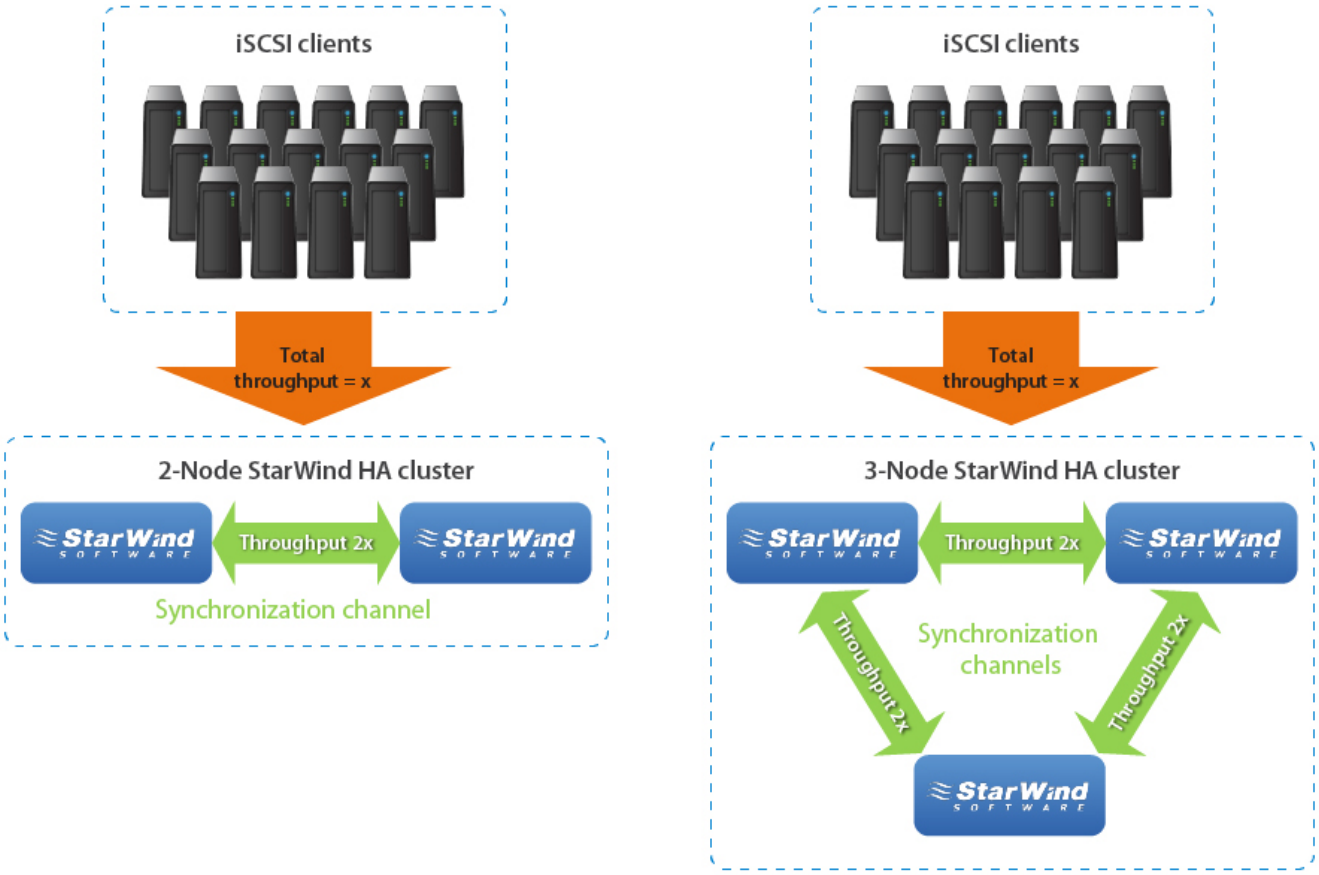
Instead, multipathing shows good results in the storage area networks. All connections pictured in the diagrams shown above (Figs.1-3) are configured to multipath IOs.

The preferred MPIO mode is round-robin. It allows you to distribute the load evenly between all the links participating in the connection. Multipathing does not guarantee uniform performance growth along with increase of network links between the servers.

Synchronization channel recommendations

The synchronization channel is a very important part of the HA configuration. StarWind uses it to mirror every write operation addressing the HA devices. It is a general requirement that the synchronization link throughput has to be equal or higher than the throughput of all client servers connected to the HA SAN cluster.

Fig. 4: Proper synchronization channel bandwidth measurement in a two- and three-node HA SAN cluster



For the HA device, the maximum write speed is limited by 3 factors:

1. Speed of the storage arrays installed on the SAN servers
2. Total throughput of round-robin multipathed iSCSI links from the client servers
3. Synchronization channel throughput

It is required that 3 ≥ 2. In 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 depends directly on the synchronization link latency. Certain hypervisor vendors have very strict requirements on the shared storage response time. Exceeding the recommended response time limits can lead to various application or virtual machine problems. In addition, certain features (e.g. Microsoft Hyper-V Live Migration or VMware vSphere HA) may fail or work incorrectly if the HA device response time is not within the recommended guidelines.

The maximum synchronization channel latency values are provided below:

- HA SANs are located in different buildings/data centers – 5 ms
- HA SANs are located in one building/data center – 3 ms

Heartbeat channel recommendations

Heartbeat is a technology that enables you to avoid the so-called "split-brain" scenario, when the HA cluster nodes continue to accept write commands without synchronizing them. If the synchronization channel fails, StarWind attempts to ping the partner node over the specified heartbeat link. If the partner node does not respond, the system assumes that it is offline. In this case, all HA devices on the remaining node flush the write cache to the disk and continue to operate in Write-Through caching mode.

If the heartbeat ping is successful, StarWind blocks one node until the synchronization channel is re-established. In order to minimize the number of network links in the HA SAN cluster, heartbeat is configured to run on the same network cards with iSCSI traffic. Heartbeat activates only if the synchronization channel fails and, therefore, it cannot affect the performance.

Since heartbeat is located on the same network links as the regular SAN traffic, it has the same bandwidth and latency requirements as the regular link in the storage area network.

It is recommended to enable the heartbeat feature for all the links between the HA SANs, except the ones used for HA device synchronization.

Benchmarking tools

A network performance issue discovered after the HA SAN deployment is a big problem. It is often nearly impossible to diagnose and fix the problem without putting the whole SAN infrastructure on hold. Therefore, every network link in the SAN has to be checked to operate at peak performance before the SAN is deployed into the production environment. Detailed information about the network benchmarking can be found in the SAN benchmarking guide.

Storage considerations

One of the two cornerstones keeping the whole SAN infrastructure together is storage. It is critical to properly measure your storage requirements. This includes two factors: the first is the actual storage capacity, and the second is performance. The number one objective of every storage administrator intending to implement a SAN is performance and capacity planning.

> **Performance**: Calculate the most exact IOPS amount for your future system. In addition, it is a good idea to add some power reserve with plans for future growth in mind.
>
> **Capacity**: Figure out how many terabytes of data you need to put on the HA SAN.

Keep in mind, that it is always possible to increase the storage capacity. Therefore, if your budget does not allow for building a big SAN in one step, you can plan for a SAN system, which will be easily scalable up to the desired capacity as time goes by. In this case, include the growth plans into the hardware specifications of your future SAN. Knowing the maximum amount of storage that your SAN can handle will enable proper scaling without the risk of exceeding the storage limits.

StarWind recommends the storage configuration used in the HA SAN cluster to be similar or identical. This includes the use of similar RAID controllers, identically configured RAID sets, and partitioning. If you cannot achieve the above mentioned recommendations and one of the SANs in the HA cluster operates slower than the rest, apply ALUA. Please refer to the Asymmetric configuration chapter of the document.

RAID controllers

There is no preferred vendor for RAID controllers. Therefore, StarWind recommends using RAID controllers from any industry-standard vendor. There are 2 basic requirements to a RAID controller for an HA SAN:

> Write-Back caching with battery backed up cache
>
> RAID0, RAID1, and RAID10 support.

StarWind iSCSI SAN does not support software RAID controllers.

Stripe Size

ISCSI uses 64K blocks for network transfers. It is recommended to align your stripe size with this value. Modern RAID controllers often show similar performance over iSCSI, independent of the stripe size used.

However, it is still recommended to keep the stripe size aligned with 64K to change a full stripe with each write operation. This increases the lifecycle of the array and ensures optimal performance.

Volumes

It is recommended to segregate the OS and the SAN storage volumes to different physical media. Use a separate hard drive, RAID1, or a USB stick to run the OS. It is a general recommendation to create a maximum of 2 RAID volumes per physical RAID controller on the SAN server.

Partitioning

It is recommended to use a GUID Partitioning Table (GPT) when initializing the disks on the SAN server. This allows you to overcome the 2TB per volume limitation for master boot record (MBR), and makes it possible to expand the partitions without any third party utilities after RAID capacity extensions.

Benchmarking tools

It is critical to benchmark the disks and RAID arrays installed in the HA SAN servers to avoid possible performance problems after deploying the HA SAN into the production environment. Make sure that the array performance does not have abnormal performance drops on mixed read and write operations, as well as on random write tests. The local array benchmark can later be used as a reference to the performance of an iSCSI HA device.

HA device considerations

HA devices are always configured according to the requirements of the client-side applications. VMware, Hyper-V, and XenServer as well as solutions designed by the other vendors, all have specific requirements for the iSCSI storage they work with. StarWind allows users to achieve the maximum level of flexibility when deploying HA storage for their production environment.

Size and provisioning considerations

There is no strict requirement for the size of the HA device that the user is creating with StarWind iSCSI SAN. Creation of one big HA device that occupies all the available space on the SAN can cause management inconvenience. It is not an issue for devices up 5-6TB but for bigger devices complications can arise due to increased full synchronization time. The use of bigger devices also increases a granular VM/application restore time after outages or major failures. Segregating your mission-critical VMs or applications to separate HA devices can simplify the management. Since the HA caching is provisioned per device, segregating the devices according to the application load profiles also allows for better utilization of the memory allocated for HA device caching.

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 synchronized with the second node's cache; only after that the HA device considers the block to be written. Along with performance improvements, Write-Back caching also brings in specific requirements to power stability. UPS units have to be installed to guarantee the proper shutdown of at least one HA SAN node in case of a power outage.

With Write-Through caching, write operations can become significantly slower, as well as much more dependent on the underlying array performance. This happens because the write is only confirmed when the block is not only cached but also 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 read cache, which balances the read/write access distribution to the underlying disk.

It is recommended to provision 256MB–1GB of caching per each terabyte of the HA device's size. Although, the maximum recommended cache size is 3GB. For most scenarios, bigger cache is not utilized effectively.

Initialization options

The HA device has 3 initialization options:

> **Initialize**: Also called "fill with zeroes", this option initiates the so-called "eager zeroing" of the HA device. This can increase the initialization time. It also makes sure that any information, that is previously written to the disk of the HA SAN node cannot be recovered when the HA device is connected to the client servers.
>
> **Don't initialize**: This option simply deploys the HA device without filling it with zeroes.
>
> **Synchronize from**: This option allows the user to synchronize the existing data from one HA SAN node to the partner(s). Simply select the synchronization source node and StarWind will carry out block-level synchronization of the whole virtual disk.

Although the "initialize" option is not mandatory, in some cases, it provides a slight performance increase. It is recommended to check it on small HA devices before initial deployment to the production environment.

CONTACTS

StarWind Software Inc.
301 Edgewater Place,
Suite 100,
Wakefield, MA 01880
Phone: +1 617-449-7717
Fax: +1 617-507-5845

E-mail: sales@starwindsoftware.com
Website: www.starwindsoftware.com