Synology

White Paper

Synology High Availability (SHA)

Based on DSM 6

Synology Inc.

Synology_SHAWP_ 20170807

Table of Contents

Chapter

1

Introduction

Uninterrupted availability is a critical goal for all businesses; however, as many as 50% of SMBs worldwide remain unprepared in the case of disaster[1]. Moreover, downtime costs a median of 12,500 USD daily. Assuming a median of six downtime events per year, the cost of unpreparedness begins to stack up.

The **Synology High Availability** solution helps users overcome this hurdle by ensuring non-stop storage services with maximized system availability to decrease the risk and impact of unexpected interruptions and costly downtime.

[1] Symantec 2011 SMB Disaster Preparedness Survey, **http://www.symantec.com/about/news/resources/press_kits/detail.jsp?pkid=dpsurvey**

Chapter 2

High-Availability Clustering

2.1 Synology High-Availability Cluster

The Synology High Availability solution is a server layout designed to reduce service interruptions caused by system malfunctions. It employs two servers to form a "**high-availability cluster**" (also called "HA cluster") consisting of two compatible Synology servers. Once this high-availability cluster is formed, one server assumes the role of the active server, while the other acts as a standby passive server.

Full data replication is required to be performed once the cluster is successfully created.

There are a few tips on reducing the time required for the initial data replication:

- **Increase the bandwidth of heartbeat connection:** Heartbeat connection is the main method for data replication from the active server to the passive server. Therefore, increasing and optimizing the bandwidth of the heartbeat connection can effectively reduce the time required for data replication.
- **Remove volumes or iSCSI LUNs that are no longer needed:** The initial data replication can sometimes be skipped if you do not plan to keep data on the active server. In this case, simply remove the volumes and iSCSI LUNs before proceeding with the SHA cluster creation. Because there is no volume on the active server, the initial replication can be skipped, thus reducing the time spent on cluster creation.

2.2 Service Continuity

Once the high-availability cluster is formed, data is continuously replicated from the active server to the passive server. All files on the active server will be copied to the passive server. In the event of a critical malfunction, the passive server is ready to take over all services. Equipped with a duplicate image of all data on the active server, the passive server will enable the high-availability cluster to continue functioning as normal, reducing downtime.

2.3 Data Replication Process

Within the high-availability cluster, all data stored on internal drives or expansion units will be replicated. Therefore when services are switched from the active to passive server, no data loss will occur.

While data replication is a continual process, it has two distinct phases spanning from formation to operation of a high-availability cluster:

- **Phase 1:** The initial data replication during cluster creation or the replication of differential data when connection to the passive server is resumed after a period of disconnection (such as when the passive server is switched off for maintenance). During this phase, the initial sync is not yet complete, and therefore switchover cannot be performed. Data changes made on the active server during this initial replication are also synced.
- **Phase 2:** Real-time data replication after the initial sync has been completed. After the initial sync, all data is replicated in real-time and treated as committed if successfully copied. In this phase, switchover can be performed at any time.

During both phases of data replication, all data syncing is performed at block-level. For example, when writing a 10 GB file, syncing and committing is broken down to block-level operations, and completed piecemeal to ensure that the active and passive servers contain identical data. As all data is constantly maintained to be up-to-date, switchover can be accomplished seamlessly.

Data and changes to be replicated include:

- **NAS Data Services:** All file services including CIFS/NFS/AFP are covered.

- **iSCSI Data Services:** High-availability clustering supports iSCSI, including iSCSI LUN and iSCSI Target services.
- **DSM and Other Services:** Management applications, including Synology DiskStation Manager (DSM) and its other services and Add-On Packages, e.g., Mail Server, Directory Server, are also covered, including all settings and service statuses.

Chapter

3

High-Availability Cluster Architecture

3.1 Physical Components

Synology's High Availability solution constructs a cluster composed of two individual storage systems: an active and a passive server. Each server comes with attached storage volumes, and the two are linked by a "Heartbeat" connection that monitors server status and facilitates data replication between the two servers.

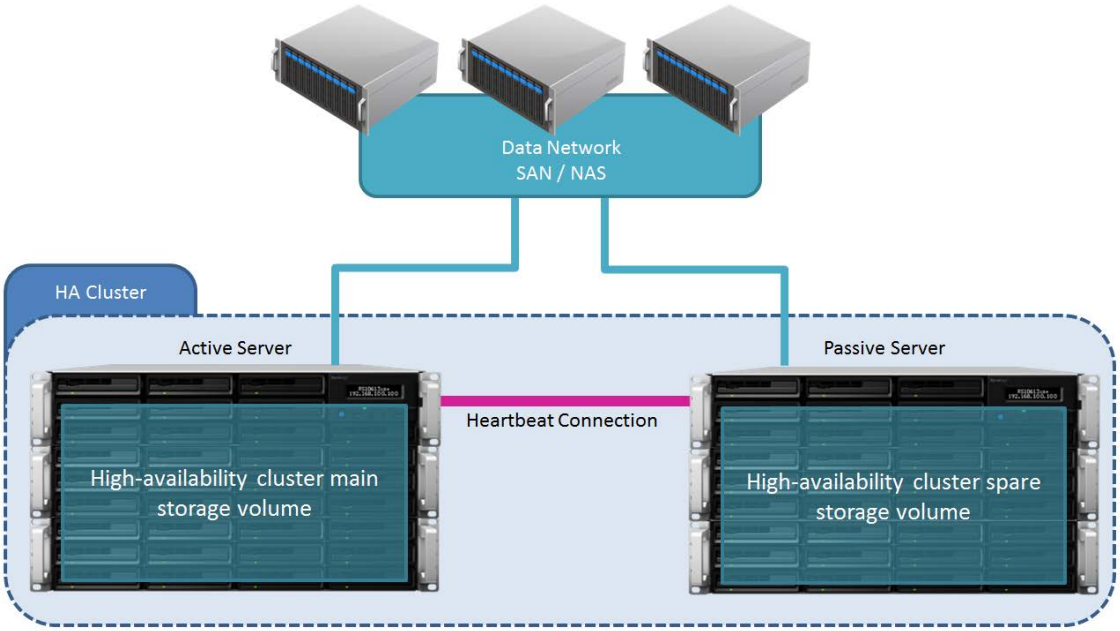


Figure 1: Physical components of a typical Synology High Availability (SHA) deployment

- **Active Server**: Under normal conditions, all services are provided by the active server. In the event of a critical malfunction, the active server will be ready to pass service provisioning to the passive server, thereby circumventing downtime.
- **Passive Server**: Under normal conditions, the passive server remains in standby mode and receives a steady stream of data replicated from the active server.
- **Heartbeat Connection**: The active and passive servers of a high-availability cluster are connected by a dedicated, private network connection known as the "Heartbeat" connection. Once the cluster is formed, the Heartbeat facilitates data replication from the active server to the passive server. It also allows the passive server to constantly detect the active server's presence, allowing it to take over in the event of active server failure. The ping response time between the two servers must be less than one ms, while the transmission speed should be at least 500 Mbps. The performance of the HA cluster will be affected by the response time and bandwidth of the heartbeat connection.
- The Heartbeat connection must be configured on the fastest network interface. For instance, if the servers are equipped with 10GbE add-on network cards, the Heartbeat connection must be configured by using 10GbE cards. In addition, it is strongly recommended that users build a direct connection (without switches) between two servers, the distance between which is usually shorter than 10 meters. If a HA cluster requires two servers with a larger distance, the heartbeat connection between two servers must have no other device in the same broadcast domain. This configuration can be achieved by configuring a separate VLAN on the Ethernet switch to isolate the traffic from other network devices. Please make sure that data connection and Heartbeat connection are in different loops lest they be interrupted at the same time when functioning.

Note: The passive server detects the presence of the active server via both the Heartbeat connection and data connection in order to prevent "split-brain" errors when the Heartbeat connection fails. A "split-brain" error occurs when both servers attempt to assume the role of active server, resulting in service errors.

- **Main Storage:** The storage volume of the active server.
- **Spare Storage**: The storage volume of the passive server, which continually replicates data received from the main storage via the Heartbeat connection.

High-availability safe mode

Instead of performing a complete replication, High-availability safe mode helps users to identify the new active server and re-build the cluster by syncing new data and modified settings from the active server to the passive server.

In high-availability safe mode, both servers and the IP addresses of the High-availability clusters will be unavailable until the split-brain error is resolved. Also, additional information will be shown, including (1) the difference of contents in the shared folders on the two servers, (2) the time log indicating when the server became active, and (3) the last iSCSI Target connection information. The information should be found in High Availability Manager or the read-only File Station. Thus, users would be able to identify the newly active server.

When the newly active server is selected, both servers will be rebooted. After that, all the modified data and settings on the active server will be synced to the passive server. Hence, a new healthy High-availability cluster shall be in place.

In addition, users can choose to make a complete replication from the active server to the passive server or they can unbind both of them.

To make a complete replication, users should choose one as the active server of the High-availability cluster and unbind the other. Once both servers are rebooted, the active server will remain in the High-availability cluster. The unbound server will keep its data and return to Standalone status. Please note that a complete replication will entail binding a new passive server onward.

When unbinding the two servers, users should save the data for each before switching to Standalone status.

Split-brain error

When a high-availability cluster is functioning normally, only one of the member servers should assume the role of active server. In this case, the passive server detects the presence of the active server via both the Heartbeat connection and data connection.

If all Heartbeat and data connections are lost, both servers might attempt to assume the role of active server. This situation is referred to as a "split-brain" error. In this case, connections to the IP addresses of the high-availability cluster will be redirected to either of the two servers, and inconsistent data might be updated or written on the two servers.

When any one of the Heartbeat or data connections is reconnected, the system will detect the split-brain error and data inconsistency between the two servers, and will enter high-availability safe mode.

Quorum server

A quorum server helps reduce the split-brain error rate. Users can assign another server to both the active and passive server as the quorum server. For example, a gateway server or DNS server is a good choice because they usually connect to both servers constantly. Please note that no application will be installed on the quorum server, which provides ping service only.

With a quorum server, the following circumstances will be controlled:

- If the passive server cannot connect to both the active and quorum servers, failover will not be performed in order to prevent split brain errors.
- If the active server cannot connect to the quorum server while passive server can, switchover will be triggered in order to achieve better availability.

3.2 Virtual Interface

When the two servers are combined into a high-availability cluster, a virtual interface -- unique server name and IP address -- shall be configured. This virtual interface allows hosts to access the cluster resources using a single namespace. Therefore, when a switchover is triggered and the provision of services is moved to the passive server, there will be no need to modify network configurations on hosts in the data network.

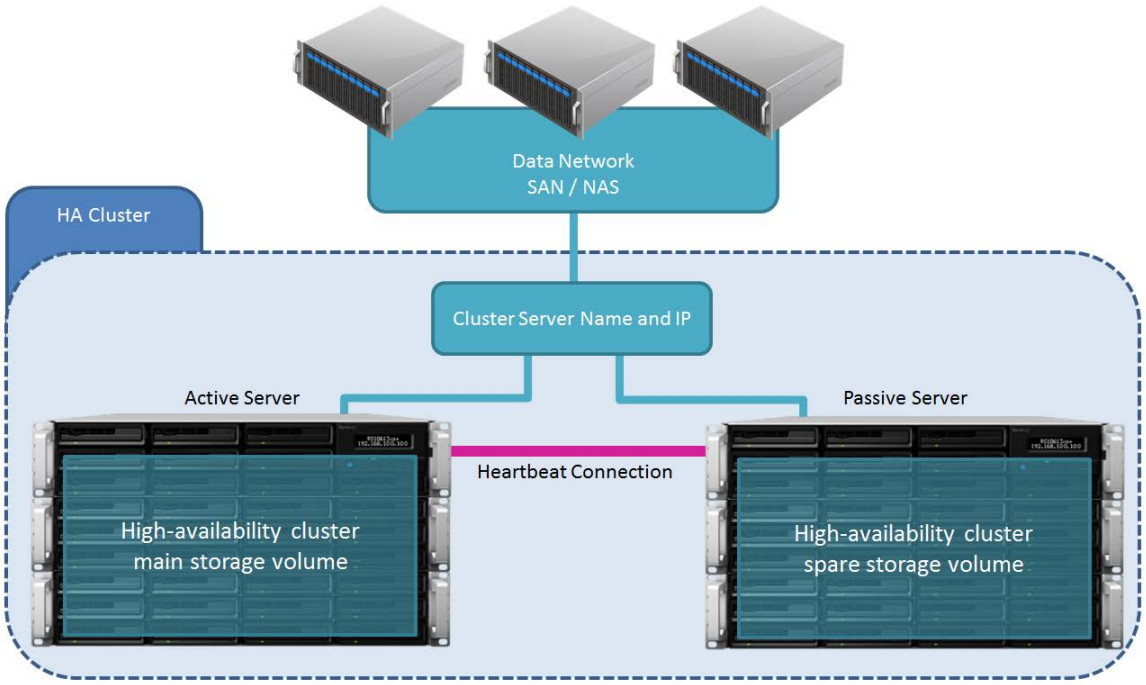


Figure 2: Hosts and NAS clients access a Synology High Availability (SHA) cluster through a single virtual interface

- **Cluster Server Name and IP addresses:** Servers in the cluster will share IP addresses and a server name, which should be used in all instances instead of the original IP addresses and individual server names.

3.3 Network Implementation

The physical network connections from the data network to the active server and passive server must be configured properly so that all hosts in the data network can seamlessly switch connections to the passive server in the event a switchover is triggered. The following section covers different configurations for various situations and Synology NAS models.

Network Implementation for Synology NAS with two LAN ports

In situations where both servers have two network ports only, one network port on each server will be occupied by the heartbeat connection, so each server will have only one port available for the HA cluster to connect to the data network. Therefore, there will not be sufficient network ports to accommodate redundant paths between the hosts in the data network and HA cluster. However, we still recommend using multiple paths to connect hosts to the data network, as well as more than one switch in your data network to provide redundancy.

Synology High Availability (SHA) provides an option to trigger a switchover when the active server detects network failure. When enabled, if connection failure occurs between the switch connected to the active server or the switch fails, service continuity will be maintained by switching over to the passive server (assuming the network connection of the passive server is healthy).

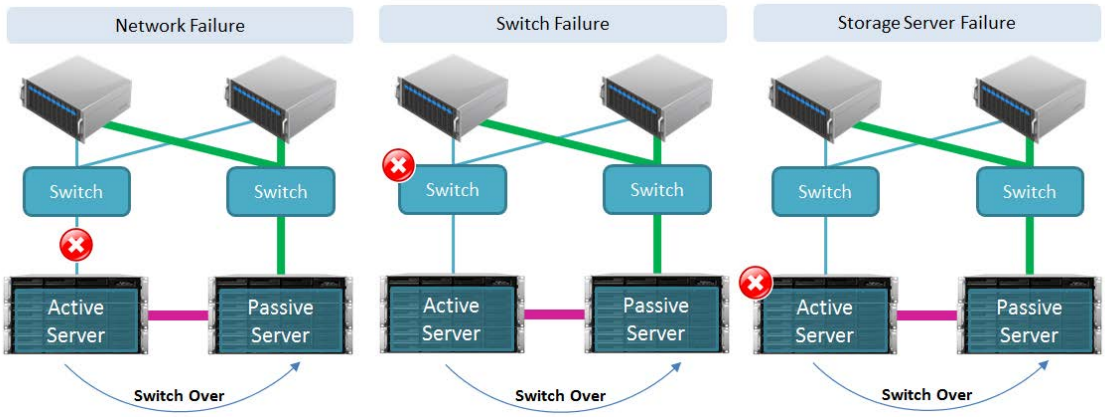


Figure 3: High-availability cluster network configuration on models with two LAN ports

Network Implementation for Synology NAS with four or more LAN ports

The best way to create a high availability environment is to use a Synology NAS with four network ports. In this instance, you can connect multiple paths between the hosts and HA cluster, providing a redundant failover path in case the primary path fails. Moreover, I/O connections between the data network and each clustered server can be connected to more than one port, providing a load balancing capability when all the connections are healthy.

Implementation for iSCSI storage

Connecting a host to more than one of the storage system's front-end ports is called "multipathing." By implementing Multipath I/O (MPIO) or Multiple Connection per Session (MC/S) on the iSCSI connection, you can deliver a high quality and reliable storage service equipped with failover and load balancing capabilities, which is also one of the best practices for IT environments. When implementing an iSCSI multipath network, **network switches should be configured on separate subnets**.

The following diagram illustrates a full HA configuration that provides contingencies for path failure, switch failure, and storage server failure.

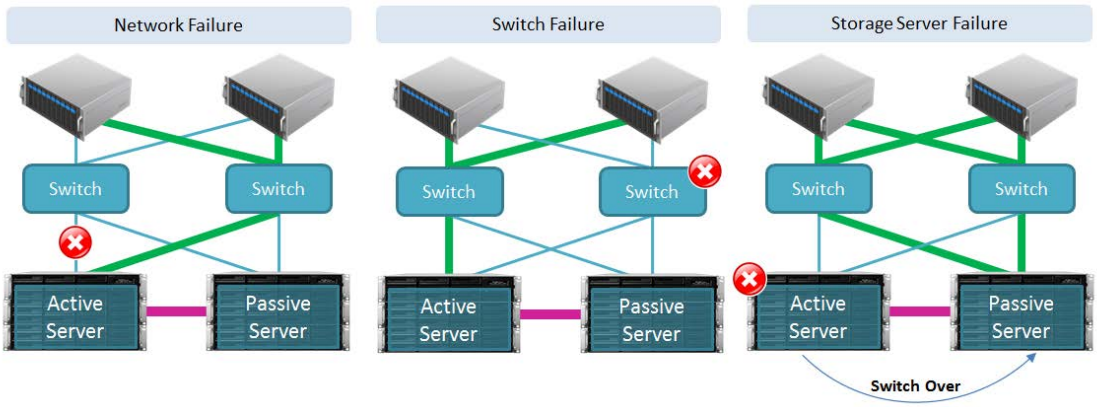


Figure 4: High-availability cluster network configuration on models with four or more LAN ports (iSCSI)

Implementation for NAS storage

The link aggregation feature on Synology NAS can be leveraged to create a resilient HA network for file transfer services such as CIFS, NFS, AFP, and FTP. Link aggregation is a method of using two Ethernet ports in parallel to provide trunking and network fault tolerance. Link aggregation with trunking enhances the connection speed beyond the limits that can be achieved with a single cable or port. Redundancy provides higher link availability and prevents possible disruptions.

Please note that when creating link aggregation on two or more switches, stacked-switches are required for this configuration.

The following diagram demonstrates how link aggregation provides contingencies for path failover and failover that is triggered when a server fails.

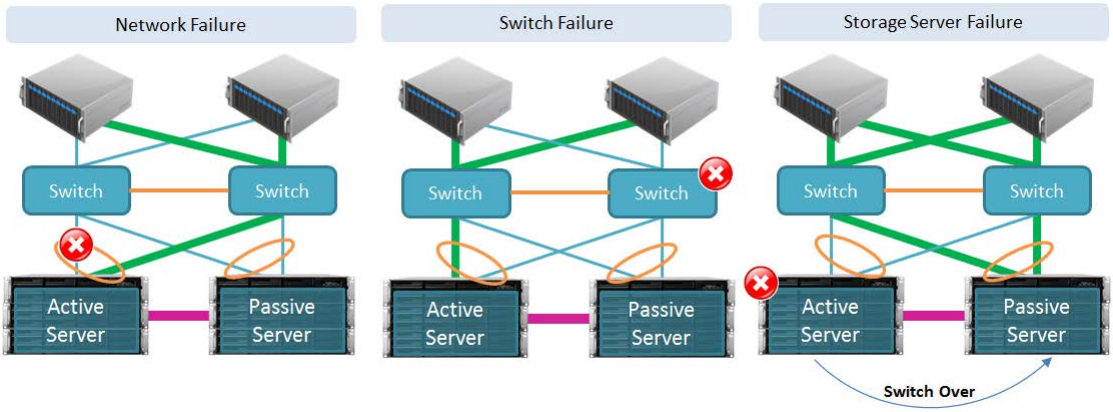


Figure 5: High-availability cluster network configuration on models with four or more LAN ports (NAS)

Path redundancy for the heartbeat connection

For Synology NAS models with four or more network ports, link aggregation may be implemented on the heartbeat connection to provide failover redundancy and load balancing. This feature does not require a switch between the connections.

Network troubleshooting

- The maximum transmission unit (MTU) and virtual LAN (VLAN) ID between DiskStation/RackStation and switcher/router must be identical. For example, if the MTU of DS/RS is 9000, do make sure the corresponding switch/router is able to measure up to that size.
- The switch/router should be able to perform multicast routing for data network. Whereas the switch/router should also be able to perform fragmentation and jumbo frame (MTU value: 9000) if the heartbeat connection goes through the switch/router.
- Ensure the firewall setting does not block the port for DSM and SHA from connection.
- Ensure that the IP of both passive server and HA are in the same subnet.
- It is suggested that both Ethernet ports connected to the same network be set in different subnets, instead of the same one. That is, avoid having these two interfaces share the same subnet; for example, setting one interface with an IP 192.168.x.x/32, whereas the other with 192.168.x.x/32.
- Unstable internet (reduced ping rate or slow internet speeds) after binding:

    1) Try connecting to another switch/router or connect an independent switch/router to DS/RS and PC/Client for testing.

    2) If the heartbeat connection goes through a switch/router, try replacing it with another switch/router or connecting two DiskStation/RackStation to each other.

- It is suggested that both Ethernet ports connected to the same network be set in different subnets, instead of the same one. That is, avoid having these two interfaces share the same subnet; for example, setting one interface with an IP 192.168.x.x/32, whereas the other with 192.168.x.x/32.
- The setting of flow control on switch/router would also induce packet loss in the network. Please check if the setting is in accordance with DS/RS, which normally will auto-detect and conform to the setting of the corresponding switch/router. User are advised to manually enable/disable the flow control if inconsistency is observed.
- Ensure that **Bypass proxy server for local addresses** in the proxy setting is enabled.

Chapter

4

Ensuring Service Continuity

4.1 Switchover Mechanism

To ensure continuous availability, service provisioning can be switched from the active server to the passive server in a normally functioning high-availability cluster at any time. Switchover can be manually triggered for system maintenance, or automatically initiated in the event of the active server malfunctioning, which is known as “failover.” After the servers exchange roles, the original active server assumes the role of the passive server and enters standby mode. As resources within the cluster are accessed using a single virtual interface, switchover does not affect the means of access.

- **Switchover:** The active and passive server can be manually triggered to exchange roles without interruption to service for occasions such as system maintenance.
- **Failover:** In the event of critical malfunction, the cluster will automatically initiate switchover to maintain service availability.

The following situations can trigger system failover:

- **Crashed storage space:** If a storage space (e.g., volume, Disk Group, RAID Group, SSD Cache, etc.) on the active server has crashed, while the corresponding storage space on the passive server is functioning normally, failover will be triggered unless there are no volumes or iSCSI LUNs (block-level) on the crashed storage space. Storage spaces are monitored every 10 seconds. Therefore, in the worst case, switchover will be triggered in 10 to 15 seconds after a crash occurs.

Note: Switchover is not possible when the storage space on the passive server is busy with a Storage Manager related process (e.g., creating or deleting a volume).

- **Service Error:** If an error occurs on a monitored service, failover will be triggered. Services that can be monitored include CIFS, NFS, AFP, FTP, and iSCSI. Services are monitored every 30 seconds. Therefore, in the worst case, switchover will be triggered 30 seconds after an error occurs.
- **Power Interruption:** If the active server is shut down or rebooted, both power units on the active server fail, or power is lost, failover will be triggered. Power status is monitored every 15 seconds. Therefore, in the worst case, switchover will be triggered 15 seconds after power interruption occurs. However, depending on the client’s protocol behavior (e.g., SAMBA), the client may not be aware of the fact that data was still in the active server's cache during power interruption. If this is the case, the data that has not been flushed into the storage might not be re-sent by the client after the power interruption, resulting in a partial data loss.
- **Data Connection Lost:** If an error occurs on the data connection, and the passive server has more healthy data connections, failover will be triggered. For example, if the active server has three data connections and two of them are down, the active server will check whether the passive server has two or more available connections. If it does, failover will be triggered in 10 to 15 seconds. Please note that for connections joined with link aggregation, each joined connection group is considered one connection.

After switchover has occurred, the faulty server may need to be replaced or repaired. If the unit is repaired, restarting the unit will bring the cluster back online and data-synchronization will automatically take place. If the unit is replaced, the cluster will need to be re-bound in order to recreate a functioning cluster. Any USB/eSATA devices attached to the active server will have to be manually attached onto the passive server once switchover is complete.

Note: When a switchover occurs, all existing sessions are terminated. A graceful shutdown of the sessions is not possible, and some data loss may occur; however, retransmission attempts should be handled at a higher level to avoid loss. Please note that if the file system created on an iSCSI LUN by your application cannot handle unexpected session terminations, the application might not be able to mount the iSCSI LUN after a failover occurs.

4.2 Switchover Time-to-Completion

When switchover is triggered, the active server becomes the passive server, at which time the original passive server will take over. During the exchange, there will be a brief period where both servers are passive and services are paused.

The time-to-completion varies depending on a number of factors:

- The number and size of volumes or iSCSI LUNs (block-level)
- The number and size of files on volumes
- The allocated percentage of volumes
- The number of running packages
- The number and total loading of services on the cluster

The following table provides estimated time-to-completion:

Settings	Switchover	Failover by data connection lost	Failover by power interruption
10 * 1T volume 10 * shared folder	36	37	28
10 * 1T volume 10 * 1T advanced LUN	38	37	28
10 * 1T block LUN	31	34	25
60 * 1T volume 60 * shared folders	103	101	33
60 * 1T volume 60 * 1T advanced LUN	105	101	36
60 * 1T block LUN	91	88	28

Unit: second

Tested on RS18016xs+ 6.0.1-7393, 10 RAID 5 groups in total. Each RAID includes 8 HDDs. RAID volumes are all in Btrfs file system. There is no user data inside. The following services are enabled: SMB, AFP, NFS, FTP, iSCSI.

The following table shows the estimated time required for SHA to continue iSCSI service on a higher-end model specially configured for mission-critical environments.

Settings	Switchover
File system: EXT4 2 * 15TB advanced LUN, 95% in use	30
File system: Btrfs 2 * 15TB advanced LUN, 95% in used	23
2 * 15TB block LUN	15

Unit: second

Tested on RS18016xs+ 6.1-15047, with 1 RAID 5 volume including 12 HDDs and 2 LUNs. The switchover time is estimated from iSCSI service stop to iSCSI service start.

4.3 Switchover Limitations

Switchover cannot be initiated in the following situations:

- **Incomplete Data Replication:** When servers are initially combined to form a cluster, a period of time is required to replicate existing data from the active to passive server. Prior to the completion of this process, switchover may fail.
- **Passive Server Storage Space Crash:** Switchover may fail if a storage space (e.g., volume, Disk Group, RAID Group, SSD Cache, etc.) on the passive server is crashed.
- **Power Interruption:** Switchover may fail if the passive server is shut down or rebooted, if both power units on the passive server malfunction, or if power is lost for any other reason.
- **DSM Update:** When installing DSM updates, all services will be stopped and then come online after DSM update installation is completed.

Chapter 5

Chapter 5: Deployment Requirements & Best Practices

Two identical Synology NAS servers that support the Synology High Availability (SHA) are required for deployment. Before the two servers are combined to form a high-availability cluster, the Synology High Availability (SHA) Wizard will check for the following hardware and software limitations to ensure compatibility.

5.1 System Requirements and Limitations

- **Synology Servers:** Both active and passive servers must install identical DSM versions and be identical models/models specifically claimed by Synology to support Synology High Availability (SHA). In addition, the memory sizes of both servers are required to be identical.

5.2 Volume and Hard Disk Requirements and Limitations

- **Storage Volume:** In order to accommodate data replication, the storage capacity of the passive server must be equal to or larger than the capacity of the active server. It is strongly advised that the storage of capacity of both servers be identical to reduce chances of inconsistencies. It is also strongly recommended that the hard drives (type, sector size, etc.) of both servers be identical.
- **Quantity of Disks:** Both active and passive servers must have the same quantity of disks. In addition, disk numbering and position must correspond.
- **Synology Hybrid Raid (SHR):** SHR format volumes are not supported.

5.3 Network Environment Requirements and Limitations

- **Network Settings:** Both servers must have static IP addresses belonging to the same subnet.
- **LAN Ports:** Both servers must have the same number of LAN ports, including the same number of additional network card interfaces.

Note: In a SHA environment, connecting via DHCP, IPv6, and PPPoE is not supported. Also, wireless and DHCP server service should be disabled. Please ensure that these situations are avoided before attempting to form a high-availability cluster.

5.4 Storage Manager Limitations

Once a high-availability cluster has been formed, Storage Manager will no longer be able to change RAID types. However, the following actions will remain available after the formation of the high-availability cluster:

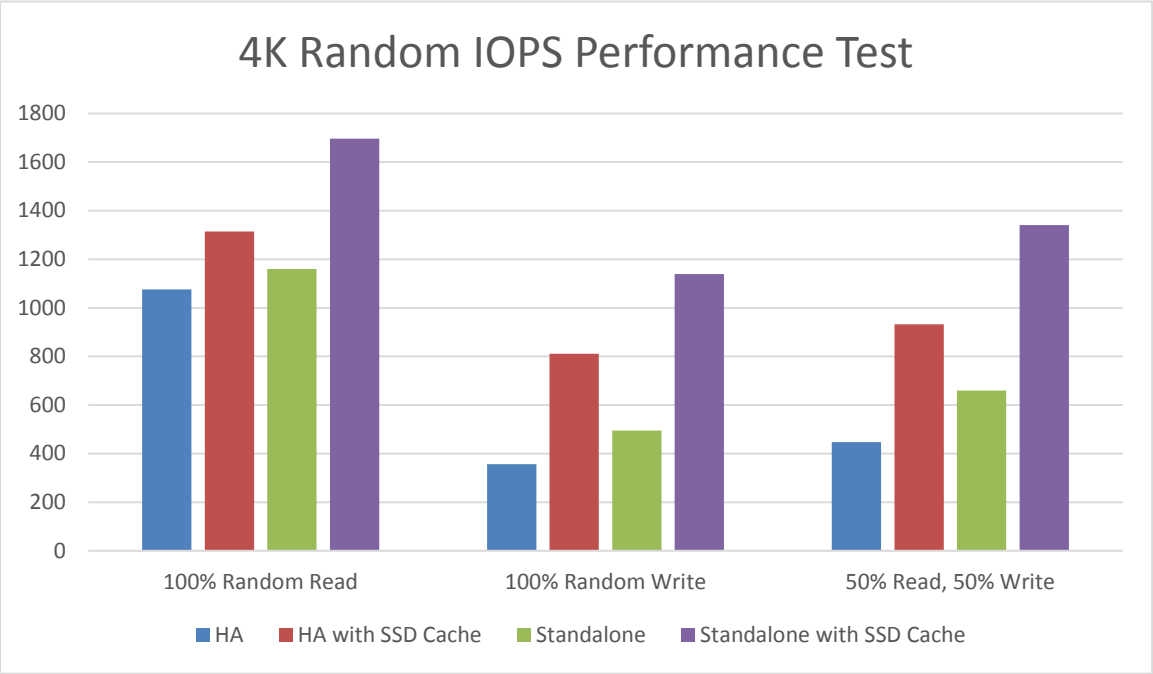
- Expand RAID Groups by adding or replacing hard disks (only for RAID Groups for multiple volumes or iSCSI LUNs).
- Expand volume or iSCSI LUN (block-level) size.
- Create, delete, or repair volumes, iSCSI LUNs, and SSD Caches.
- Change iSCSI LUN (file-level) size and location.
- Change iSCSI LUN target.

5.5 Expansion Units Requirements

Expansion units can be added to existing high-availability cluster configurations in order to increase storage capacity. As with other hardware requirements, identical expansion units are required for both the active and passive servers.

5.6 Performance Considerations

Synology High Availability employs the synchronous commit approach by acknowledging write operations after data has been written on both the active and passive servers at the cost of performance. To enhance the random IOPS performance and reduce latency, it's recommended to enable the SSD cache for volumes that require high performance for random IO workloads.



	HA	HA with SSD Cache	Standalone	Standalone with SSD Cache
100% Random Read	1075.58	1313.84	1160.5	1696.2
100% Random Write	355.97	811.58	495.37	1138.5
50% Read, 50% Write	447.2	932.24	659.1	1341.21

Test Environment: RS2414+; Hit Rate: 50%; Read-Write Cache Enabled.

Summary

Synology's High Availability solution provides a cost-effective and reliable means of insuring against service downtime. This white paper has outlined the basic principles and benefits of Synology High Availability (SHA). For more information and customized consultation on deployment, please contact Synology at **www.synology.com**.