



**SGI NAS
HA Cluster User Guide
Release 3.1.x**

Copyright © 2013 SGI. All rights reserved; provided portions may be copyright in third parties, as indicated elsewhere herein. No permission is granted to copy, distribute, or create derivative works from the contents of this electronic documentation in any manner, in whole or in part, without the prior written permission of SGI.

SGI reserves the right to make changes to this document at any time without notice and assumes no responsibility for its use.

Refer to the latest product announcement or contact your local SGI representative for information on feature and product availability.

This document includes the latest information available at the time of publication.

TRADEMARKS AND ATTRIBUTIONS

SGI, Silicon Graphics, Supportfolio and the SGI logo are trademarks or registered trademarks of Silicon Graphics International Corp. or its subsidiaries in the United States and other countries. Solaris and OpenSolaris are trademarks or registered trademarks of Sun Microsystems, Inc. in the U.S. and other countries. Microsoft and Windows are registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries. All other trademarks mentioned herein are the property of their respective owners.

Document Number: 007-5899-002

Table of Contents

1 Overview.....	1
1.1 Product Features.....	1
1.2 Basic Concepts.....	1
2 Installation & Setup.....	3
2.1 Pre-requisites.....	3
2.2 Sample Network Architecture.....	3
3 Administration using NMC.....	5
3.1 Configuring the Cluster and Heartbeat Interfaces.....	5
3.2 Configuring the Cluster's Shared Volumes.....	5
4 Administration using NMV.....	9
4.1 Configuring the Cluster and Heartbeat Interfaces.....	9
4.2 Configuring the Cluster's Shared Volumes.....	12
5 Heartbeat and Network Interfaces.....	15
5.1 Quorum Disk.....	16
5.2 Serial Link.....	16
5.3 IPMP.....	17
6 Ensuring Exclusive Access to Storage.....	21
6.1 SCSI-3 PGR.....	21
7 Storage Failover.....	23
7.1 Cluster Configuration Data.....	24
7.2 Mapping Information.....	25
7.3 NFS/CIFS Failover.....	25
7.4 Configuring iSCSI targets for Failover.....	25
7.5 Configuring Fibre channel targets for Failover	30
8 System Operations.....	33
8.1 Check status of cluster.....	33
8.2 Checking Cluster Failover Mode.....	34
8.3 Failure Events.....	34

8.4 Service repair.....	35
8.5 Replacing a faulted node.....	36
8.6 Maintenance.....	37
8.7 System Upgrades.....	37
8.7.1 Upgrade procedure.....	37
9 Service Failover.....	39
10 Advanced Setup.....	41
10.1 Name Resolution.....	41
10.2 Cache devices.....	42
11 Testing and Troubleshooting.....	45
11.1 Verify DNS entries.....	45
11.2 Verify moving a resource between nodes of a cluster.....	46
11.3 Verify failing service back to original node.....	47
11.4 Gathering Support Logs.....	47
12 Contact information.....	49
12.1 Support request	49
12.2 Other resources.....	50

1 Overview

1.1 Product Features

HA Cluster provides a storage volume sharing service. One or more shared volumes are made highly available by detecting system failures and transferring ownership of shared volumes to the cluster pair. An HA Cluster consists of two SGI NAS appliances. Neither system is specifically designated to be the 'primary' or 'secondary system'. Both systems can be actively managing shared storage, although any given volume is owned by only one system at any given time.

HA Cluster is based on the RSF-1 (Resilient Server Facility), an industry-leading high-availability and cluster middleware application that ensures critical applications and services are kept running in the event of system failures.

1.2 Basic Concepts

An HA cluster consists of SGI NAS appliances running a defined set of services and monitoring each other for failures. These SGI NAS appliances are interconnected by means of various communication channels, through which they exchange heartbeats that provide information about their states and the services running on them.

RSF-1 cluster service - a transferable unit consisting of application start-up and shutdown code, its network identity and its data. Services can be migrated between cluster appliances either manually or automatically upon failure of one appliance.

An HA Cluster is a group of SGI NAS appliances and therefore provides a superset of the corresponding SGI NAS "basic group" functionality. In particular, you could still use the generic 'switch' command, to switch the management console to operate in a group mode - that is, execute NMC commands on all appliances in the group (in this case - in the cluster). See NMC command 'switch group' for details. To view the existing groups of appliances, run 'show group' (or view the existing configured groups of appliances via NMV).

HA Cluster provides server monitoring and failover. Protection of services such as iSCSI

involves cooperation with other modules such as the SCSI Target Plugin.

2 Installation & Setup

The installer should review the latest release notes and user guide on the support pages at www.sgi.com before beginning the customer installation.

2.1 Pre-requisites

High availability capability for SGI NAS is provided by the HA Cluster Plugin. You will need this software installed on each SGI NAS clustered appliance. SCSI and iSCSI failover services use the SCSI Target Plugin, which is included with the SGI NAS software.

HA Cluster assumes that there is shared storage between the SGI NAS clustered systems. Additional items that need to get set up are:

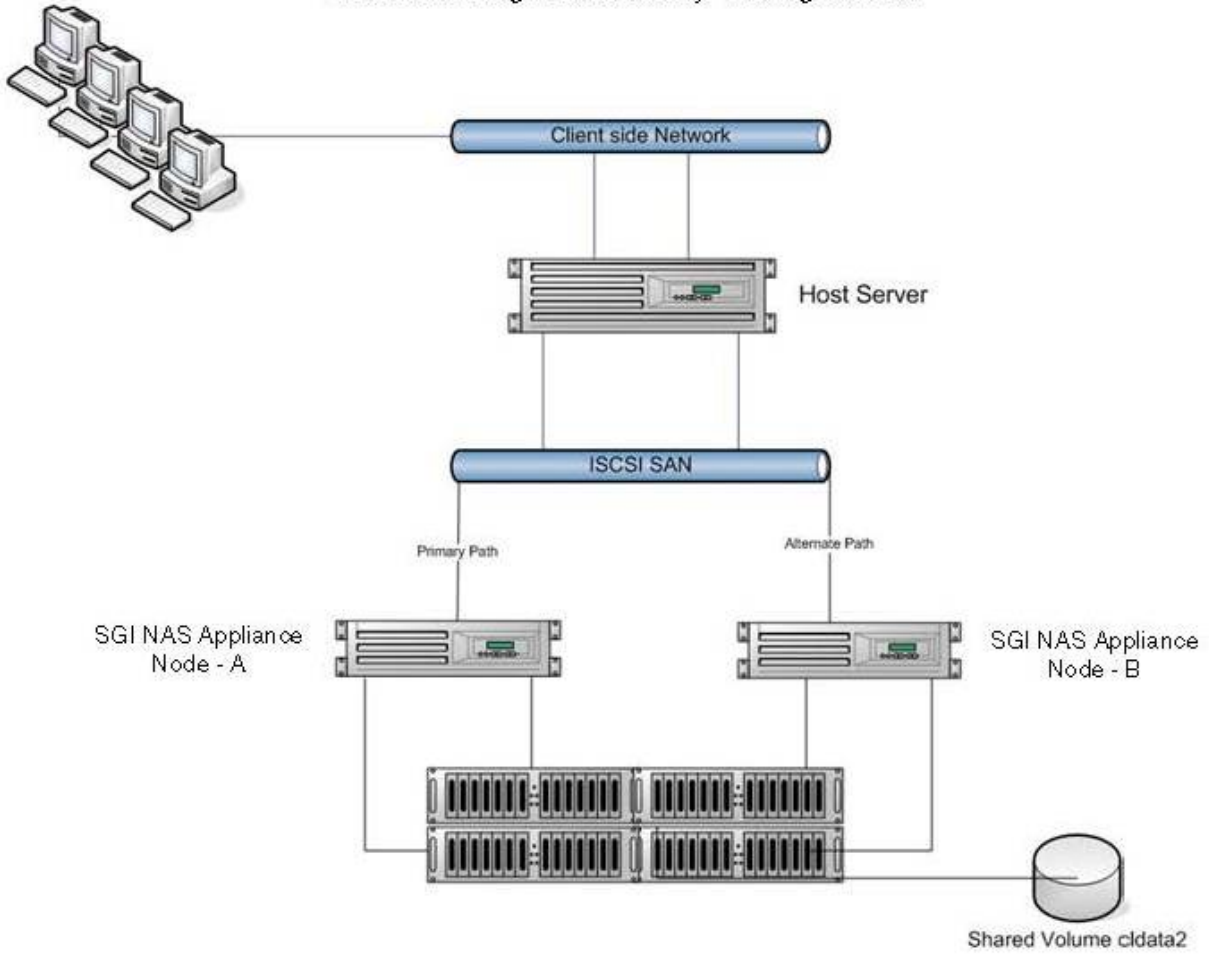
- One IP address for each cluster service unit (Zvol, NAS folder or iSCSI LUN)
- Multiple NICs on different subnets for cluster heartbeat and NMV management
- DNS entry for each service name in the cluster

SGI NAS supports the use of a separate device as a transaction log for committed writes. HA Cluster assumes that this ZFS Intent Log (ZIL) is part of the same storage system as the shared volume.

2.2 Sample Network Architecture

The cluster hardware setup could be two x86/64 boxes with a SAS-connected JBOD and two network interface cards. The picture below is an example of an HA Cluster deployment of a SGI NAS iSCSI environment. The host server has iSCSI LUNS coming from the SGI NAS appliances “nodeA” and “nodeB”. The SGI NAS appliances are using the Active / Active function of the HA cluster and nodeA is servicing one group of iSCSI luns and nodeB is presenting a NAS storage LUN.

SGI NAS High Availability Configuration



3 Administration using NMC

The cluster can be configured and managed via both SGI NAS Management Console (NMC) and the appliance's web GUI (NMV). This section describes command line interfaces for configuring the cluster.

For instructions on using the GUI to configure the cluster, see “Administration Using NMV”.

3.1 Configuring the Cluster and Heartbeat Interfaces

To define the cluster use the create group command. Note that an SGI NAS appliance can not belong to more than one cluster.

```
nmc:/$ create group rsf-cluster
Group name                : cluster-example
Appliances                : nodeA, nodeB
Description               : some description
Scanning for disks accessible from all appliances ...
Heartbeat disk            : c2t4d0
Enable inter-appliance heartbeat via dedicated heartbeats disk? No
Enable inter-appliance heartbeat via primary interfaces? Yes
Enable inter-appliance heartbeat via serial ports? No
Custom properties        :
Bringing up the cluster nodes, please wait ...
Jun 20 12:18:39 nodeA RSF-1[23402]: [ID 702911 local0.alert] RSF-1 cold restart: All
services stopped.
RSF-1 cluster 'cluster-example' created. Initializing ..... done.
```

For more information on configuring network interfaces, see the following section [5 Heartbeat and Network Interfaces](#).

3.2 Configuring the Cluster's Shared Volumes

After setting up the cluster, you need to create one or more shared volumes. The shared logical hostname is a name associated with the failover IP interface that will be moved to the alternate node as part of failover.

```
nmc:/$ setup group rsf-cluster cluster-example shared-volume add
Scanning for volumes accessible from all appliances ...
Shared volume           : cldata2
VIP1 Shared logical hostname : rsf-data
VIP1 Network interface at   : nodeA
e1000g1
VIP1 Network interface at   : nodeB
e1000g1
VIP1 Failover Netmask      :
The IP configuration to be used with 'cldata2' is: 172.16.3.22. Please confirm that
this configuration is correct ?  Yes
Stop adding VIPs?  Yes
Main node                 : nodeA
Initial timeout           : 60
Standard timeout          : 60
Use SCSI reservation (SCSI PGR) for additional protection? (Issue scsi reservati ons
on the devices in a volume before importing it; this is done to enforce data
integrity (i.e. prevents the pool being imported on two nodes at any one time).
Under normal circumstances this option should be left enabled and should only b e
turned off if instructed to do so by UIK support staff.)  No
Adding shared volume 'cldata2', please wait ...
Feb 10 06:12:37 myhost20 RSF-1[15827]: [ID 702911 local0.alert] RSF-1 cold resta rt:
All services stopped.
Waiting for add operation to complete ..... done.

HA CLUSTER STATUS: cluster-example
nodeA:
  cldata2      running      auto      unblocked  rsf-data  e1000g1   60  60
nodeB:
  cldata2      stopped      auto      unblocked  rsf-data  e1000g1   60  60
```

Note, that at this point you may get an error, that the shared logical hostname is not resolvable. For information on how to resolve it, see the section [10.1 Name resolution](#).



Note, that although all the shared volumes are accessible from all appliances, only imported on the current appliance will be shown in the list. In other words, if you want to create a new cluster service with a certain shared volume, this volume need to be imported on the current node.

Next, the cluster interconnect gets validated and the volume get shared:

```
About to verify interconnect between appliances in the group. Caution! To skip this
check, say No. Proceed to verify appliances interconnect?  Yes

Initial timeout           : 60
Standard timeout         : 60

Adding new shared volume, please wait ...

Jun 20 12:40:06 nodeA RSF-1[25914]: [ID 702911 local0.alert] RSF-1 cold restart: All
services stopped.

Waiting for add operation to complete ..... done.
```

Starting from this point on, the volume 'cldata2' is shared. In the event of a system failure the volume will remain accessible to users as long as one of the systems remains up.

Note, that you can add few virtual IPs/hostnames per volume service.

In NMC, run:

```
nmc:/$ setup group rsf-cluster <service name> vips add
```

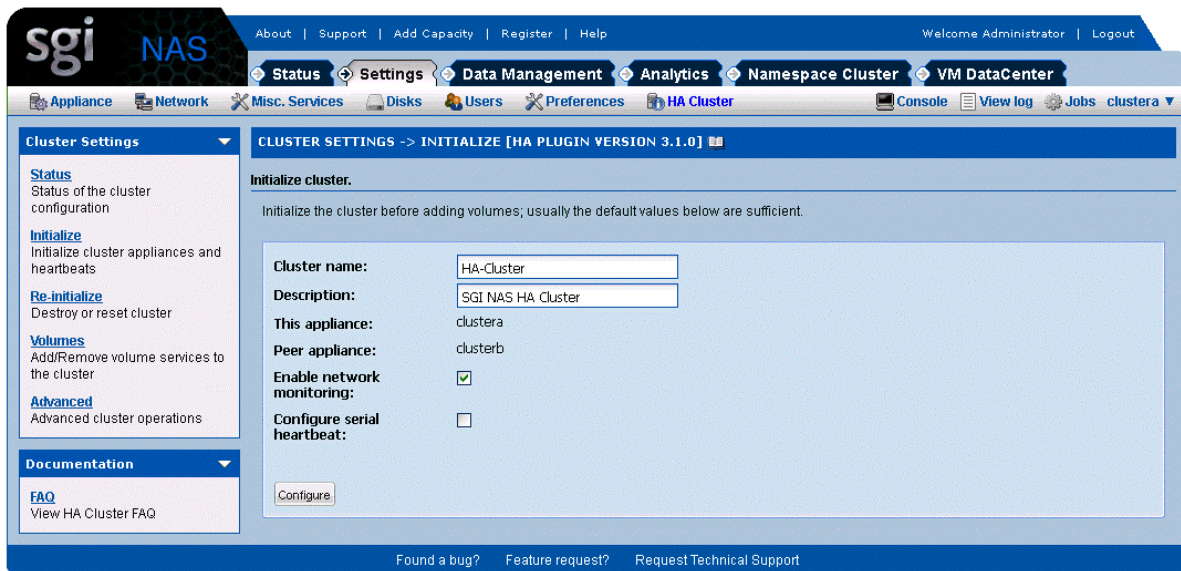
NMV provides the same functionality.

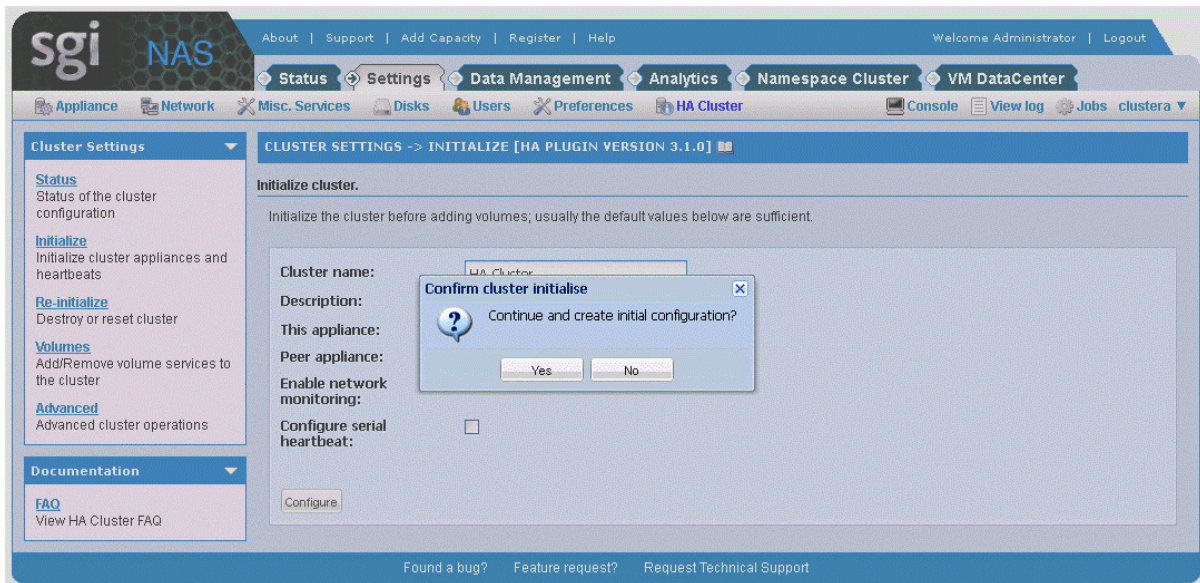
4 Administration using NMV

The cluster configuration steps that were shown for NMC can also be done using our Web GUI.

4.1 Configuring the Cluster and Heartbeat Interfaces

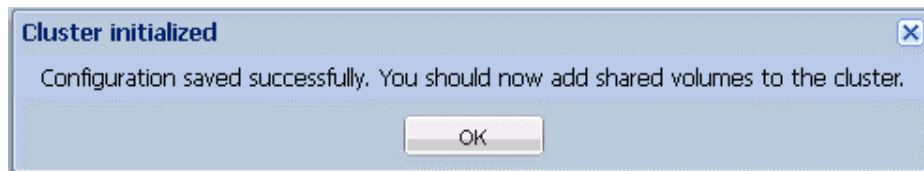
Selecting **Settings**→**HA Cluster**→**Initialize** in NMV is equivalent to using '**create group rsf-cluster**' command in NMC. Remember that an SGI NAS appliance can not belong to more than one cluster.





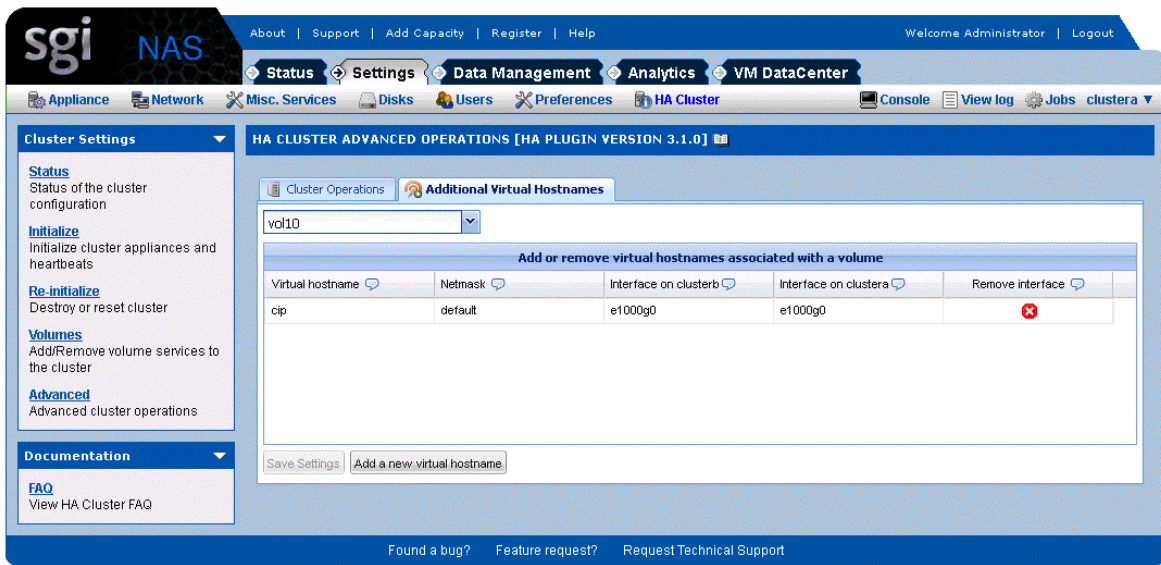
Click 'yes'.

For more information on configuring network interfaces, see section [5. Heartbeat and Network Interfaces](#).



As soon as you get the above message, the cluster is initialized. And you can start adding shared volumes to cluster.

To add additional hostname to volume service click on [Advanced](#) → Additional Virtual Hostnames:



Click on 'Add a new virtual hostname' to see the following window. Fill the required fields:

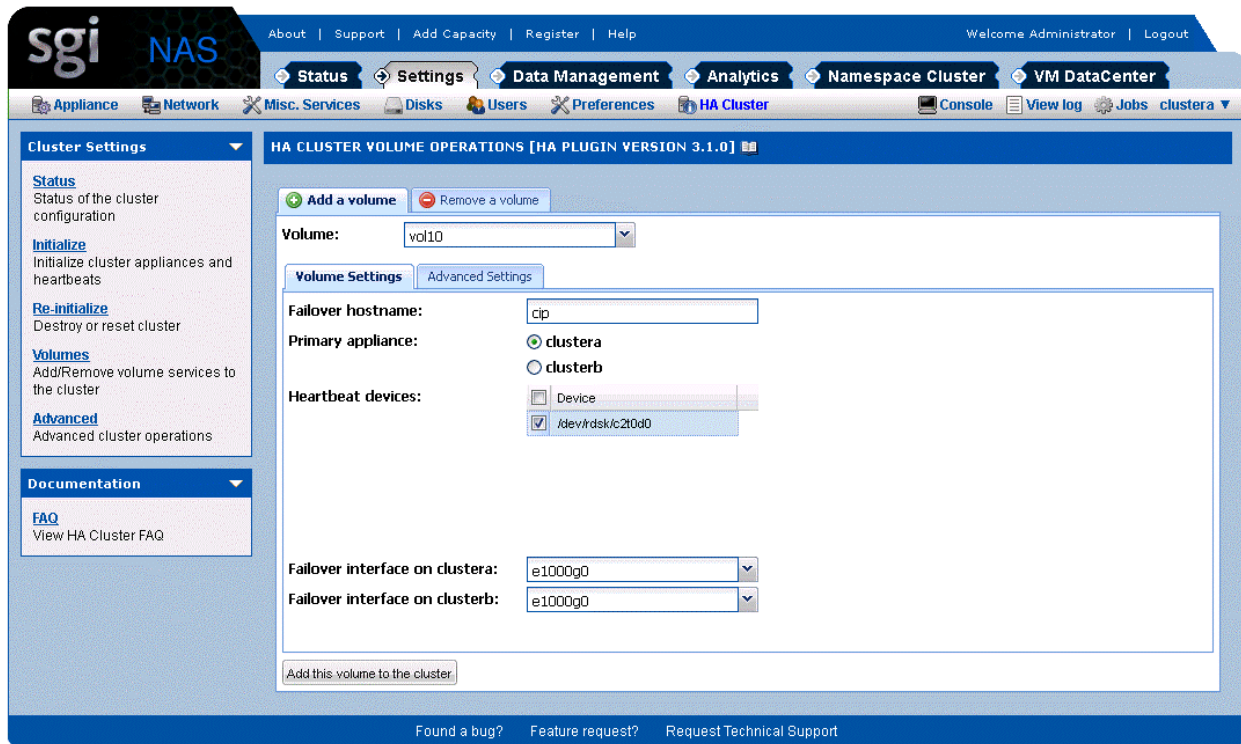
The dialog box is titled 'Add virtual hostname' and contains the following fields:

- Virtual hostname:** Text input field with placeholder text 'Please enter a host name'.
- Netmask:** Text input field with value 'default'.
- Interface on clusterb:** Dropdown menu with value 'e1000g0'.
- Interface on clustera:** Dropdown menu with value 'e1000g0'.

Buttons at the bottom are 'Cancel' and 'Save'.

4.2 Configuring the Cluster's Shared Volumes

After cluster initialization you'll be automatically redirected to adding volume services page:



Note that creation of a new volume on a cluster has a limitation. During the process of creating a new volume service, shared volumes are seen in NMV drop-down list only on the node they are currently imported on. In other words, before creating a new cluster service required volumes need to be imported on the desired appliance. This limitation is planned to be removed in the future releases.

HA Cluster User Guide

The screenshot displays the SGI NAS management console for an HA Cluster. The interface includes a navigation menu with options like Status, Settings, Data Management, Analytics, and VM DataCenter. The main content area is titled 'APPLIANCE SUMMARY FOR HA-CLUSTER [HA PLUGIN VERSION 3.1.0]' and is divided into several sections:

- Cluster Settings:** A sidebar menu with options: Status (Status of the cluster configuration), Initialize (Initialize cluster appliances and heartbeats), Re-initialize (Destroy or reset cluster), Volumes (Add/Remove volume services to the cluster), and Advanced (Advanced cluster operations).
- Documentation:** A sidebar menu with an FAQ link (View HA Cluster FAQ).
- APPLIANCE SUMMARY FOR HA-CLUSTER [HA PLUGIN VERSION 3.1.0]:**
 - Appliance clustera is up:** A table showing volume service 'vol10' with state 'Available', failover mode 'Automatic', and network interface 'e1000g0: cip [192.168.100.115]'. The last state change was on Wed Aug 29 19:47:44.
 - Appliance clusterb is up:** A table showing volume service 'vol10' with state 'Exported', failover mode 'Automatic', and network interface 'e1000g0: cip [192.168.100.115]'. The last state change was on Wed Aug 29 19:46:06.
 - Heartbeats to appliance clustera:** A table showing heartbeat status from clusterb to clustera.
 - Heartbeats to appliance clusterb:** A table showing heartbeat status from clustera to clusterb.

At the bottom of the interface, there are links for 'Found a bug?', 'Feature request?', and 'Request Technical Support'.

5 Heartbeat and Network Interfaces

SGI NAS appliances in the HA Cluster constantly monitor each other states and statuses, via heartbeats. Because HA Cluster servers must be certain that an appliance (member of the cluster) is down before taking over its services, the cluster is configured to use several communication channels through which to exchange heartbeats.

Only the loss of all heartbeat channels represents a failure. If an appliance wrongly detects a failure, it may attempt to start a service that is already running on another server, leading to so-called “split brain” syndrome. This can result in confusion and data corruption. Multiple, redundant heartbeats prevent this from occurring.

HA Cluster supports 3 types of heartbeat communication channels:

- Shared disk accessible and writable from all appliances in the cluster (also sometimes called quorum device) or VDEV labels of the devices in the shared volume.
- Network interfaces, including configured interfaces, unconfigured interfaces, and link aggregates
- Serial link

Later heartbeat properties can be changed by running the following nmc command:

```
nmc:/$ setup group resf-cluster <cluster_group_name> hb_properties
```

If no services are shared between two particular SGI NAS appliances, then no direct heartbeats are required between them. However, at least one heartbeat must be transmitted to each member of a cluster for control and monitoring requests to be propagated. The heartbeat monitoring logic is defined by two parameters: X and Y, where:

- X equals number of heartbeats the interface is observed to be down before action is taken
- Y represents the number of heartbeats an interface must be observed as up before making it available again to the cluster.

The current heartbeat defaults are 3 and 2 heartbeats, respectively.

With SGI NAS, we also provide protection for network interfaces through link aggregation. You can set up aggregated network interfaces using NMC or NMV.

5.1 Heartbeat mechanism

Starting with SGI NAS 3.1 the necessity for dedicated heartbeat device (or quorum disk) is removed. Now the heartbeat is performed through VDEV labels of devices in the shared volume. If a shared volume consists of few disks, VDEV labels of only two disks will be used for heartbeating and the user may specify which disk to be used.

Though the quorum disk option still remains in the configuration file, it is recommended to use the shared volume's labels.

The heartbeat mechanism uses sectors 512 and 518 in the blank 8K space of the VDEV label on each of the shared disks. There is also no speed requirement, 5400rpm drives are more than capable.

5.2 Serial Link

Serial heartbeat packets are exchanged via a dedicated RS232 serial link between any two appliances, using a custom protocol. To prevent routing problems affecting this type of heartbeat, IP is not used on this link.

For the serial link, you will require:

- Spare RS232 serial ports on each HA Cluster server (if you do not have any, you can purchase serial expander cards);
- A crossover, or null modem RS232 cable with an appropriate connector on each end.

Null modem cables are commonly used to connect pieces of Data Terminal Equipment (DTE) together. They are sometimes used to attach console terminals.

On each server, you should enable the relevant serial port device but disable any login, modem or printer services running on it. The serial port must not be used for any other purpose.

To configure serial port heartbeats, you should answer **'Yes'** to this question during HA cluster group creation:

'Enable inter-appliance heartbeat via serial ports?'

5.3 IPMP

Here are some steps to configure IPMP multi-pathing and HA Cluster using two network interfaces and the reserved class C address range.

1. Identify the network interfaces on your machine you want to use for IPMP. In this example we're using two, hme0 and hme1 in the IPMP group rsfnafo.
2. Obtain four fixed IP addresses in the same local LAN segment to be used as fixed IPMP addresses, this example uses the reserved class C range 192.168.20.* with a simple naming convention for clarity; change these names to suit your installation. Update /etc/inet/hosts with the IP addresses obtained:

```
192.168.20.101 DUMMY0
192.168.20.102 DUMMY1
192.168.20.103 REALHOST
192.168.20.104 RSF-DATA
```

The DUMMY0 and DUMMY1 addresses are fixed to hme0 and hme1 for use by IPMP, REALHOST is a floating address used to refer to the node itself, and the RSF-DATA address will be used by RSF-1 to provide an address for clients to access services in the cluster.

3. Next configure the two interfaces using the **/etc/hostname.hme[01]** files:

/etc/hostname.hme0

```
DUMMY0 netmask + broadcast +
group rsfnafo deprecated -failover up
addif REALHOST netmask + broadcast + failover up
```

/etc/hostname.hme1

```
DUMMY1 netmask + broadcast +
```

```
group rsfnafo deprecated -failover standby up
```

4. Next configure unique MAC addresses on all interfaces. To do this, run `ifconfig -a` and note the MAC address on the first interface:

```
hme0: flags=1000843 mtu 1500 index 2
inet 298.178.99.141 netmask fffffff0 broadcast 298.178.99.143
ether 8:0:20:ca:ff:eb
```

Next, enable local mac addresses, plumb in any unconfigured interfaces and then assign them a new unique MAC address. The usual way to do this is by slightly modifying an existing mac address on the system, so in this case we change the hme0 address 8:0:20:ca:ff:eb to 8:0:20:ca:ff:ec, i.e. add one to the final hex number:

```
# eeprom 'local-mac-address?=true'
# /sbin/ifconfig hme1 plumb
# /sbin/ifconfig hme1 ether 8:0:20:ca:ff:ec
```

5. Enable IP failure detection in RSF-1 by adding the following line at the top of the RSF-1 configuration file (`/opt/HAC/RSF-1/etc/config`) in the global section:

```
#####
#### Optional global defaults & definitions come first. #####
#####
CLUSTER_NAME IPMP_example
IPDEVICE_MONITOR 5,5
POLL_TIME 2
REALTIME 1
#####
##### End of global section, start of machines section. #####
#####
```

6. The RSF-DATA should then be declared in the RSF-1 configuration file for a

single service VIP:

SERVICE example RSF-DATA "IPMP Service Example"

```
INITIMEOUT 60
RUNTIMEOUT 60
SERVER REALHOST
IPDEVICE "hem0"
SERVER OTHERHOST
IPDEVICE "hme0"
```

7. Finally, reboot the server and check that the appropriate addresses have been enabled on the interfaces.

6 Ensuring Exclusive Access to Storage

At any point in time a given shared volume is accessed exclusively via the appliance that is currently providing the corresponding volume-sharing service. To ensure this, HA Cluster provides reliable fencing through the utilization of multiple types of heartbeats; the most important of these is the disk heartbeat, in conjunction with any other type. Generally, additional heartbeat mechanisms increase reliability of the cluster's fencing logic; the disk heartbeats however are essential.

HA Cluster also has the ability to reboot the failed appliance in certain cases. One such case would be a failure to export the shared volume from a failed appliance - that is, from the appliance that has failed to provide the (volume-sharing) service. This functionality is analogous to STONITH.

In addition, SGI NAS RSF-1 cluster provides a number of other failsafe mechanisms.

When a (volume sharing) service is to be started, the IP address associated with that service should NOT be attached to any interface. The cluster automatically detects and reports the case when this is not so - that is, when the IP address is in use. In this latter case, the local service start-up is not performed.

On disc systems which support it, a SCSI reservation can be placed on a disc before accessing the file systems, and the system is set to panic should that reservation be lost. This also serves to protect the data on a disc system. SCSI-2 reservations are supported.

6.1 SCSI-3 PGR

HA Cluster employs SCSI-2 PGR. SCSI-3 PGR is not supported for HA Cluster. It won't work with SATA drives, and has certain other limitations. It is our recommendation to always deploy cluster with a shared disk (quorum device) and at least one more heartbeat type of a channel (Ethernet or Serial). If this is done, the cluster logic itself will ensure exclusive access, independently of the storage interconnects used in the cluster.

7 Storage Failover

The primary benefit of HA Cluster is to detect storage system failures and transfer ownership of shared volumes to the alternate SGI NAS appliance. HA Cluster ensures service continuity in the presence of service level exceptional events, including power outage, disk failures, appliance running out of memory or crashing, etc.

This section discusses some of the details associated with failover.

Currently the minimum time to detect that an appliance has failed is 10 seconds. The failover and recovery time is then largely dependent on the amount of time it takes to re-import the data volume on the alternate appliance. Best practices to reduce the failover time include using fewer zvols and file systems per data volume. When using fewer file systems you may then want to use other properties such as reservations and quotas to control resource contention between multiple applications.

In the default configuration, HA Cluster will also failover storage services if network connectivity is lost. HA Cluster automatically works out which network device to monitor based on the services bound to an interface so no further configuration is required. Checking is done on all nodes in the cluster, so even if a node is not running any services, HA Cluster will continue to monitor the unused interfaces, and, should one go offline, prevent fail over to this node for services bound to that interface (as there is little point in failing over to a machine with an unusable interface for a service). Should the interface subsequently recover then HA Cluster will re-enable fail over for that interface.

Other types of failure protection include link aggregation for network interfaces and MPxIO for protection against SAS link failures.

In addition to failover of the shared storage, HA Cluster will also failover the storage services. However not all configuration settings are moved. For example, any local users that have been configured for the SGI NAS appliance will not be moved. It is highly recommended that a directory service such as LDAP is used in this particular case.

7.1 Cluster Configuration Data

When configuring SCSI targets in a cluster environment, you want to make sure that configurations and mappings are consistent across the cluster members. All SCSI Target operations are automatically propagated in the general case. However, there can be issues if the alternate node is not available at the time of the configuration change. By default, the operation will result in a warning to the user that the remote update failed. You can also set HA Cluster to synchronous mode. In this case the action will fail completely, if the remote update fails.

To protect local configuration information that is not migrated, you can periodically save this configuration to a remote site (perhaps the alternate node) and then use NMC commands to restore it in the event of a failover. Use the command '**setup application configuration**' with the save and restore subcommands.

Here are some examples of using NMC commands to re-synchronize the cluster configuration, if one of the nodes is not up-to-date. In these two examples, nodeA has the latest configuration and nodeB needs to be updated.

One option is to run this command from nodeA:

```
nmc:/$ setup iscsi config restore-to nodeB
```

This would save the configuration of nodeA, copy it over to nodeB, and restore it on nodeB.

Another option is to run this command from nodeB:

```
nmc:/$ setup iscsi config restore-from nodeA
```

Note: Restore operations are destructive and should be performed only during a planned downtime window.

Key configuration data that is saved includes target groups and host groups (stmf.config) and targets, initiators, and target portal groups (iscsi.conf).

If CHAP authentication is being used, and the CHAP configuration was done through NMC or NMV, then it can be safely saved and restored.

7.2 Mapping Information

SCSI Target is used to map Zvols from the cluster nodes to client systems. It is critical that the cluster nodes have the same mapping information. Mapping information is specific to the volume and is stored with the volume itself.

Manual maintenance tasks can be performed on the mapping information using the 'mapmgr' command.

7.3 NFS/CIFS Failover

HA Cluster can be used to ensure the availability of NFS shares to users. However, it should be noted that HA Cluster does not detect the failure of the NFS server software.

7.4 Configuring iSCSI targets for Failover

HA Cluster can be used to failover iSCSI volumes from one cluster node to another. The target IQN is moved as part of the failover.

Setting up iSCSI failover involves first setting up Zvol in the shared volume. Setting up a Zvol is shown in the following screenshot. The main point to note is that the process of creating Zvol and sharing it via iSCSI is done separately from HA Cluster configuration.

If iSCSI Zvols are created before the Zvol's volume is marked as a shared cluster volume, then at the time the cluster volume is shared active iSCSI sessions may experience some delays. Depending on the network and application environment and active workload, it is possible that you may also see command level failures or disconnects during this period.

When adding a shared volume to a cluster which will have zvols created as backing storage for iSCSI targets, it is vital all client iSCSI initiators, irrespective of operating system, are configured to access those targets using the shared logical hostname specified when the volume service was created, rather than a real hostname associated with one of the appliances.

It is important to note, that the cluster handles all aspects of the shared logical hostname

configuration; therefore, you should never attempt to configure the shared logical hostname manually. Furthermore, unless the shared volume service is running, the shared logical hostname should not be present on the network, this can be verified with the ICMP ping command.

To configure iSCSI targets on the appliance where the volume service is currently running go to **Data Management** → **SCSI Target** → **Zvols** → Create to create a virtual block device (zvol01) of the required size (200MB) using the shared volume (vol01).

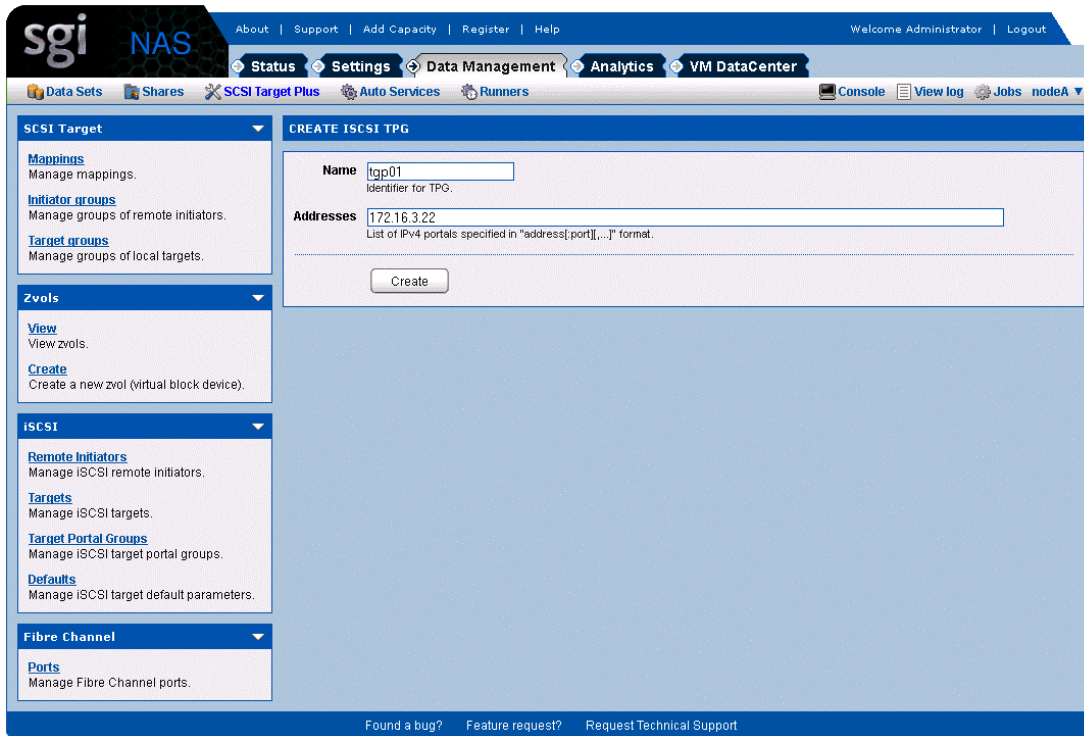
The screenshot displays the SGI NAS web interface. The top navigation bar includes 'About | Support | Add Capacity | Register | Help' and 'Welcome Administrator | Logout'. The main menu shows 'Status', 'Settings', 'Data Management', 'Analytics', and 'VM DataCenter'. The left sidebar contains navigation options for 'SCSI Target', 'Zvols', 'iSCSI', and 'Fibre Channel'. The main content area is titled 'CREATE A NEW ZVOL (VIRTUAL BLOCK DEVICE)' and contains the following configuration fields:

- Volume:** A dropdown menu set to 'vol01'. Below it, the text reads 'Zvol's volume.'
- Name:** A text input field containing 'zvol01'. Below it, the text reads 'Unique path within the ZFS namespace. LUN name can only contain alphanumeric characters as well as underscore ("_"), dash ("-"), period ("."). Maximum length of a dataset name is 256 minus length of selected volume.'
- Description:** An empty text input field. Below it, the text reads 'Human-readable description for this zvol.'
- Size:** A text input field containing '200m'. Below it, the text reads 'Maximum size of the LUN, e.g.: 2TB, 100GB, 500M, 100K. If "sparse" mode is not used, the entire specified size is allocated; otherwise the virtual block device will start small and then may grow up to the specified size. Minimum size of zvol that can be shared - 1M.'
- Initial Reservation:** A dropdown menu set to 'Yes'. Below it, the text reads 'Say "No" to create a "sparse" (that is, thinly provisioned) zvol with no initial reservation. The effective used size is limited by the specified size. Default is "Yes".'
- Block size:** A dropdown menu set to '8K'. Below it, the text reads 'Specifies a suggested block size for the LUN. Default is 8K.'
- Compression:** A dropdown menu set to 'on'. Below it, the text reads 'Controls the compression algorithm used for this dataset. Default is "on".'
- Deduplication:** A dropdown menu set to 'off'. Below it, the text reads 'Controls the deduplication option for the volume. If enabled, it will optimize use of duplicate copies of data. Default is "off".'
- Log Bias:** A dropdown menu set to 'latency'. Below it, the text reads 'Provide a hint to ZFS about handling of synchronous requests in this dataset. If logbias is set to latency (the default), ZFS will use pool log devices (if configured) to handle the requests at low latency. If logbias is set to throughput, ZFS will not use configured pool log devices. ZFS will instead optimize synchronous operations for global pool throughput and efficient use of resources.'
- Number of copies:** A dropdown menu set to '1'. Below it, the text reads 'Controls the number of copies of data stored for this dataset. Default is "1".'
- Sync:** A dropdown menu set to 'standard'. Below it, the text reads 'Controls synchronous requests (standard - ensure all synchronous requests are written to stable storage; always - every file system transaction will be written and flushed to stable storage by system call return; disabled - synchronous requests are disabled). Default is standard.'

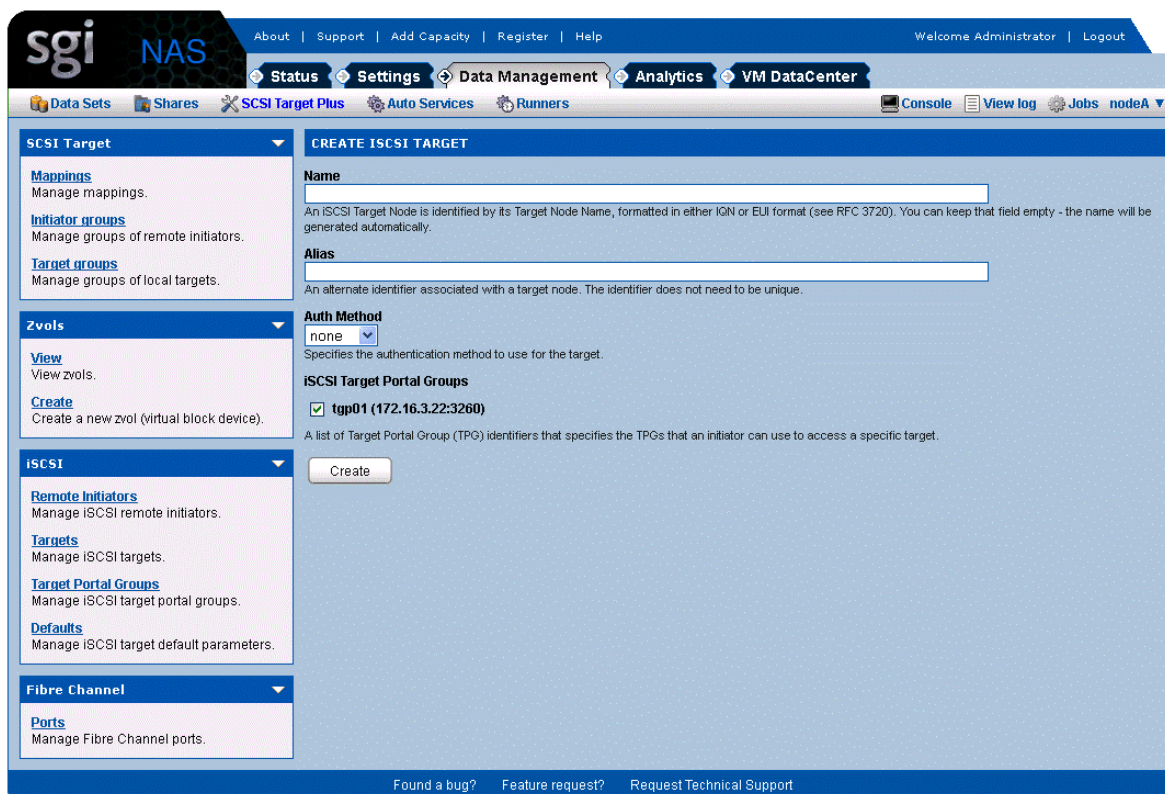
At the bottom of the configuration area is a 'Create' button. The footer of the interface contains the text 'Found a bug? Feature request? Request Technical Support'.

The newly created zvol is automatically migrated to the other appliance on failover, therefore it does not need to be duplicated manually.

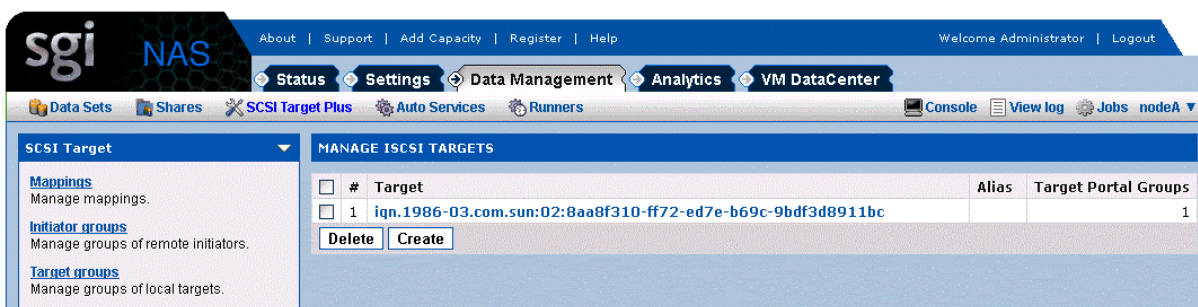
Once the zvol has been created, select Target Portal Groups from the iSCSI pane to define a target portal group (tpg01); it is vital the IPv4 portal address is the shared logical hostname (172.16.3.22) specified when the volume service was created, not a real hostname associated with one of the appliances. The newly created target portal group is automatically replicated to the other appliance, therefore it does not need to be duplicated manually.



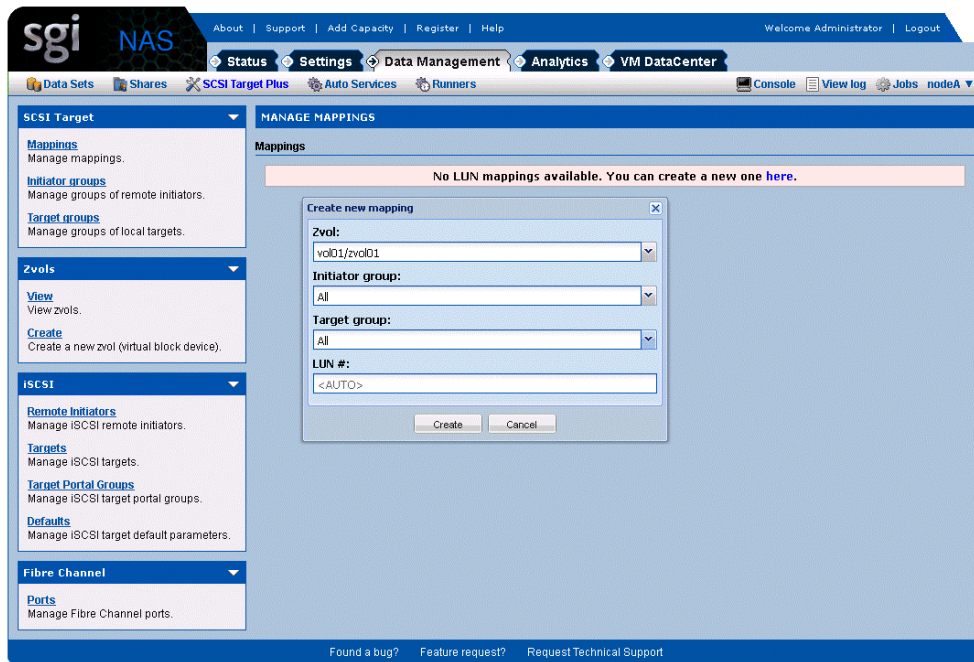
Once the target portal group has been defined, select 'Targets' from the iSCSI pane to create an iSCSI target and add the target portal group defined in the previous step to limit zvol visibility from client initiators to the target portal group. The newly created iSCSI target is automatically replicated to the other appliance, therefore it does not need to be duplicated manually.



The newly created iSCSI target (iqn.1986-03.com.sun:02:8aa8f310-ff72-ed7e-b69c-9bdf3d8911bc) is now displayed in the Targets page:



Once the iSCSI target has been created with the target portal group, select 'Mappings' from the SCSI Target pane to create a LUN mapping to the zvol to be used as backing storage for the iSCSI target. The newly created LUN mapping is automatically migrated to the other appliance on failover, therefore it does not need to be duplicated manually.



Finally, on the client regardless of operating system, it is vital the iSCSI initiator is configured to use both the IQN (iqn.1986-03.com.sun:02:9d5ba857-c064-e538-9e5a-eac99840dd0d) of the iSCSI target created and the shared logical hostname (172.16.3.22) associated with both the volume service and target portal group (tpg01) to access the zvol (zvol01) via iSCSI.

For example, on OpenSolaris:

```


root@nextest2:~# iscsiadm list discovery
Discovery:
  Static: disabled
  Send Targets: disabled
  iSNS: disabled
root@nextest2:~# iscsiadm modify discovery -s enable
root@nextest2:~# iscsiadm add static-config iqn.1986-03.com.sun:02:8aa8f310-ff72-ed7e-b69c-9bdf3d8911bc, 172.16.3.22:3260
root@nextest2:~# devfsadm
root@nextest2:~# format
Searching for disks...done

AVAILABLE DISK SELECTIONS:
 0. c1t0d0 <DEFAULT cyl 30391 alt 2 hd 255 sec 63>
   /pci@0,0/pci1028,177@1f,2/disk@0,0
 1. c2t600144f0ee48850000004c3729fa0002d0 <DEFAULT cyl 198 alt 2 hd 64 sec 32>
   /scsi_vhci/disk@g600144f0ee48850000004c3729fa0002
 2. c3t2d0 <FUJITSU-MAJ3182M SUNI8G-0804-16.87GB>
   /pci@0,0/pci8086,244e@1e/pci1000,1000@2/sd@2,0
 3. c3t3d0 <FUJITSU-MAC3091L SUN9.0G-1111-8.43GB>
   /pci@0,0/pci8086,244e@1e/pci1000,1000@2/sd@3,0
 4. c3t4d0 <SEAGATE-ST39102LCSUN9.0G-0828-8.43GB>
   /pci@0,0/pci8086,244e@1e/pci1000,1000@2/sd@4,0
Specify disk (enter its number): █
    
```

Failover time varies depending on the environment. As an example, initiating failover for a pool containing six zvols, the observed failover time is 32 seconds. Clients hang while the failover is occurring, but otherwise recover quickly.

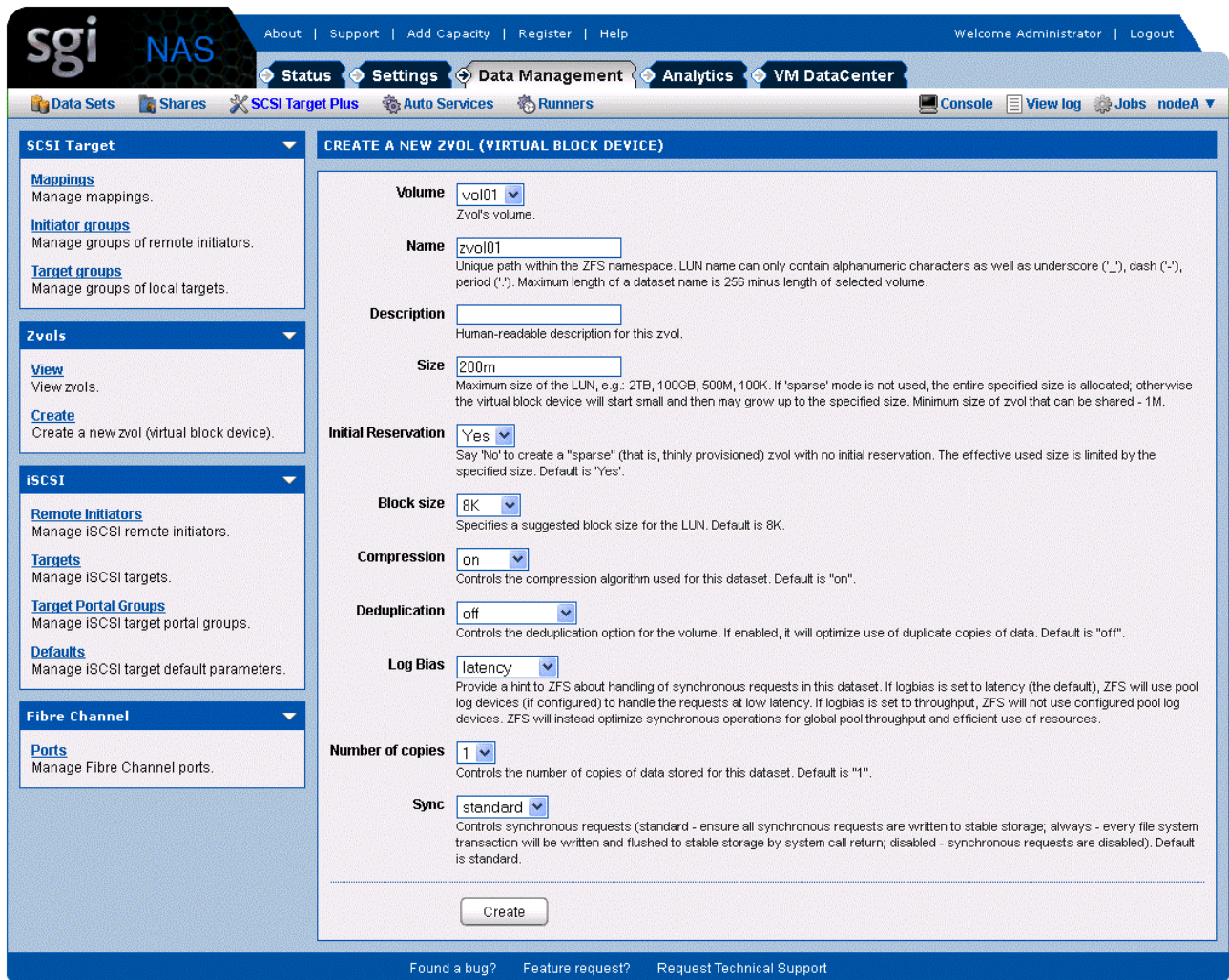
7.5 Configuring Fibre channel targets for Failover

As a prerequisite for configuring Fibre Channel targets, the HBA port modes of both appliances need to be changed from Initiator mode to Target mode. To change HBA port mode, go to **Data Management** → **SCSI Target Plus** → **Fibre Channel** → **Ports** and select **Target** from the drop-down mode menu box.

FIBRE CHANNEL PORTS								
	WWN	State	ID	Mode	Speed	Supported Speeds	Model	Manufacturer
	2100001B321B4F9C	online	11100	Initiator	1Gb	1Gb 2Gb 4Gb	QLE2460	QLogic Corp.

Once the HBA port modes of both appliances have been changed from Initiator mode to Target mode, reboot both appliances for the Target mode changes to come into effect.

To configure Fibre Channel targets on the appliance where the volume service is currently running, go to **Data Management** → **SCSI Target Plus** → **Zvols** → **Create in NMV** to create a virtual block device (in this example `zvol101`) of the required size (`200MB`) using the shared volume (`vol101`). The newly created zvol is automatically migrated to the other appliance on failover; therefore it does not need to be duplicated manually.



Once the Fibre Channel target has been created, select **Mappings** from the **SCSI Target Plus** pane to create a LUN mapping to the zvol to be used as backing storage for the Fibre Channel target. Again, the newly created LUN mapping is automatically migrated to the other appliance on failover; therefore it does not need to be duplicated manually.

The screenshot displays the SGI NAS web management interface. At the top, there is a navigation bar with the SGI logo and 'NAS' text, followed by links for 'About', 'Support', 'Add Capacity', 'Register', and 'Help'. On the right side of the navigation bar, it says 'Welcome Administrator' and 'Logout'. Below the navigation bar, there are several tabs: 'Status', 'Settings', 'Data Management', 'Analytics', and 'VM DataCenter'. Underneath these tabs, there are icons for 'Data Sets', 'Shares', 'SCSI Target Plus', 'Auto Services', and 'Runners'. On the far right of this bar, there are icons for 'Console', 'View log', 'Jobs', and 'nodeA'. The main content area is divided into a left sidebar and a main panel. The sidebar has four sections: 'SCSI Target' (with sub-links for Mappings, Initiator groups, and Target groups), 'Zvols' (with sub-links for View and Create), 'iSCSI' (with sub-links for Remote Initiators, Targets, Target Portal Groups, and Defaults), and 'Fibre Channel' (with a sub-link for Ports). The main panel is titled 'MANAGE MAPPINGS' and shows a message: 'No LUN mappings available. You can create a new one here.' A 'Create new mapping' dialog box is open in the center, containing the following fields: 'Zvol:' with a dropdown menu showing 'vol01/zvol01', 'Initiator group:' with a dropdown menu showing 'All', 'Target group:' with a dropdown menu showing 'All', and 'LUN #:' with a text input field containing '<AUTO>'. At the bottom of the dialog box are 'Create' and 'Cancel' buttons. At the bottom of the main interface, there are links for 'Found a bug?', 'Feature request?', and 'Request Technical Support'.

8 System Operations

There are a variety of commands and GUI screens to help you with day-to-day cluster operations. There is a set of cluster-specific commands to supplement NMC. To see a list of these available NMC commands, type “help keyword cluster”.

Note that although a shared volume is accessible from both cluster nodes, the “show volume” command will show the volume only if it's running on the node currently owning that volume.

8.1 Check status of cluster

Similar to NMC, NMV can also be used to check overall cluster status.

The screenshot shows the SGI NAS HA Cluster GUI. The main content area is titled "APPLIANCE SUMMARY FOR HA-CLUSTER [HA PLUGIN VERSION 3.1.0]". It contains several sections:

- Appliance clusterA is up:** A table showing the status of volume services for node A.

Volume service	Volume state	Failover mode	Action	Network interfaces + virtual hostnames	Last state change	Logs
ha-test	Available	Automatic	Select action...	e1000g1:rsf-data[172.16.3.22]	Wed Aug 29 19:47:44	View
- Appliance nodeB is up:** A table showing the status of volume services for node B.

Volume service	Volume state	Failover mode	Action	Network interfaces + virtual hostnames	Last state change	Logs
ha-test	Exported	Manual	Select action...	e1000g1:rsf-data[172.16.3.22]	Wed Aug 29 19:46:06	View
- Heartbeats to appliance nodeA:** A table showing heartbeat information for node A.

From	Type	Sent to	State	Updated
nodeB	disc	/dev/rdisk/c2t2d0s0:512,/dev/rdisk/c2t2d0s0:518	Up	Thu Aug 30 13:25:35
nodeB	disc	/dev/rdisk/c2t2d0s0:518,/dev/rdisk/c2t2d0s0:512	Up	Thu Aug 30 13:25:35
nodeB	net	nodeA	Up	Thu Aug 30 13:25:35
- Heartbeats to appliance nodeB:** A table showing heartbeat information for node B.

From	Type	Sent to	State	Updated
nodeA	disc	/dev/rdisk/c2t2d0s0:518,/dev/rdisk/c2t2d0s0:512	Up	Thu Aug 30 13:25:35
nodeA	disc	/dev/rdisk/c2t2d0s0:512,/dev/rdisk/c2t2d0s0:518	Up	Thu Aug 30 13:25:35
nodeA	net	nodeB	Up	Thu Aug 30 13:25:35

8.2 Checking Cluster Failover Mode

HA Cluster can be configured to detect failures and alert the user, or to actually failover the shared volumes automatically. To check which mode you have configured, you can use the “show group rsf-cluster” command as shown below.

```
nmc@nodeA:/$ show group rsf-cluster HA-Cluster
PROPERTY                                VALUE
name                                     : HA-Cluster
appliances                               : [nodeB nodeA]
hbipifs                                  : nodeB:nodeA: nodeA:nodeB:
netmon                                   : 1
info                                     : UIK"PCU HA Cluster
generation                               : 1
refresh_timestamp                        : 1297423688.31731
hbdisks                                  : nodeA:c2t1d0 nodeB:c2t0d0
type                                     : rsf-cluster
creation                                 : Feb 11 03:28:08 2011

SHARED VOLUME: ha-test
svc-ha-test-ipdevs                       : rsf-data nodeB:e1000g1 nodeA:e1000g1
svc-ha-test-main-node                     : nodeA
svc-ha-test-shared-vol-name               : ha-test

HA CLUSTER STATUS: HA-Cluster
nodeA:
  ha-test    running    auto    unblocked  rsf-data    e1000g1    60    60
nodeB:
  ha-test    stopped    auto    unblocked  rsf-data    e1000g1    60    60
```

8.3 Failure Events

SGI NAS keeps track of various appliance components, and their state. If and when failover occurs (or any service changes to a "broken" state), an email is sent to the administrator

describing the event.



NOTE: you should have correctly setup SMTP configuration, and previously tested that you are indeed receiving mails from the appliance.

8.4 Service repair

There are two broken states:

- **broken_safe** - A problem occurred while starting the service on the server, but it has been stopped safely and may be run elsewhere.
- **broken_unsafe** - A fatal problem occurred while starting or stopping the service on the server. The service cannot be run on any other server in the cluster until it has been repaired.

To repair the shared volume which is in "broken" state run the NMC command:

```
nmc:/$ setup group rsf-cluster shared-volume repair
```

This will go through the repair process.

In NMV you can mark a service as repaired.

The screenshot displays the SGI NAS HA Cluster management interface. The top navigation bar includes 'About', 'Support', 'Add Capacity', 'Register', and 'Help'. The main menu shows 'Status', 'Settings', 'Data Management', 'Analytics', and 'VM DataCenter'. The left sidebar contains 'Cluster Settings' (Status, Initialize, Re-initialize, Volumes, Advanced) and 'Documentation' (FAQ). The main content area is titled 'APPLIANCE SUMMARY FOR HA-CLUSTER [HA PLUGIN VERSION 3.1.0]' and shows the following sections:

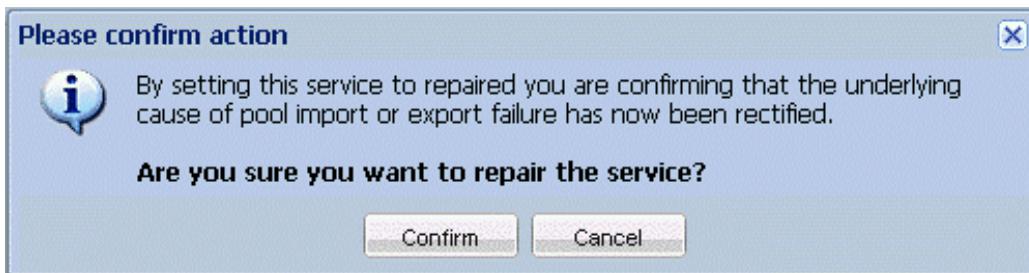
- Appliance cluster is up:** A table with columns: Volume service, Volume state, Failover mode, Action, Network interfaces + virtual hostnames, Last state change, and Logs.

Volume service	Volume state	Failover mode	Action	Network interfaces + virtual hostnames	Last state change	Logs
ha-test	Import Failure	Automatic	Mark repaired	e1000g1:rsf-data[172.16.3.22]	Wed Aug 29 19:47:44	View
ha-test2	Available	Automatic	Select action...	e1000g1:rsf-data[172.16.3.22]	Wed Aug 29 19:47:44	View
- Appliance nodeB is up:** A similar table to the one above, showing the status of nodeB.

Volume service	Volume state	Failover mode	Action	Network interfaces + virtual hostnames	Last state change	Logs
ha-test	Import Failure	Automatic	Select action...	e1000g1:rsf-data[172.16.3.22]	Wed Aug 29 19:47:44	View
ha-test2	Available	Automatic	Select action...	e1000g1:rsf-data[172.16.3.22]	Wed Aug 29 19:47:44	View
- Heartbeats to appliance nodeA:** A table showing heartbeat details for nodeA.

From	Type	Sent to	State	Updated
nodeB	disc	/dev/rdisk/c2t0d0s0:512,/dev/rdisk/c2t0d0s0:518	Up	Thu Aug 30 13:25:35
nodeB	disc	/dev/rdisk/c2t2d0s0:518,/dev/rdisk/c2t2d0s0:512	Up	Thu Aug 30 13:25:35
nodeB	disc	/dev/rdisk/c2t3d0s0:518,/dev/rdisk/c2t3d0s0:512	Up	Thu Aug 30 13:25:35
nodeB	net	nodeA	Up	Thu Aug 30 13:25:35
- Heartbeats to appliance nodeB:** A table showing heartbeat details for nodeB.

From	Type	Sent to	State	Updated
nodeA	disc	/dev/rdisk/c2t1d0s0:518,/dev/rdisk/c2t1d0s0:512	Up	Thu Aug 30 13:25:35
nodeA	disc	/dev/rdisk/c2t2d0s0:512,/dev/rdisk/c2t2d0s0:518	Up	Thu Aug 30 13:25:35
nodeA	disc	/dev/rdisk/c2t3d0s0:512,/dev/rdisk/c2t3d0s0:518	Up	Thu Aug 30 13:25:35
nodeA	net	nodeB	Up	Thu Aug 30 13:25:35



8.5 Replacing a faulted node.

SGI NAS provides advanced capability to restore node in a cluster, in case it accidentally became out of service. There is no need to delete the cluster group on another node and reconfigure the cluster and all the cluster services.

Faulted node is replaced with the following NMC command:


```
nmc:/$ setup group rsf-cluster <group_name> replace_node
```

After executing the command system asks to choose which node to be excluded from the cluster and which will be used instead. System is checking the host parameters of the new node and if they matches the requirements of the cluster group replaces the old one.



Note: Before performing replace node operation, identical configuration must be set up on the new or restored hardware, which will be used to replace the old faulted node. Otherwise, operation fails. Serial port hearbeats configuration should be the same as well.

8.6 Maintenance

To take the cluster offline for maintenance, without triggering a failover, use the 'manual' subcommand. Here is an example:

```
nmc:/$ setup group rsf-cluster (name-of-the-cluster) shared-volume (name-of-the-volume) manual
```

The switchover mode defines whether or not an appliance will attempt to start a service when it is not running. There are separate switchover mode settings for each appliance that can run a service.

The switchover modes can be set to automatic or manual. In automatic mode, the appliance will attempt to start the service in question when it detects that no sibling appliance in the cluster is available or running it. In manual mode, it will not attempt to start the service but will generate warnings when it is unavailable. If the appliance cannot obtain a definitive answer regarding the state of the service (because it cannot contact its siblings in the cluster) or the service is not running anywhere else, the appropriate timeout must expire before any action can be taken. The primary service switchover modes are typically set to automatic to ensure that a appliance starts its primary service(s) on boot up. Note that putting a service into manual mode when the service is already running does not stop that service, it only prevents the service from being started on that appliance.

8.7 System Upgrades

Occasionally, you may need to upgrade SGI NAS software on the appliance. Since this may require a reboot it needs to be managed carefully in a cluster environment. The user is reminded that the cluster service will not be available during the upgrade.

8.7.1 Upgrade procedure

Let's assume we have nodeA and nodeB; nodeA is active and nodeB is passive:

```
nmc@nodeA:/$ setup group rsf-cluster HA-Cluster show
PROPERTY                               VALUE
name                                    : HA-Cluster
appliances                              : [nodeA nodeB]
hbipifs                                 : nodeA:nodeB: nodeB:nodeA:
netmon                                  : 1
info                                    : SGI NAS HA Cluster
generation                              : 1
refresh_timestamp                       : 1369784179.6838
type                                     : rsf-cluster
creation                                 : May 28 16:36:19 2013
SHARED VOLUME: tank1
svc-tank1-shared-vol-name               : tank1
svc-tank1-ipdevs                        : sginas nodeA:igb3 nodeB:igb3
svc-tank1-main-node                     : nodeA
svc-tank1-inittimeout                   : 20
svc-tank1-runtimeout                   : 8
svc-tank1-mhdc-disable                  : n
HA CLUSTER STATUS: HA-Cluster
nodeA:
  tank1      running      auto      unblocked  sginas     igb3       20      8
nodeB:
  tank1      stopped     manual   unblocked  sginas     igb3       20      8
```

Use the following sequence of actions to perform the upgrades on both nodes:

1. Login to nodeB and upgrade by running:

```
nmc:/$ setup appliance upgrade
```

2. Make sure the upgrade on nodeB successfully finished. Login to nodeA and failover to nodeB by running the following command:

```
nmc:/$ setup group rsf-cluster <group name> <shared volume name> nodeB
```

3. After failover finishes, nodeA becomes passive. Run the upgrade command on nodeA:

```
nmc:/$ setup appliance upgrade
```

4. Run the failover command on nodeB:

```
nmc:/$ setup group rsf-cluster <group name> <shared volume name> nodeA
```

Now, nodeA becomes active again.

9 Service Failover

As discussed previously, system failures will result in the failover of ownership of the shared volume to the alternate node. Additionally the systems are likely to be running storage services such as auto-snap, auto-sync, and auto-tier. As part of the failover process, HA Cluster migrates the storage services associated with the shared volume(s) and restarts the services on the alternate node.

10 Advanced Setup

10.1 Name Resolution

Appliances in the HA cluster group must be "resolvable" from each other, as far as host name resolution is concerned. To achieve this, you can either configure your DNS server accordingly, or add records to `/etc/hosts`. If you don't want to edit `/etc/hosts` manually, you may do this automatically at volume sharing (service) creation time. You will be asked to enter virtual shared service hostname and virtual IP address. Definition of these parameters will allow the software to modify `/etc/hosts` tables automatically on all HA-cluster group members.

Note, that you can use a virtual IP address instead of a shared logical hostname.

These host records may be added automatically during shared service creation time. If the given shared logical hostname is not resolved from one group member, then you need to define the IP address for it and records will be added automatically.

```
nmc@nodeA:/$ setup group rsf-cluster HA-Cluster shared-volume add
Scanning for volumes accessible from all appliances ...
Shared volume           : ha-test2
VIP1 Shared logical hostname : rsf-data2
Warning! Failed to resolve logical hostname 'rsf-data2' on appliance 'nodeA'
Shared logical hostname 'rsf-data2' must be resolvable from
all appliances in the cluster.

You can choose to manually configure your DNS server, or local hosts
tables on the appliances (see 'setup appliance hosts -h' for details).

Alternatively, you could allow this cluster configuration logic to update
your local hosts records automatically.
Press No to leave the appliances' local hosts tables intact.
Proceed to modify the hosts table(s)? Yes
IP address for shared logical hostname 'rsf-data2': 172.16.3.23
System host file was changed all group members. Backup stored as /etc/hosts.rsfc.
VIP1 Network interface at nodeB: e1000g1
```

```
VIP1 Network interface at nodeA: e1000g1
Verifying logical (failover) IP address '172.16.3.23' ...Success.
VIP1 Failover Netmask      :
The IP configuration to be used with 'rsf-data2' is: 172.16.3.23. Please confirm
that this configuration is correct ?   Yes
Stop adding VIPs?   Yes
Main node            : nodeA
Initial timeout     : 60
Standard timeout    : 60
Use SCSI reservation (SCSI PGR) for additional protection? (Issue scsi reservations
on the devices in a volume before importing it; this is done to enforce data
integrity (i.e. prevents the pool being imported on two nodes at any one time).
Under normal circumstances this option should be left enabled and should only be
turned off if instructed to do so by UIK support staff.)   Yes
Adding shared volume 'ha-test2', please wait ...
Feb 14 06:54:28 nodeA RSF-1[8543]: [ID 702911 local0.alert] RSF-1 cold restart: All
services stopped.
Waiting for add operation to complete ... done.

HA CLUSTER STATUS: HA-Cluster
nodeB:
  ha-test      stopped  auto   unblocked  rsf-data    e1000g1   60  60
  ha-test2     stopped  auto   unblocked  rsf-data2   e1000g1   60  60
nodeA:
  ha-test      running  auto   unblocked  rsf-data    e1000g1   60  60
  ha-test2     running  auto   unblocked  rsf-data2   e1000g1   60  60
```

If you decide to do the work manually, you need to make sure that shared cluster volume is configured on each cluster node's /etc/hosts file, as shown below:

172.16.3.20 nodeA nodeA.mydomain

172.16.3.21 nodeB nodeB.mydomain

172.16.3.22 rsf-data

172.16.3.23 rsf-data2

10.2 Cache devices

SGI NAS allows you to configure specific devices in a data volume to be cache devices.

For example, using solid-state disks as cache devices can improve performance for random

read workloads of mostly static content. Cache devices can be specified when the volume is created, or added later using the “setup volume grow” NMC command.

Cache devices are also available for shared volumes in the HA Cluster. However, it should be noted that local disks used as cache can't be failed over since they are not accessible by the alternate node. After failover therefore the volume will be marked “Degraded” because of the missing devices. If local cache is critical for performance, then you should set up local cache devices for the shared volume on each cluster node when the volume is first configured. This involves setting up local cache on one node, and then manually failing over the volume to the alternate so that local cache can be added there as well. This will ensure the volume will have cache devices available automatically after a failover, but a drawback is that the data volume will always be “Degraded” because there will always be cache devices that are unavailable.

Additionally users can control the cache settings for Zvols within the data volume. In a cluster, the Zvol cache policy needs to be “write-through” not “write-back”. The following steps in NMC can be used to administer the cache policy.

```
nmc:/$ setup volume <name> zvol cache
nmc:/$ setup zvol cache
nmc:/$ show zvol cache
```


11 Testing and Troubleshooting

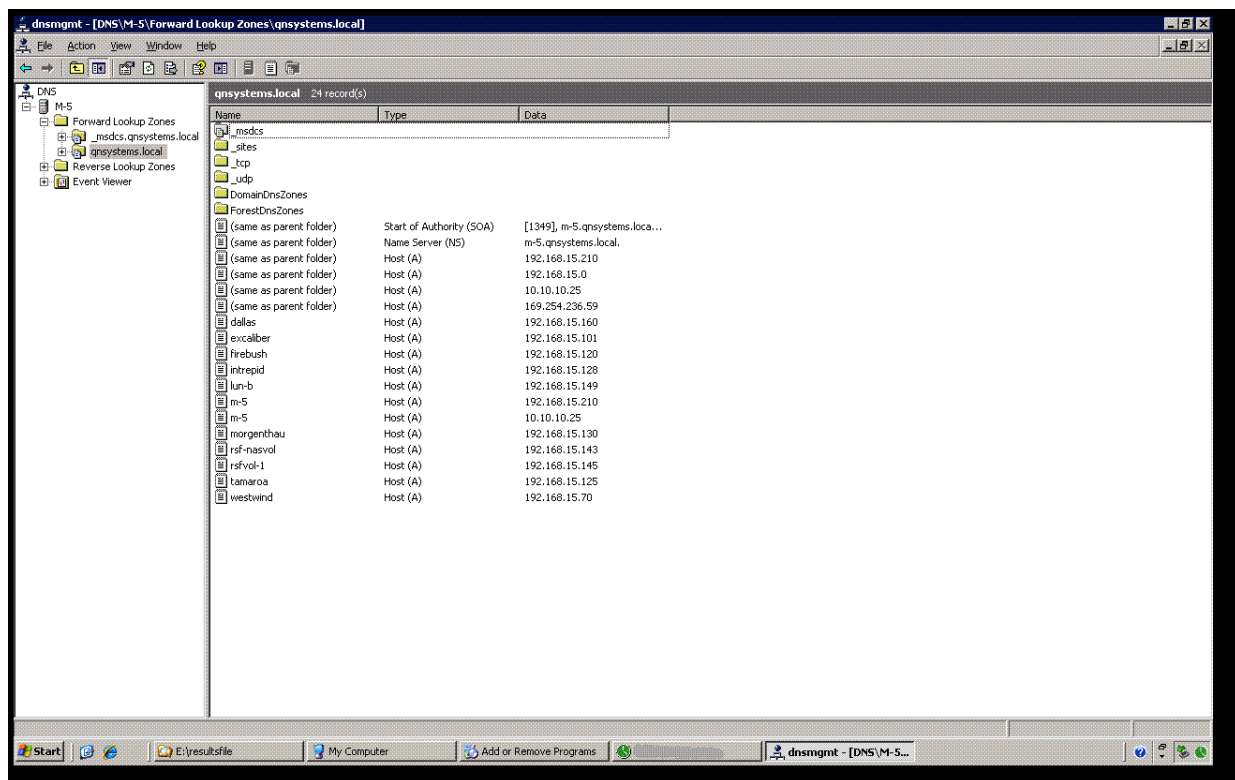
A storage service failover can be triggered manually using NMC. This is done using the “failover” subcommand, as shown here:

```
nmc:/$ setup group rsf-cluster (name-of-the-cluster) (name-of-the-appliance) failover
```

Initiate administrative (manual) failover. Perform failover from the current appliance to the specified appliance. This will cause the volume sharing service to be stopped (and the volume getting exported) on the appliance that is currently providing volume-sharing services, and the opposite actions taking place on the specified appliance.

11.1 Verify DNS entries

There is a name associated with the cluster that is referred to as a shared logical hostname. This name needs to be resolvable by the clients that will be accessing the cluster. One way to do this is to use DNS. The following screen shows running the DNS management application on Windows and viewing the host record of the shared cluster hostname to verify that it was setup properly.



11.2 Verify moving a resource between nodes of a cluster

A manual failover can be used to move a resource from one SGI NAS node to another node in the cluster. The following steps illustrate moving the shared volume “ha-test2” from nodeA to nodeB:

```
nmc@nodeA:/$ setup group
Option ?  rsf-cluster
Option ?  HA-Cluster
Option ?  shared-volume
Option ?  ha-test2
Option ?  show
volume: ha-test2
state: ONLINE
scan: none requested
config:

      NAME          STATE      READ WRITE CKSUM
      ha-test2      ONLINE    0     0     0
      c2t3d0        ONLINE    0     0     0

errors: No known data errors

HA CLUSTER STATUS: HA-Cluster
nodeA:
  ha-test2      running      auto      unblocked  rsf-data2   e1000g1    60  60
nodeB:
  ha-test2      stopped     manual    unblocked  rsf-data2   e1000g1    60  60
```

```
nmc@nodeA:/$ setup group
Option ?  rsf-cluster
Option ?  HA-Cluster
Option ?  shared-volume
Option ?  ha-test2
Option ?  failover
Appliance      : nodeB
```

```
Waiting for failover operation to complete ..... done.

nodeB:
  ha-test2      running      auto      unblocked  rsf-data2   e1000g1    60  60
```

11.3 Verify failing service back to original node

```
nmc@nodeA:/$ setup group
Option ?  rsf-cluster
Option ?  HA-Cluster
Option ?  shared-volume
Option ?  ha-test2
Option ?  failover
Appliance      : nodeA
SystemCallError: (HA Cluster HA-Cluster): cannot set mode for cluster node 'nodeA':
Service ha-test2 is already running on nodeA (172.16.3.20)
```

```
nmc@nodeA:/$ setup group
Option ?  rsf-cluster
Option ?  HA-Cluster
Option ?  shared-volume
Option ?  ha-test2
Option ?  failover
Appliance      : nodeB
Waiting for failover operation to complete ..... done.

nodeB:
  ha-test2      running      auto      unblocked  rsf-data2   e1000g1    60  60
```

11.4 Gathering Support Logs

Click  button under '**Logs**' column to view support logs.

Start-up and Shut-down log for cluster nodeA, volume ha-test

- The cluster logs can be sent to support technicians via configured [SMTP mail server](#).

To submit logs for support please fill out the following details and click or

Company

Contact Email

Brief Description

Startup log

```
Service ha-test start log capture file created at Tue Aug 28 16:48:06 2012
[3547 Aug 28 16:48:06] [ha-test rsfexec] ping_ip_address: checking rsf-data does not exist on the
network
[3547 Aug 28 16:48:08] [ha-test rsfexec] Address rsf-data is not in use, OK to start services.
[3547 Aug 28 16:48:09] [ha-test rsfexec] Running S01announce start 1

*****
[3636 Aug 28 16:48:09] [ha-test S01announce] Startup of service:ha-test started - attempt:1
[3547 Aug 28 16:48:09] [ha-test rsfexec] Running S02ApplianceStarting start 1
[3662 Aug 28 16:48:10] [ha-test S02ApplianceStarting] ifconfig state before interface plumbing:
lo0: flags=2001000849<UP,LOOPBACK,RUNNING,MULTICAST,IPv4,VIRTUAL> mtu 8232 index 1
    inet 127.0.0.1 netmask ff000000
```

12 Contact information

12.1 Support request

To contact support at SGI, click on 'Support' in NMV as it's marked with red square on the screen below.

The screenshot shows the SGI NAS web interface. At the top, there is a navigation bar with links for 'About', 'Support', 'Add Capacity', 'Register', and 'Help'. The 'Support' link is highlighted with a red square. Below the navigation bar, there are several tabs: 'Status', 'Settings', 'Data Management', 'Analytics', 'Namespace Cluster', and 'VM DataCenter'. The 'Support' tab is active, and the page title is 'REQUEST FOR TECHNICAL SUPPORT'. The form contains the following fields:

- Company:** Text input field.
- Contact E-Mail:** Text input field with the value 'root@localhost'.
- Category:** Dropdown menu with 'Other' selected. Below it, the text 'General SGI NAS issue -> Other' is visible.
- Subject:** Text input field.
- Verbosity:** Dropdown menu with 'Verbose' selected. Below it, the text 'includes extended logging and diagnostics.' is visible.
- Comment:** Large text area for entering the support request details.

At the bottom of the form, there is a 'Send Request' button. Below the form, there are three links: 'Found a bug?', 'Feature request?', and 'Request Technical Support'.

or type the following NMC command:

```
nmc:/$ support
```

which will then prompt for a subject and message.

12.2 Other resources

For licensing questions, please contact your SGI sales or support representative.

Product Support

SGI provides a comprehensive product support and maintenance program for its products. For a full description of this program, do one of the following:

- See <http://www.sgi.com/support/>.
- If you are in North America, contact the Technical Assistance Center at 1 (800) 800 4SGI or contact your authorized service provider.
- If you are outside North America, see the following website for the appropriate Customer Service phone number: <http://www.sgi.com/support/supportcenters.html>.