

# HA Cluster Plugin User Guide

## Table of Contents

### [High Availability - HA Cluster](#)

[Definitions & Terms](#)

[Heartbeats](#)

[Network Monitoring](#)

[Fencing](#)

[Getting Started](#)

[Usage](#)

[F.A.Q.](#)

[Sample HW Configuration](#)

[References](#)

# High Availability – HA Cluster

NexentaStor Version 2.0 includes fully automated High Availability cluster.

NexentaStor rsf-cluster [plugin](#) is based on the RSF-1 (<http://www.high-availability.com>) The latter, first released in 1995, is an industry-leading high-availability and cluster middleware application that ensures critical applications and services are kept running in the event of system failures.



For information on NexentaStor plugins, please visit NexentaStor [F.A.Q.](#) pages, or see Section "[Frequently Asked Questions](#)" in this document.

RSF-1 (Resilient Server Facility) "sits" between the storage volume management and application layers of typically: web servers, application servers, firewall servers and database servers, providing support for most leading applications in those.

This article provides further definitions and describes the capabilities in detail.

## Definitions & Terms

RSF-1 cluster is defined as a number of NexentaStor appliances running a defined set of services and monitoring each other for failures. These NexentaStor appliances are interconnected by means of various communication channels, through which they exchange information about their states and the services running on them (heartbeats).

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

NexentaStor RSF-1 cluster provides the basic volume sharing service. RSF-1 cluster makes a given shared volume (or multiple volumes) Highly Available (HA), by providing access on a near-continuous basis, regardless of certain common types of hardware and software failure, and maintenance requirements.

RSF-1 cluster may contain two or more NexentaStor appliances. None of the appliances in the cluster is specifically designated to be the "primary" or "active", as opposed to being "secondary" or "passive". In fact, any appliance in the cluster may be replaced by any other appliance in this same cluster, as far as cluster provided service is concerned.

Because RSF-1 servers must be certain that an appliance (member of the cluster) is down before taking over its services, RSF-1 is normally configured to use several communication channels to exchange heartbeats.

NexentaStor appliances in the RSF-1 cluster constantly monitor each other states via 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.

## Heartbeats

If no services are shared between two particular NexentaStor 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.

RSF-1 cluster supports 3 types of heartbeat communication channels:

1. Shared disk accessible and writeable from all appliances in the cluster (also sometimes called "**quorum** device")
2. Ethernet link

### 3. Serial link

For more information, please see a note on Fencing (below).

Appliances in the RSF-1 cluster periodically synchronize, by exchanging their respective configurations. No manual intervention required - the respective configurations are continuously stored and backed up on-change, upon creation (and for the lifetime) of a RSF-1 cluster.

RSF-1 cluster is capable of providing a variety of High Availability (HA) services. RSF-1 ensures service continuity in presence of service level exceptional events, including power outage, disk failures, appliance running out of memory or crashing, etc.

In its first release, the NexentaStor rsf-cluster plugin provides a fundamental service of sharing a given ZFS volume. In the NexentaStor rsf-cluster, a given shared volume will be accessible (and all its network shares will be available) even if one of the appliances goes down or becomes unresponsive.

Note that RSF-1 cluster IS A group of appliances, and therefore provides a superset of the corresponding NexentaStor "basic group" functionality. In particular, you could still use the generic 'switch' command, to switch Management Console to operate in a group mode - that is, execute CLI 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)

### Network Monitoring

NexentaStor RSF-1 cluster constantly monitors availability, as far as (NFS, CIFS, iSCSI, etc.) clients on the network are concerned. The 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 (default 3)
- Y represents the number of heartbeats an interface must be observed as up before marking it available again to the cluster (default 2).

The current defaults are 3 and 2 heartbeats, respectively.

### Fencing

NexentaStor RSF-1 cluster provides reliable fencing through the utilisation 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.

In addition, NexentaStor RSF-1 cluster provides a number of other failsafe mechanisms:

1. 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.
2. 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.

## Getting Started

1. Send request for rsf-plugin to sales at nexenta.com. Specify two license keys for the appliances that you intend to use in a cluster.
2. The plugin will be uploaded into your repository and will be available for installation. Use web GUI (NMV) or management console to install the cluster. Note that the cluster needs to be installed on all appliances in the cluster. If you are using NMC, run 'setup plugin install rsf-plugin' to install the software. See F.A.Q. article [What is NexentaStor plugin?](#) for more information.
3. Once installed, execute the following NMC command: 'create group rsf-cluster'. For manual pages, use -h (help) option: 'create group rsf-cluster -h'
4. Add shared volume. See extended example below.

## Usage

The following example<sup>1</sup> demonstrates creation of a two-node cluster, with the subsequent addition of a volume-sharing service. Note that cluster supports multiple volume-sharing services (which is the same as saying that more than one volume can be shared).

```
nmc:/$ setup group rsf-cluster create
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 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.
```

Next, we add a volume-sharing service:

```
nmc:/$ setup group rsf-cluster cluster-example shared-volume add
Scanning for volumes accessible from all appliances ...
Shared volume              : clldata2
Shared logical hostname    : clserv2
nodeA                      : e1000g1
nodeB                      : e1000g1
```

---

<sup>1</sup> Currently, NexentaStor does not provide web GUI to manage the HA cluster. For a quick reference of all related console commands, please run 'help keyword rsf', or see 'help keyword' in this document.

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. Should the 'nodeA' fail the volume will be accessible through the 'nodeB' transparently for the users, and vice versa.

## F.A.Q.

- **Question:** how does the cluster prevent access to the same shared volume from two appliances in the cluster?
- Answer: at any point in time a given shared volume is accessed only and exclusively via the appliance that is currently providing the corresponding volume-sharing service (see above). The exclusivity is ensured by multiple heartbeat channels that connect appliances in the cluster, as well as the cluster's 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](#).
- **Question:** does the cluster use SCSI-3 PGR to ensure exclusive access?
- Answer: not at the time. SCSI-3 PGR 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. Having said that, we do plan to provide a configurable ability to make use of PGR (Note: please see related F.A.Q. entry on SCSI-2 PGR).
- **Question:** Is SCSI-2 reservations supported?
- Answer: Yes, SCSI-2 reservation has been added in [NexentaStor version 2.1](#).
- **Question:** What is a "shared logical hostname"?
- A hostname that maps onto the failover IP interface on all the appliances in the cluster. This hostname is expected to be used by NFS/CIFS/iSCSI/etc. clients to reliably access their respective destinations. The cluster provides a built-in network monitor (see "Network Monitoring" above) to ensure uninterrupted access to clustered resources (such as shared volumes).
- **Question:** After installing the plugin and setting up the cluster, can I simulate or try

## failover manually?

- Yes, simply use the 'failover' subcommand. The corresponding man page follows below<sup>2</sup>:

```
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.

- **Question:** I need to take appliance in the cluster offline for maintenance, but don't want to trigger failover. How can I do that?
- Use 'manual' subcommand<sup>3</sup>:

```
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.

- **Question:** Does the cluster handle a failure of storage (eg., SAS) interconnect?
- Answer: Generally, cluster provides node level redundancy. Controller level redundancy is provided via MPxIO multipathing. For more information, please see the following F.A.Q. entry: [Is it possible to use I/O multipathing? How?](#)
- **Question:** What will happen if NFS server crashes? Will this trigger failover?
- Answer: We are planning to integrate a number of application-specific agents, to provide (application-specific) high-availability for NFS, CIFS, and iSCSI ("applications") in the first place. The first release of the cluster (which is part of NexentaStor 2.0 release) does not include this functionality.
- **Question:** Are users replicated in the cluster?
- Answer: No, local Unix users are not replicated. This feature is currently not being planned. We would recommend to use LDAP (or Active Directory) for centralized user management.
- **Question:** I have auto-snap (automated scheduled snapshots) running on my appliance.

---

<sup>2</sup> As always, use -h (help) option to display manual pages

<sup>3</sup> Use -h (help) option to display manual pages

### Will the snapshots continue upon failover?

- Answer: Immediately after NexentaStor 2.1 release we'll start working to provide high-availability for a number of SMF services, including certainly auto-snap, auto-sync, and auto-tier. One immediate workaround would be to use NMC command 'setup appliance configuration' to save the appliance's configuration, then copy the corresponding file over to the other appliance, and use the same command 'setup appliance configuration' to recreate the services from their persistent state. This logic is in fact automated with the existing [Simple Failover plugin](#).
- **Question:** Setting up an HA cluster is complicated due to the number of interconnects and components.. Do you provide turn-key support?
- Answer: Yes. In fact, it is **strongly recommend** to get turn-key support from us or our partners that provided the hardware for the cluster.
- **Question:** Are there any guidelines for characteristics of the quorum drive, size, speed etc?<sup>4</sup>
- Answer: The only requirement is a single small partition per volume - for heartbeating 1Mb is more than sufficient (in fact it's generally configured to write to blocks 34, 36 and 38 only). There is also no speed requirement, 5400rpm drives are more than capable. At present the NexentaStor appliance utilizes one disk for this purpose although this may change at future releases or could be "tuned" by a certified NexentaStor channel partner.
- **Question:** How do I configure HA Cluster with IPMP?
- Answer: The work is underway right now to add IPMP multi-pathing to the NexentaStor management interfaces. Separately, for NIC level redundancy consider using link aggregation (Section "[Link Aggregation](#)").

In addition, this article provides a simple step-by-step configuration for IPMP 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_VIP
```

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\_VIP 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
```

---

<sup>4</sup> This and the rest F.A.Q. articles in this section are quoted from our partner's website <http://www.high-availability.com/FAQ> with little or no modifications

**/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 ffffffff 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_VIP` should then be declared in the RSF-1 configuration file for a single service VIP:

```
SERVICE example RSF_VIP "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.

• **Question:** Where can I find the RSF-1 Cluster Documentation?

• **Answer:** You can find all the RSF-1 documentation on the high-availability.com website at <http://www.high-availability.com/links/4-0-support.php>. This will include the RSF-1 Administration Guide, The RSF-1 Quickstart Guide and the RSF-1 Agent Framework Documentation.

• **Question:** How do I Install the HA Plugin for NexentaStor?

• **Answer:** Please see the [HA Cluster](#) plugin web page.

• **Question:** How do I check, if our HA Cluster services are in Auto or Manual failover mode?

• **Answer:** Use NMC 'show group rsf-cluster' command. Here's an example that shows two shared volumes:

```
nmc$ show group rsf-cluster test
name                : test
appliances          : [host1.nexenta.com host2.nexenta.com]
hbipifs             : host1.nexenta.com:host2.nexenta.com:
                    host2.nexenta.com:host1.nexenta.com:
netmon              : 1
hbdisks             : host1.nexenta.com:c2t4d0 host2.nexenta.com:c2t4d0
type                : rsf-cluster
creation            : Aug 10 15:31:58 2009

SHARED VOLUME: cldata
svc-cldata-failover-hostname : clserv
svc-cldata-ipdevs           : host2.nexenta.com:e1000g1 host1.nexenta.com:e1000g1
svc-cldata-shared-vol-name  : cldata

SHARED VOLUME: cldata2
svc-cldata2-failover-hostname : clserv2
svc-cldata2-ipdevs           : host2.nexenta.com:e1000g1 host1.nexenta.com:e1000g1
svc-cldata2-shared-vol-name  : cldata2

HA CLUSTER 1.0 STATUS:
host1:
  cldata2    stopped    manual  unblocked  clserv2    e1000g1   60  60
  cldata     running   auto    unblocked  clserv     e1000g1   60  60
host2:
  cldata2    running   auto    unblocked  clserv2    e1000g1   60  60
  cldata     stopped   auto    unblocked  clserv     e1000g1   60  60
```

• **Question:** Could you list all CLI commands available to manage HA Cluster?

• **Answer:** Use 'help keyword' to locate the corresponding NMC commands. For instance: 'help keyword rsf' or 'help keyword cluster'

• **Question:** Can HA Cluster be configured to fail over services if network connectivity is lost.

• **Answer:** Yes, and this is also the default setting (netmon = 1 in the 'show group' printout above). You may choose to disable network monitoring at cluster creation time - this, however, is not recommended.

RSF-1 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, RSF-1 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 RSF-1 will re-enable fail over for that interface.

- **Question:** What does broken(safe) and broken(unsafe) mean in the output of NMC command 'show rsf-cluster', and how do I fix it?
- Answer:  
<http://www.high-availability.com/FAQ/index.php?action=artikel&cat=1&id=3&artlang=en>

### Sample HW Configuration

- Two x86/64 boxes with SAS-connected JBOD, 2 Network Interface Cards
- One shared disk must be allocated as quorum device (heartbeat)
- One NIC must be allocated for shared logical hostname
- Second NIC may be used for heartbeat (optional)
- Serial ports can be used for heartbeat (optional)

### References

- <http://www.high-availability.com>
- [Simple Failover](#)
- [AutoCDP](#)
- [What is a NexentaStor plugin?](#)