



**Red Hat Reference Architecture Series**

# **Demonstrating Red Hat<sup>®</sup> Enterprise Linux<sup>®</sup> (RHEL) 5 Virtualization**

**Volume 2: Clustering**

Version 1.0

September 2008





## Demonstrating Red Hat® Enterprise Linux® 5 Virtualization Volume 2: Clustering

Copyright © 2008 by Red Hat, Inc.

1801 Varsity Drive  
Raleigh NC 27606-2072 USA  
Phone: +1 919 754 3700  
Phone: 888 733 4281  
Fax: +1 919 754 3701  
PO Box 13588  
Research Triangle Park NC 27709 USA

"Red Hat," Red Hat Linux, the Red Hat "Shadowman" logo, and the products listed are trademarks or registered trademarks of Red Hat, Inc. in the United States and other countries. Linux is a registered trademark of Linus Torvalds.

All other trademarks referenced herein are the property of their respective owners.

© 2008 by Red Hat, Inc. This material may be distributed only subject to the terms and conditions set forth in the Open Publication License, V1.0 or later (the latest version is presently available at <http://www.opencontent.org/openpub/>).

The information contained herein is subject to change without notice. Red Hat, Inc. shall not be liable for technical or editorial errors or omissions contained herein.

Distribution of substantively modified versions of this document is prohibited without the explicit permission of the copyright holder.

Distribution of the work or derivative of the work in any standard (paper) book form for commercial purposes is prohibited unless prior permission is obtained from the copyright holder.

The GPG fingerprint of the [security@redhat.com](mailto:security@redhat.com) key is:  
CA 20 86 86 2B D6 9D FC 65 F6 EC C4 21 91 80 CD DB 42 A6 0E



# Table of Contents

<u>1 About this document.....</u>	<u>5</u>
<u>1.1 Audience.....</u>	<u>5</u>
<u>1.2 References .....</u>	<u>5</u>
<u>1.3 Document Conventions.....</u>	<u>5</u>
<u>1.4 Terms and Acronyms.....</u>	<u>6</u>
<u>2 Introduction.....</u>	<u>6</u>
<u>2.1 What is Virtualization?.....</u>	<u>6</u>
<u>2.2 What is a Cluster?.....</u>	<u>6</u>
<u>2.3 Why combine Virtualization and Clusters?.....</u>	<u>7</u>
<u>3 Hardware Requirements.....</u>	<u>7</u>
<u>3.1 Servers.....</u>	<u>7</u>
<u>3.2 Storage Infrastructure.....</u>	<u>8</u>
<u>3.2.1 Layout using the NFS share of an NFS server or NAS filer.....</u>	<u>8</u>
<u>3.2.2 Layout using a SAN.....</u>	<u>10</u>
<u>4 Operating System Installation and Configuration.....</u>	<u>11</u>
<u>4.1 Partitioning.....</u>	<u>11</u>
<u>4.2 Software Selection.....</u>	<u>11</u>
<u>4.3 Network Configuration.....</u>	<u>12</u>
<u>4.3.1 Fixed Ethernet order.....</u>	<u>12</u>
<u>4.3.2 Client LAN Connection.....</u>	<u>12</u>
<u>4.3.3 Further Network Configuration.....</u>	<u>13</u>
<u>4.3.4 Xen Network configuration.....</u>	<u>14</u>
<u>4.4 NTP configuration.....</u>	<u>14</u>
<u>5 Shared Storage Configuration for NFS Share Based Solution.....</u>	<u>15</u>
<u>6 Shared Storage Configuration for SAN based solution.....</u>	<u>15</u>
<u>6.1 Configuring Device Mapper Multipathing.....</u>	<u>15</u>
<u>6.1.1 DM/multipath software.....</u>	<u>15</u>
<u>7 Red Hat Cluster Configuration.....</u>	<u>17</u>
<u>7.1 Cluster Infrastructure – fencing of dom0.....</u>	<u>17</u>



<a href="#">7.2 Cluster Infrastructure – fencing of virtual guests.....</a>	<a href="#">17</a>
<a href="#">7.3 Cluster Infrastructure – Quorum Disk Configuration.....</a>	<a href="#">18</a>
<a href="#">7.4 Cluster Volume Manager configuration for Virtual Guests.....</a>	<a href="#">19</a>
<a href="#">7.5 GFS configuration for Virtual Guests.....</a>	<a href="#">19</a>
<a href="#">8 Xen preparation and guest installation.....</a>	<a href="#">20</a>
<a href="#">8.1 Prepare cluster nodes to handle Xen Live Migration.....</a>	<a href="#">20</a>
<a href="#">8.2 Xen Guest installation.....</a>	<a href="#">21</a>
<a href="#">8.3 Implementation of Xen guest as a cluster service.....</a>	<a href="#">22</a>
<a href="#">8.4 Testing the cluster.....</a>	<a href="#">23</a>
<a href="#">8.4.1 Test disable SAN / NFS connection.....</a>	<a href="#">23</a>
<a href="#">8.4.2 Test disable Ethernet connection.....</a>	<a href="#">23</a>
<a href="#">8.4.3 Test Xen guest Live Migration.....</a>	<a href="#">23</a>
<a href="#">9 Backup &amp; Restore.....</a>	<a href="#">24</a>
<a href="#">9.1 Backup.....</a>	<a href="#">24</a>
<a href="#">9.2 Restore.....</a>	<a href="#">24</a>
<a href="#">Appendix A – Sample kickstart file.....</a>	<a href="#">25</a>
<a href="#">Appendix B – Complete sample cluster.conf.....</a>	<a href="#">26</a>
<a href="#">Appendix C – Conga Cluster Management.....</a>	<a href="#">27</a>



# 1 About this document

This document is the second of a planned series detailing the use of server virtualization on Red Hat Enterprise Linux (RHEL) 5. This volume provides a guideline for implementing Xen virtual guests on top of Red Hat Cluster Suite. Volume 1 provides details in installing and managing standalone Xen guests.

The goal of this document is to provide documented step by step details on installing, configuring and running a Red Hat Cluster with GFS having Xen virtual guests handled by the Red Hat Cluster Manager.

## 1.1 Audience

The intended audience for this volume is Red Hat Global Professional Services and Red Hat Partners, however an end user will value the instructions detailed.

## 1.2 References

Configuring and Managing a Red Hat Cluster

[http://www.redhat.com/docs/en-US/Red\\_Hat\\_Enterprise\\_Linux/5.2/html/Cluster\\_Administration/index.html](http://www.redhat.com/docs/en-US/Red_Hat_Enterprise_Linux/5.2/html/Cluster_Administration/index.html)

Installation Guide

[http://www.redhat.com/docs/en-US/Red\\_Hat\\_Enterprise\\_Linux/5.2/html/Installation\\_Guide/index.html](http://www.redhat.com/docs/en-US/Red_Hat_Enterprise_Linux/5.2/html/Installation_Guide/index.html)

Virtualization Guide

[http://www.redhat.com/docs/en-US/Red\\_Hat\\_Enterprise\\_Linux/5.2/html/Virtualization/index.html](http://www.redhat.com/docs/en-US/Red_Hat_Enterprise_Linux/5.2/html/Virtualization/index.html)

Deployment Guide

[http://www.redhat.com/docs/en-US/Red\\_Hat\\_Enterprise\\_Linux/5.2/html/Deployment\\_Guide/index.html](http://www.redhat.com/docs/en-US/Red_Hat_Enterprise_Linux/5.2/html/Deployment_Guide/index.html)

**Demonstrating Red Hat® Enterprise Linux® 5 Virtualization**

Volume 1: Installation and Management Basics

## 1.3 Document Conventions

As in most documents with procedural citations, certain paragraphs in this manual are represented in different fonts, typefaces, sizes, and color. This highlighting is helpful in determining command line user input from text file content. Information represented in this manner includes the following:

- command names



Linux commands like `iptables` or `yum` will be differentiated by font.

- user input  
User entered commands and their respective output will be displayed as seen below.

```
# echo "This is an example of command line input and output"  
This is an example of command line input and output  
#
```

- file content  
Listing the content or partial content of a Linux ASCII file will be displayed as seen below.

! This is the appearance of text contained within a file

## 1.4 Terms and Acronyms

CLVMD	Cluster Logical Volume Manager Daemon
DLM	Distributed Lock Manager
dom0	Virtualization host running the Xen kernel
domU	Virtualized guest machine
GFS	Global File System
HBA	Host Bus Adapter

## 2 Introduction

### 2.1 What is Virtualization?

Virtualization allows multiple operating system instances to run concurrently on a single computer; it is a means of separating hardware from a single operating system. Each “guest” OS is managed by a Virtual Machine Monitor (VMM), also known as a hypervisor. Because the virtualization system sits between the guest and the hardware, it can control the guests’ use of CPU, memory, and storage, even allowing a guest OS to migrate from one machine to another.

### 2.2 What is a Cluster?

A cluster is essentially a group of two or more computers working together which, from an end user’s perspective, appear as one server. The high availability aspect of any cluster indicates that it provides services configured in a manner such that a monitored failure on any cluster member will not prevent the continued availability of the service itself.



## 2.3 Why combine Virtualization and Clusters?

A customer can benefit to using a cluster infrastructure to underlay virtualization by:

- increased application uptime due to Live Migration of virtual guests
- the automatic fail-over of virtual guests when failure is detected and handled by Red Hat Cluster Suite
- cost savings derived from using virtualization to consolidate existing servers

## 3 Hardware Requirements

### 3.1 Servers

The minimum and recommended server hardware configurations for the cluster nodes are shown in **Table 1**. At least two servers are needed to provide the benefits of increased availability using clusters.

Requirement	Minimum	Recommended
Processor	1 x PIII 800MHz or equivalent with PAE support	2 – 4 x Xeon/Opteron, 32bit or 64 bit with PAE support
RAM	1024 MB	4 - 8 GB (32 bit) 8-64 GB (64 bit)
On board disk	1 x 20GB single IDE	2 x 73GB SCSI in hardware RAID 1
Network	1 x 100BaseT	4 x 1000BaseT
External Storage Connection	1 x NIC for iSCSI	2 x 2Gbit Fibre Channel HBA
Fencing device	At least 1 of the Red Hat Cluster Suite supported fencing devices	To provide redundant fencing, make sure there exists a 2 <sup>nd</sup> supported fence device

**Table 1: Minimum and recommended server configuration**

These recommendations are not based on performance criteria – rather they are based on what actual customers are deploying.

Red Hat recommends having a minimum of 512 MB RAM available per virtual guest. Have in mind a single server has to run all virtual guests in case of fail-over. Since a single server must be able to run all guests in case of a failover, the total memory required per server would be the sum of that required all virtual guests plus 256MB for the hypervisor.



## 3.2 Storage Infrastructure

The storage must be accessible using the same technology for all the physical system in the cluster. Currently, the supported storage technologies are:

- NFS Server / NAS-Filer
- Multi-initiator SAS
- iSCSI
- Fibre attached storage

This document provides example configurations using an NFS share and SAN storage, the most common deployments.

### 3.2.1 Layout using the NFS share of an NFS server or NAS filer

Using a NFS share eliminates the need to use the following Red Hat Cluster Suite components:

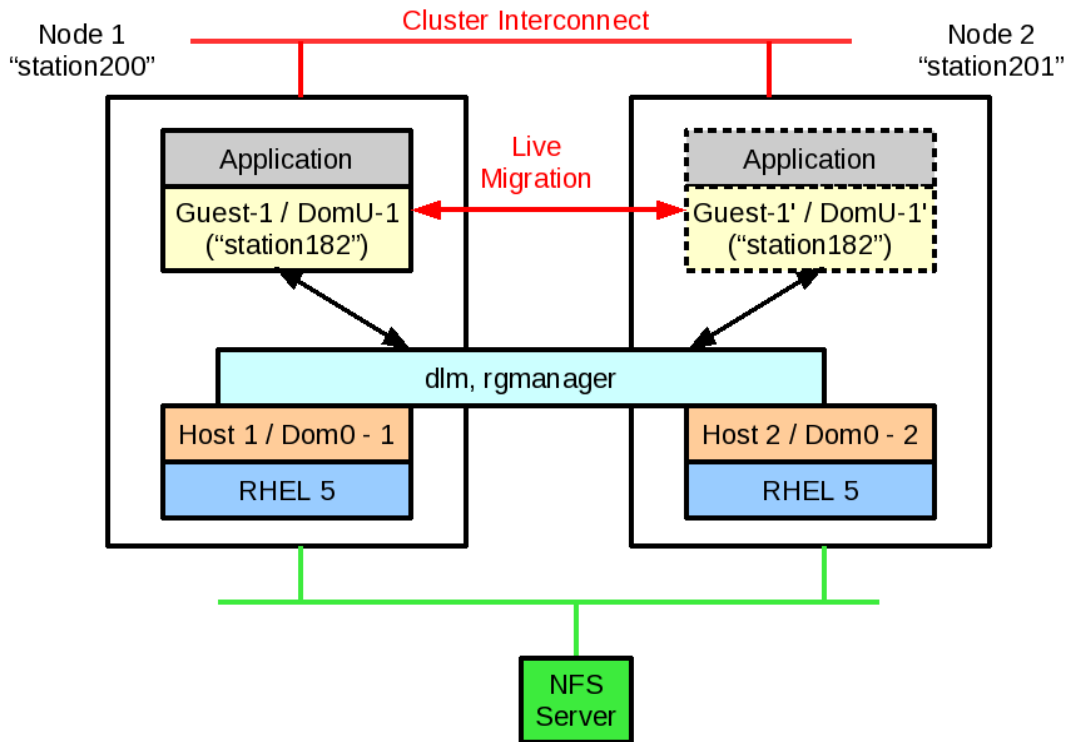
- Quorum Disk
- Cluster Volume Manager
- GFS

With the use of an NFS share the following points should be considered:

- With virtual guest images will reside on the NFS share, therefore the NFS Server becomes a single point of failure.
- While Red Hat supports a configuration of a two-node cluster without a quorum disk, a fencing race will occur when regular heartbeats are not detected. Configuring a quorum disk requires shared SAN or iSCSI storage.



The design layout and necessary services of a configuration using an NFS share are shown in **Figure 1**. The virtual guest file images reside on an NFS share and the Red Hat Cluster Suite will handle virtual guest fail-over.

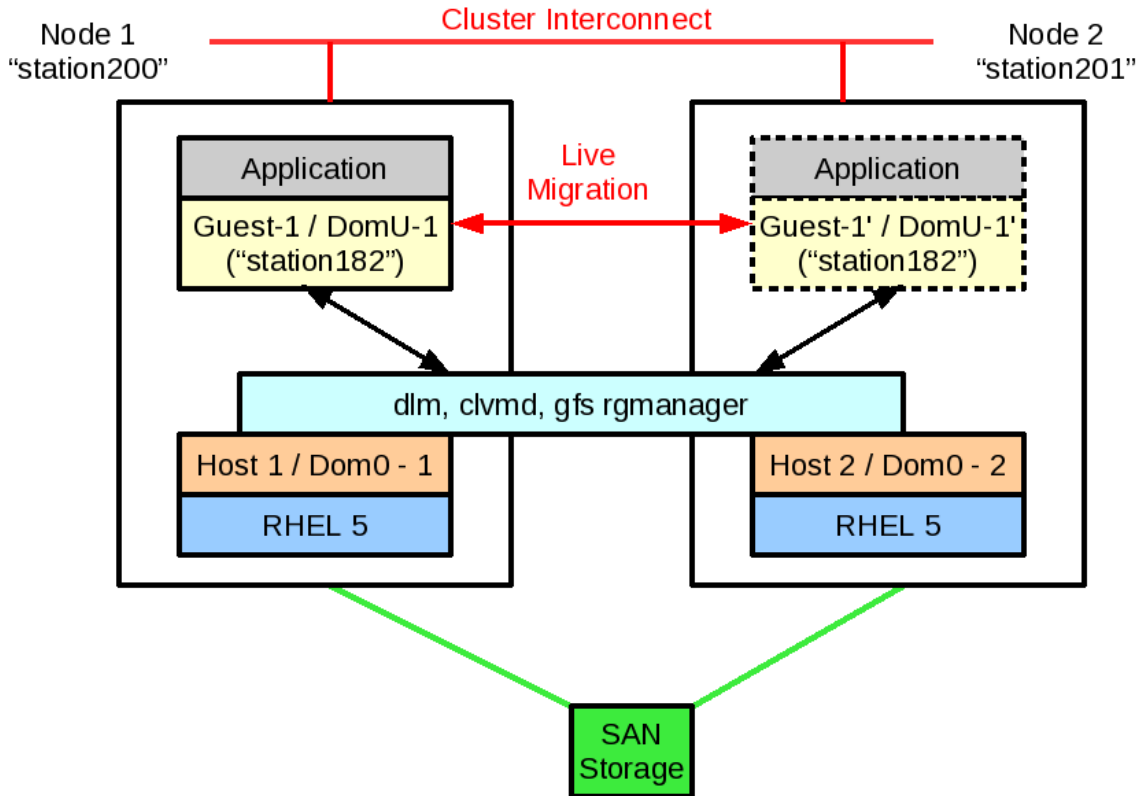


**Figure 1: NFS Layout**



### 3.2.2 Layout using a SAN

The SAN layout in the **Figure 2** can provide multiple redundant paths between each server and the storage. Depending on several factors, such as how many storage processors there are in the storage array, the number of redundant paths per node could vary. Note that with the correct software, the paths can be used for load balancing, therefore increasing throughput between the storage and the servers.



**Figure 2: SAN configuration**



**Table 2** summarizes the hardware requirements for a redundant SAN architecture.

Component	Recommendation
HBA	2 x supported Fibre Channel HBAs per server.
Fibre Channel switches	2 x FC switches. These switches do not necessarily have to be dedicated to the cluster nodes – they can be part of the broader corporate SAN architecture.
FC-attached storage array	Any standard storage array is acceptable. There is no need to dedicate an array to the cluster nodes – space on an existing corporate storage array is fine. Arrays with more than 1 storage processor are preferred as this removes a single point of failure.
Multipathing Software	Red Hat Enterprise Linux includes with device-mapper-multipath support.

**Table 2:** Requirements for redundant SAN

## 4 Operating System Installation and Configuration

Red Hat strongly recommends to using Red Hat Enterprise Linux 5.2 including latest available Errata updates.

Install the operating system on to the two servers using one of these methods:

- Use kickstart from an HTTP server containing the installation tree
- Use (if existing) a RHN Satellite for kickstarting
- Use a PXE provisioning infrastructure based on cobbler
- Use manual installation by CD/DVD set

Refer to **Appendix B** for a sample kickstart file containing all necessary packages.

**Note:** If using SAN storage, ensure that connections to the external storage are disconnected during installation. Otherwise there is a possibility that you might install the operating system to the shared storage rather than to the local disk.

### 4.1 Partitioning

The default partitioning layout provided by the anaconda installer can be used for the host. The additional space used by the virtual guest images will reside on the NFS share or SAN storage.

### 4.2 Software Selection

Choose the default software group selection with the addition of the following package groups and packages as shown from this packages section of a *kickstart* file:

```
@ base
sysstat
iscsi-initiator-utils
@ cluster-storage
kmod-gfs-PAE
```



```
kmod-gfs-xen
@ clustering
@ virtualization
@ X Window System
@ GNOME Desktop Environment
ntp
perl-Crypt-SSLeay
```

## 4.3 Network Configuration

The network configuration described is based on two bonded interfaces, bond0 (eth0/eth2 for Xen and cluster communication) and bond1 (eth1/eth3 for application traffic).

### 4.3.1 Fixed Ethernet order

To prevent the ethernet interface from switching when adding new network card hardware, create and use an additional udev rule at every cluster node which maintains interfaces to MAC addresses.

```
[/etc/udev/rules.d/70-custom.rules]
```

```
ACTION=="add", KERNEL=="eth*", SYSFS{address}=="00:16:55:81:EF:C0", NAME="eth0"
ACTION=="add", KERNEL=="eth*", SYSFS{address}=="00:16:55:81:EF:A3", NAME="eth1"
ACTION=="add", KERNEL=="eth*", SYSFS{address}=="00:16:55:81:EF:E4", NAME="eth2"
ACTION=="add", KERNEL=="eth*", SYSFS{address}=="00:16:55:81:EF:B1", NAME="eth3"
```

### 4.3.2 Client LAN Connection

The client LAN connection will usually consist of two ethernet connections bonded using the shipped bonding driver. The recommendation is to use static IP address configuration rather than DHCP.

If more than one bonded ethernet connections is used, for example a second for a management interface, then change the value of "max\_bonds" accordingly.

This will configure eth0 and eth2 to be slave devices of a virtual device 'bond0' – the server's static IP address on the client LAN is on this interface:

```
[/etc/modprobe.conf - partial]
```

```
options bonding max_bonds=2
alias bond0 bonding
alias bond1 bonding
```

```
[/etc/sysconfig/network-scripts/ifcfg-bond0]
```

```
DEVICE=bond0
IPADDR=192.168.10.XXX
NETMASK=255.255.255.0
ONBOOT=yes
BOOTPROTO=static
BONDING_OPTS="mode=1 miimon=100 max_bonds=2"
```

```
[/etc/sysconfig/network-scripts/ifcfg-eth0]
```

```
DEVICE=eth0
ONBOOT=yes
MASTER=bond0
```



```
SLAVE=yes  
BOOTPROTO=none
```

```
[/etc/sysconfig/network-scripts/ifcfg-eth2]  
DEVICE=eth2  
ONBOOT=yes  
MASTER=bond0  
SLAVE=yes  
BOOTPROTO=none
```

This will configure eth1 and eth3 to be slave devices of a virtual device 'bond1' – the server's static IP address on the client LAN is on this interface:

```
[/etc/sysconfig/network-scripts/ifcfg-bond1]  
DEVICE=bond1  
IPADDR=10.0.0.XXX  
NETMASK=255.255.255.0  
ONBOOT=yes  
BOOTPROTO=static  
BONDING_OPTS="mode=1 miimon=100 max_bonds=2"
```

```
[/etc/sysconfig/network-scripts/ifcfg-eth1]  
DEVICE=eth1  
ONBOOT=yes  
MASTER=bond1  
SLAVE=yes  
BOOTPROTO=none
```

```
[/etc/sysconfig/network-scripts/ifcfg-eth3]  
DEVICE=eth3  
ONBOOT=yes  
MASTER=bond1  
SLAVE=yes  
BOOTPROTO=none
```

### 4.3.3 Further Network Configuration

Edit */etc/resolv.conf* and */etc/sysconfig/network* as normal to set the namervers, hostname and default gateway. Red Hat Enterprise Linux 5 Clustering uses fully qualified domain named names (fqdn), thus, for a two-node cluster with an application interface, the following configuration is needed:

```
[/etc/hosts - partial]  
# Public addresses  
10.0.0.1      node1.publicclan.com node1  
10.0.0.2      node2.publicclan.com node2  
  
# Private Xen + Cluster addresses  
192.168.10.200  station200.example.com station200  
192.168.10.201  station201.example.com station201  
  
# Xen virtual guest addresses  
192.168.10.182  station182.example.com station182  
192.168.10.183  station183.example.com station183
```



```
[/etc/sysconfig/network]
```

```
NETWORKING=yes
NETWORKING_IPV6=yes
HOSTNAME=stationXXX.example.com
GATEWAY=192.168.10.254 # set it to your default gateway
```

### 4.3.4 Xen Network configuration

To properly configure a Xen network bridge (xenbr#) to use the enslaved bonded interfaces perform the following changes:

a) create `/etc/xen/scripts/network-bridge-bonding` like the following and make it executable:

```
#!/bin/bash
dir=$(dirname "$0")
"$dir/network-bridge" "$@" vifnum=0 netdev=bond0 bridge=xenbr0
"$dir/network-bridge" "$@" vifnum=1 netdev=bond1 bridge=xenbr1
```

b) modify `/etc/xen/xend-config.sxp`:

```
(xend-unix-server yes)
(xend-relocation-server yes)
(xend-unix-path /var/lib/xend/xend-socket)
(xend-relocation-port 8002)
(xend-relocation-hosts-allow '^localhost$ ^localhost\\.localdomain$
*\\.example\\.com$')
(network-script network-bridge-bonding)
(vif-script vif-bridge)
(dom0-min-mem 256)
(dom0-cpus 0)
(vncpasswd '')
```

Verify the bridge is working as expected by performing `brctl show`:

```
[root@station200 ~]# brctl show
bridge name bridge id          STP enabled interfaces
xenbr0          8000.fefffffffffff          no          pbond0
               vif0.0
xenbr1          8000.fefffffffffff          no          pbond1
               vif0.1
```

### 4.4 NTP configuration

To guarantee the time and date of the cluster nodes is in sync, configure NTP accordingly. Change `/etc/ntp.conf` to look like the following example:

```
driftfile /var/lib/ntp/ntp.drift
server 192.168.10.254 # IP of NTP server
restrict 192.168.10.254 # IP of NTP server
restrict 127.0.0.1
restrict default notrust nomodify nopeer
```

Start NTP service and make sure it gets started after reboot:

```
# service ntpd start
# chkconfig ntpd on
```



## 5 Shared Storage Configuration for NFS Share Based Solution

To use an NFS Server or an NAS Filer for shared storage, the two mount points, `/etc/sysconfig/sharedvm` and `/var/lib/xen/images`, must be mounted automatically.

Make the directory for the mount.

```
# mkdir -p /etc/sysconfig/sharedvm
```

Edit `/etc/fstab`.

```
192.168.10.254:/data/xen_configs /etc/sysconfig/sharedvm nfs defaults 0 0
192.168.10.254:/data/xen_images /var/lib/xen/images nfs defaults 0 0
```

By choosing an NFS share based solution, the following sections of configuration are skipped:

- Quorum Disk
- device Mapper Multipathing
- Cluster Volume Manager
- GFS

## 6 Shared Storage Configuration for SAN based solution

The shared storage will be used by Xen and Red Hat Cluster Suite as a solution for the following:

- The quorum disk daemon, `qdiskd`, requires at least 15 MB of shared storage
- Red Hat Global File System (GFS) is needed to provide access to the same Xen disk image files through every node of the cluster to be able to perform Live Migration
- GFS will provide access to the same Xen configuration files to remove the need to keep them synchronized manually

The shared storage can be either iSCSI based, Multi-initiator SAS or SAN based.

### 6.1 Configuring Device Mapper Multipathing

Device Mapper Multipath (DM/multipath) is described here since it is shipped and supported with Red Hat Enterprise Linux. Other multipath solutions may be available.

#### 6.1.1 DM/multipath software

Edit `/etc/multipath.conf` and adopt the configuration to suitable to the storage used. Some examples for most common SAN storage configurations can be found at `"/usr/share/doc/device-mapper-multipath*/multipath.conf.defaults"`.

It is recommended that aliases be configured for the LUNs used by the Red Hat Cluster Suite. A sample `/etc/multipath.conf` is shown:

```
multipaths {
    multipath {
```



```
        wwid                360060160e9523400d245f3234343dd11
        alias                quorum
    }
    multipath {
        wwid                360060160e952340a29154234443dd11
        alias                gfs_xenimages
    }
    multipath {
        wwid                360060160e952340a29154234443ff12
        alias                gfs_xenconfigs
    }
}
```

After establishing the configuration above, start multipathing using the following commands:

```
# chkconfigmultipathd on
# service multipathd start
# multipath-v3
```

Next use `fdisk` to create a partition on `/dev/mpath/quorum`, `/dev/mpath/gfs_xenimages` and `/dev/mpath/gfs_xenconfigs` and label for LVM use:

```
# fdisk /dev/mpath/gfs_xenconfigs
The number of cylinders for this disk is set to 1274.
There is nothing wrong with that, but this is larger than 1024,
and could in certain setups cause problems with:
 1) software that runs at boot time (e.g., old versions of LILO)
 2) booting and partitioning software from other OSS
   (e.g., DOS FDISK, OS/2 FDISK)

Command (m for help): n
Command action
  e   extended
  p   primary partition (1-4)
p
Partition number (1-4): 1
First cylinder (1-1274, default 1): <cr>
Using default value 1
Last cylinder or +size or +sizeM or +sizeK (1-1274, default 1274): <cr>
Using default value 1274

Command (m for help): t
Selected partition 1
Hex code (type L to list codes): 8e
Changed system type of partition 1 to 8e (Linux LVM)

Command (m for help): w
The partition table has been altered!

Calling ioctl() to re-read partition table.

WARNING: Re-reading the partition table failed with error 22: Invalid
argument.
The kernel still uses the old table.
The new table will be used at the next reboot.
```



Syncing disks.

To force re-reading the partition table and to update dm-multipath devices run the following commands:

```
# partprobe
# multipath -F
# multipath -v3
```

Repeat these steps for the other devices as well. After this is done, all necessary devices should be in `/dev/multipath/`:

```
# ls -l /dev/multipath/
total 0
lrwxrwxrwx 1 root root 7 Sep  4 01:28 gfs_xenconfigs -> ../dm-1
lrwxrwxrwx 1 root root 7 Sep  4 01:28 gfs_xenconfigspl -> ../dm-3
lrwxrwxrwx 1 root root 7 Sep  4 01:28 gfs_xenimages -> ../dm-0
lrwxrwxrwx 1 root root 7 Sep  4 01:28 gfs_xenimagespl -> ../dm-9
lrwxrwxrwx 1 root root 8 Sep  3 23:58 quorum -> ../dm-10
lrwxrwxrwx 1 root root 8 Sep  3 23:58 quorump1 -> ../dm-11
```

## 7 Red Hat Cluster Configuration

This chapter describes the steps necessary to configure the host/dom0 cluster.

### 7.1 Cluster Infrastructure – fencing of dom0

Fencing is the mechanism used to prevent split brain situations in a cluster. Configure the fence device to be accessible from the cluster nodes by TCP/IP and create a user with the permissions to power off the node. The `cluster.conf` below illustrates an example of a fence configuration via APC power-switches. Red Hat does not recommend relying on only one fence mechanism. If there is only one supported fence device available, `fence_manual` should be added as second method.

A full list of supported fence devices can be found at:

[http://www.redhat.com/cluster\\_suite/hardware/](http://www.redhat.com/cluster_suite/hardware/)

When using Management Board fencing devices, make sure the following is performed on every node:

```
chkconfig acpid off
```

### 7.2 Cluster Infrastructure – fencing of virtual guests

For virtual guest fencing there exist a special fence agent called `fence_xvm`. To configure this agent a key file must first be created:

```
# dd if=/dev/urandom of=/etc/cluster/fence_xvm.key bs=4k count=1
```

Distribute this key file to the other node(s) and add the appropriate entry to the `cluster.conf` file.



## 7.3 Cluster Infrastructure – Quorum Disk Configuration

Execute the following commands to create the quorum disk:

```
mkqdiskd -c /dev/mpath/quorump1 -l quorum
```

The quorum disk supports heuristic methods to verify the TCP/IP connection -- usually used is a gateway IP.

Create the file `/etc/cluster/cluster.conf` with the following content for basic cluster node setup. [Quorum disk part is printed blue, `fence_xvm` is printed brown, `fencing_APC` is printed red and the second fence device, `fence_manual`, is printed green].

```
<?xml version="1.0"?>
<cluster name="dom0cluster1" alias="dom0cluster1" config_version="1">
  <quorumd device="/dev/mpath/quorump1" interval="3" min_score="1" tko="4" votes="1">
    <heuristic interval="4" program="ping 192.168.10.254 -c3 -t1" score="1"/>
  </quorumd>
  <fence_daemon clean_start="1" post_fail_delay="5" post_join_delay="20"/>
  <cman expected_votes="1" two_node="1"/>
  <clusternodes>
    <clusternode name="station200.example.com" nodeid="1" votes="1">
      <fence>
        <method name="1">
          <device name="fence_station200" port="1" switch="1"/>
        </method>
        <method name="2">
          <device name="manualfence" nodename="station200.example.com"/>
        </method>
      </fence>
    </clusternode>
    <clusternode name="station201" nodeid="2" votes="1">
      <fence>
        <method name="1">
          <device name="fence_station201" port="2" switch="1"/>
        </method>
        <method name="2">
          <device name="manualfence" nodename="station201.example.com"/>
        </method>
      </fence>
    </clusternode>
  </clusternodes>
  <fencedevices>
    <fencedevice agent="fence_apc" hostname="172.16.73.12"
      login="fenceuser" name="fence_station200" passwd="PASSWORD"/>
    <fencedevice agent="fence_apc" hostname="172.16.73.13"
      login="fenceuser" name="fence_station201" passwd="PASSWORD"/>
    <fencedevice agent="fence_manual" name="manualfence"/>
    <fencedevice agent="fence_xvm" name="fence_xen"/>
  </fencedevices>
</cluster>
```

Now copy the `/etc/cluster/cluster.conf` to the second node and perform the following steps on each member to start the cluster and make it permanent:

```
# chkconfig cman on
# chkconfig qdiskd on
```



```
# service cman start
# service qdiskd start
```

Verify the basic cluster infrastructure is running as expected by issuing `clustat`:

```
Member Status: Quorate
```

Member Name	ID	Status
-----	----	-----
station200.example.com	1	Online, Local
station201.example.com	2	Online
/dev/mpath/quorumpl	0	Online, Quorum Disk

## 7.4 Cluster Volume Manager configuration for Virtual Guests

After the basic cluster infrastructure is configured, the next step is to create and add resources to the cluster. First, the cluster volumes have to be created then the GFS filesystems.

Enable the Cluster Volume Manager on both nodes:

```
# chkconfig clvmd on
# service clvmd start
```

Create the physical volume, volume group and logical volume for the Xen images and Xen configuration files. These steps only have to be performed on one of the members.

```
# pvcreate /dev/mpath/gfs_xenimagesp1
Physical volume "/dev/mpath/gfs_xenimagesp1" successfully created
# pvcreate /dev/mpath/gfs_xenconfigsp1
Physical volume "/dev/mpath/gfs_xenconfigsp1" successfully created
# vgcreate -s 64M vg_gfs_xenimages /dev/mpath/gfs_xenimagesp1
Volume group "vg_gfs_xenimages" successfully created
# vgcreate -s 64M vg_gfs_xenconfigs /dev/mpath/gfs_xenconfigsp1
Volume group "vg_gfs_xenconfigs" successfully created
# lvcreate -L 100G -n lv_gfs_xenimages vg_gfs_xenimages
Logical volume "lv_gfs_xenimages" created
# lvcreate -L 1G -n lv_gfs_xenconfigs vg_gfs_xenconfigs
Logical volume "lv_gfs_xenconfigs" created
```

## 7.5 GFS configuration for Virtual Guests

Format the logical volumes with GFS file-system using the following command:

```
# mkfs.gfs -b 4096 -j 2 -p lock_dlm -t dom0_cluster1:gfs_xenimages
/dev/vg_gfs_xenimages/lv_gfs_xenimages
```

This will destroy any data on `/dev/vg_gfs_xenimages/lv_gfs_xenimages`.

Are you sure you want to proceed? [y/n] y

```
Device: /dev/vg_gfs_xenimages/lv_gfs_xenimages
Blocksize: 4096
Filesystem Size: 2490068
Journals: 2
Resource Groups: 38
Locking Protocol: lock_dlm
```



```
Lock Table:                dom0_cluster1:gfs_xenimages
```

```
Syncing...  
All Done
```

```
# mkfs.gfs -b 4096 -j 2 -p lock_dlm -t dom0_cluster1:gfs_xenconfigs  
/dev/vg_gfs_xenconfigs/lv_gfs_xenconfigs  
...  
All Done
```

The mount point for the shared virtual guest configuration files must be create at each cluster member.

```
# mkdir -p /etc/sysconfig/sharedvm
```

Add the entries to mount GFS volumes automatically to `/etc/fstab`:

```
/dev/vg_gfs_xenimages/lv_gfs_xenimages /var/lib/xen/images gfs defaults 0 0  
/dev/vg_gfs_xenconfigs/lv_gfs_xenconfigs /etc/sysconfig/sharedvm gfs defaults 0 0
```

To complete the dom0 cluster infrastructure and resources configuration, start and permanently enable GFS:

```
# service gfs start  
# chkconfig gfs on
```

## 8 Xen preparation and guest installation

The following steps prepare Xen on the cluster nodes to handle Live Migration. The Xen guests are created then implemented as cluster resources.

### 8.1 Prepare cluster nodes to handle Xen Live Migration

To enabling Xen Live Migration support, a couple of lines in `/etc/xen/xend-config.sxp` must be changed. Change the `xend-relocation-hosts-allow` entry to include the dom0 system addresses of the configuration being deployed. These changes should have been already completed in the **Network Configuration** chapter, so only verification should be necessary.

```
(xend-unix-server yes)  
(xend-relocation-server yes)  
(xend-unix-path /var/lib/xen/xend/socket)  
(xend-relocation-port 8002)  
(xend-relocation-hosts-allow '^localhost$ ^localhost\\.localdomain$  
*\\.example\\.com$')  
(network-script network-bridge-bonding)  
(vif-script vif-bridge)  
(dom0-min-mem 256)  
(dom0-cpus 0)  
(vncpasswd '')
```

Once this has been completed, restart both dom0 cluster nodes:

```
# reboot
```



## 8.2 Xen Guest installation

Xen guest installation can be done in several ways:

- directly through virt-install
- via graphical virt-manager
- via koan using cobbler
- via RHN Satellite Server

It depends on a customer's existing infrastructure and preference which installation method to choose, but the following must be addressed:

- Verify that "tap:aio" is used rather than "file" for the storage config of Xen guest(s)
- After the installation is complete, the guest will be shutdown to allow the config file(s) to be copied to the shared GFS directory `/etc/sysconfig/sharedvm/`
- Configure the dom0 cluster management the guest(s)

As an generic working example the installation way using "virt-install" is shown:

```
# virt-install
What is the name of your virtual machine? station182.example.com
How much RAM should be allocated (in megabytes)? 512
What would you like to use as the disk (path)?
    /var/lib/xen/images/station182.example.com-disk0.img
Would you like to enable graphics support? (yes or no) no
What is the install location? nfs:install.server.example.com:/rhel5-server-x86

Starting install...
Creating domain...                                0 B 00:04
Bootdata ok (command line is    method=nfs:install.server.example.com:/rhel5-server-x86)
Linux version 2.6.18-58.el5xen (brewbuilder@hs20-bc2-2.build.redhat.com) (gcc version 4.1.2 20070626 (Red Hat 4.1.2-14)) #1 SMP Tue Nov 27 17:15:58 EST 2007
...
```

Copy the Xen guest file to the GFS config Volume making it available on both cluster nodes:

```
# cp /etc/xen/station182.example.com /etc/sysconfig/sharedvm/
```

Edit the guest configuration file:

- change the line starting with: `'disk = [ "file:"` to `'tap:aio:'`
- if desired, bind the guest network to the external Xen bridge interface `xenbr1` as shown below

```
name = "station182.example.com"
uuid = "0d64fcd5-9f51-1f53-855e-fb0fb81df160"
maxmem = 280
memory = 280
vcpus = 1
bootloader = "/usr/bin/pygrub"
on_poweroff = "destroy"
on_reboot = "restart"
```



```
on_crash = "restart"
vfb = [ "type=vnc,vncunused=1" ]
disk = [ "tap:aio:/var/lib/xen/images/station182.example.com-disk0,xvda,w" ]
vif = [ "mac=00:16:3e:00:01:82,bridge=xenbr1" ]
```

### 8.3 Implementation of Xen guest as a cluster service

Extending the *cluster.conf* is necessary to add the Xen guests to cluster control. The “*rm*” section has to be added to the of the *cluster.conf* within the “cluster” tag. The complete *cluster.conf* in **Appendix B** will show the exact placement of the section within the file.

Within the resource manager tag of the *cluster.conf*:

- increase the **version number** in the second line
- add two **failoverdomains** with the created **virtual guests as resources** with opposing priorities

```
<?xml version="1.0"?>
<cluster name="dom0cluster1" alias="dom0cluster1" config_version="2">
...
  <rm>
    <failoverdomains>
      <failoverdomain name="domain1" ordered="1" restricted="1">
        <failoverdomainnode name="station200.example.com" priority="1"/>
        <failoverdomainnode name="station201.example.com" priority="2"/>
      </failoverdomain>
      <failoverdomain name="domain2" ordered="1" restricted="1">
        <failoverdomainnode name="station201.example.com" priority="1"/>
        <failoverdomainnode name="station200.example.com" priority="2"/>
      </failoverdomain>
    </failoverdomains>
    <resources/>
    <vm autostart="0" domain="domain1" exclusive="0"
      name="station182.example.com" path="/etc/sysconfig/sharedvm"
      migrate="live" recovery="restart"/>
    <vm autostart="0" domain="domain2" exclusive="0"
      name="station183.example.com" path="/etc/sysconfig/sharedvm"
      migrate="live" recovery="restart"/>
  </rm>
</cluster>
```

To apply the new cluster configuration to the running cluster, perform the following command:

```
# ccs_tool update /etc/cluster/cluster.conf
```

Now start and permanently enable the cluster resource group manager:

```
# service rgmanager restart
# chkconfig rgmanager on
```

Enable the virtual guests:

```
# clusvcadm -e vm:station182.example.com -m station200.example.com
# clusvcadm -e vm:station183.example.com -m station200.example.com
```

Invoking `clustat` on either dom0 member should show both Xen guests as cluster



resources:

```
# clustat
Member Status: Quorate

Member Name                                ID    Status
-----
station200.example.com                    1 Online, Local,
rgmanager
station201.example.com                    2 Online, rgmanager
/dev/mpath/quorumpl                       0 Online, Quorum Disk

Service      Name                                Owner(Last) State
-----
vm:station182.example.com                station200.example.com started
vm:station183.example.com                station200.example.com started
```

In the above example both Xen guests have been started at the same node, *station200*.

## 8.4 Testing the cluster

Proper testing of the cluster consists of passing the following scenarios on each member:

- Test disable SAN / NFS connection
- Test disable Ethernet connection
- Test Xen guest Live migration

### 8.4.1 Test disable SAN / NFS connection

1. Test single SAN path removal to test dm-multipathing (transparent path fail-over should occur)
2. Test complete SAN removal or loss of NFS share access to test failover (cluster node should be fenced / virtual guest relocation should occur)

### 8.4.2 Test disable Ethernet connection

1. Test bonding fail-over if one interface is disabled (transparent bonding fail-over should occur)
2. Test cluster fail-over if both bonded interfaces are down (cluster node should get fenced / virtual guest relocation should occur)

### 8.4.3 Test Xen guest Live Migration

Test Live Migration by issuing a command similar to the following example:

```
# clusvcadm -M vm:station183.example.com -m station201.example.com
Trying to migrate vm:station183.example.com to
station201.example.com...Success
```

**Note:** Use always `clusvcadm` to perform Live Migration or enabling or disabling of a guest. Since `xm`, `virsh`, and `virt-manager` commands are not cluster aware, do not use them to perform these operations on guests that are cluster resources!



## 9 Backup & Restore

### 9.1 Backup

Two different backup strategies are possible:

- For backing up the dom0 hypervisor and the virtual guests, Red Hat recommends to use a filesystem based backup software like amanda or use the customer's currently implemented backup software.
- For the GFS based image files for virtual systems, Red Hat recommends to stop virtual guests and perform backup. It is not recommend to backup virtual guest file images while guests are running or paused.

### 9.2 Restore

For restoring virtual guest images to the cluster, perform the following steps:

1. Make sure the clustered service is shown as “*disabled*” in output of `clustat`; if this is not the case, use: `clusvcadm -d vm:stationXXX.example.com`
2. Recover the image file of the virtual guest to `/var/lib/xen/images/`
3. If necessary, recover `/etc/xen/<XEN_GUEST_NAME>` configuration file
4. Enable the virtual machine using: `clusvcadm -e vm:stationXXX.example.com`



## Appendix A – Sample kickstart file

A *kickstart* file can be used to install the cluster nodes. Below is an example, please modify to your specific needs:

```
install
# choose one of the following 2:
cdrom
url --url http://install.server.example.com/rhel5.2-i386 -DIRECTORY
# Installation key is commented, so the installer will ask for a valid one:
#key --skip
lang en_US.UTF-8
keyboard en_US
skipx
network --device eth0 --bootproto dhcp
rootpw --iscrypted $1$Bd5.N$FwrRk.ZWB2bDS0T0f2PPU1
firewall --disabled
authconfig --useshadow --enablemd5
selinux --disabled
timezone Europe/Berlin
bootloader --location=mbr --driveorder=sda --append="rhgb quiet"
clearpart --all --initlabel
part / --fstype ext3 --size=1024 --grow --ondisk=sda --asprimary
part swap --size=recommended --ondisk=sda --asprimary

%packages
@ base
sysstat
iscsi-initiator-utils
@ cluster-storage
kmod-gfs-PAE
kmod-gfs-xen
@ clustering
@ virtualization
@ X Window System
@ GNOME Desktop Environment
ntp
perl-Crypt-SSLeay
```



## Appendix B – Complete sample cluster.conf

```
<?xml version="1.0"?>
<cluster name="dom0cluster1" alias="dom0cluster1" config_version="1">
  <quorumd device="/dev/mpath/iscsip1" interval="3" min_score="1" tko="4"
votes="1">
    <heuristic interval="4" program="ping 192.168.10.254 -c3 -t1"
score="1"/>
  </quorumd>
  <fence_daemon clean_start="1" post_fail_delay="5" post_join_delay="20"/>
  <cman expected_votes="1" two_node="1"/>
  <clusternodes>
    <clusternode name="station200.example.com" nodeid="1" votes="1">
      <fence>
        <method name="1">
          <device name="fence_station200" port="1" switch="1"/>
        </method>
        <method name="2">
          <device name="manualfence" nodename="station200.example.com"/>
        </method>
      </fence>
    </clusternode>
    <clusternode name="station201" nodeid="2" votes="1">
      <fence>
        <method name="1">
          <device name="fence_station201" port="2" switch="1"/>
        </method>
        <method name="2">
          <device name="manualfence" nodename="station201.example.com"/>
        </method>
      </fence>
    </clusternode>
  </clusternodes>
  <fencedevices>
    <fencedevice agent="fence_apc" hostname="172.16.73.12"
login="fenceuser" name="fence_station200" passwd="PASSWORD"/>
    <fencedevice agent="fence_apc" hostname="172.16.73.13"
login="fenceuser" name="fence_station201" passwd="PASSWORD"/>
    <fencedevice agent="fence_manual" name="manualfence"/>
    <fencedevice agent="fence_xvm" name="fence_xen"/>
  </fencedevices>
  <rm>
    <failoverdomains>
      <failoverdomain name="domain1" ordered="1" restricted="1">
        <failoverdomainnode name="station200.example.com" priority="1"/>
        <failoverdomainnode name="station201.example.com" priority="2"/>
      </failoverdomain>
      <failoverdomain name="domain2" ordered="1" restricted="1">
        <failoverdomainnode name="station201.example.com" priority="1"/>
        <failoverdomainnode name="station200.example.com" priority="2"/>
      </failoverdomain>
    </failoverdomains>
  </resources/>
  <vm autostart="0" domain="domain1" exclusive="0"
name="station182.example.com" path="/etc/sysconfig/sharedvm"
migrate="live" recovery="restart"/>
  <vm autostart="0" domain="domain2" exclusive="0">
```



```
name="station183.example.com" path="/etc/sysconfig/sharedvm"
migrate="live" recovery="restart"/>
</rm>
</cluster>
```

## Appendix C – Conga Cluster Management

Red Hat Cluster Suite in Red Hat Enterprise Linux 5 provides Conga, a smart client / server infrastructure to install, configure and manage Red Hat Clusters. Conga consist out of two services:

1. `ricci` – the cluster node agent
2. `luci` – the management backend incl. a web front end

Conga can be used to remotely monitor and manage the status or actions of the virtual cluster.

If third node is available, install it the same way as the cluster nodes. Some notes regarding the `luci` management server:

- The management server running `luci` can be any Red Hat Enterprise Linux 5 system having TCP/IP access to the cluster nodes
- One `luci` management server can handle several clusters
- The management node does not need to be very strong (a small system with 512 MB of RAM is sufficient)
- It does not need to have SAN access

Make sure the service `ricci` is running on every cluster member:

```
# service ricci status
ricci (pid 7178) is running...
```

Next start `luci` on the management node:

```
# service luci start
```



Now point a web-browser to the `luci` frontend of the management server and login with the credentials.

The screenshot shows a Mozilla Firefox browser window with the title "luci - Mozilla Firefox". The address bar contains the URL "https://station202:8084/luci/acl\_users/credentials\_cookie\_auth/require\_login". The page header includes the Red Hat logo and the text "CLUSTER AND STORAGE SYSTEMS". Below the header, there is a "Please log in" section with a message: "To access this part of the site, you need to log in with your user name and password." The login form is titled "Account details" and contains the following fields and elements:

- Login Name:** A text input field containing "admin". A note below it says "Login names are case sensitive, make sure the caps lock key is not enabled."
- Password:** A password input field with masked characters ".....". A note below it says "Case sensitive, make sure caps lock is not enabled."
- Log in:** A button to submit the login form.

Below the login form, there is a message: "Please log out or exit your browser when you're done." At the bottom of the page, there is a footer with the text: "The Conga Cluster and Storage Management System is Copyright © 2000-2008 Red Hat, Inc. Distributed under the GNU GPL license." The browser's status bar at the bottom shows "Done" on the left and "station202:8084" on the right.



At *homebase* choose *Add an Existing Cluster* to start the import the running cluster:

The screenshot shows a Mozilla Firefox browser window with the title "homebase — luci - Mozilla Firefox". The address bar contains "https://station202:8084/luci/homebase". The page header features the Red Hat logo and the text "CLUSTER AND STORAGE SYSTEMS". Below the header, there are tabs for "homebase", "cluster", and "storage", and a "help log out" link. The main content area is titled "Luci Homebase" and includes a sidebar with the user "admin" and a list of actions: "Add a System", "Add an Existing Cluster", and "Add a User". The main text area says "Welcome to Luci, admin. Select an action from the list on the left." At the bottom, a footer contains the copyright information: "The Conga Cluster and Storage Management System is Copyright © 2000-2008 Red Hat, Inc. Distributed under the GNU GPL license."

https://station202:8084/luci/homebase?pagetype=7

station202:8084



Enter one of the dom0 cluster node names, the corresponding root password and select *Submit*:

The screenshot shows a web browser window titled "Luci — homebase — Add a running cluster to be managed by Luci - Mozilla Firefox". The address bar shows the URL "https://station202:8084/luci/homebase?pagetype=7". The page header includes the Red Hat logo and "CLUSTER AND STORAGE SYSTEMS". Below the header, there are navigation tabs for "homebase", "cluster", and "storage". A sidebar on the left shows a user menu for "admin" with options: "Add a System", "Add an Existing Cluster", and "Add a User". The main content area is titled "Add an Existing Cluster" and contains the instruction: "Enter one node from the cluster you wish to add to the Luci management interface." Below this, there is a form with three input fields: "System Hostname" (containing "station200.example.com"), "Root Password" (masked with "\*\*\*\*\*"), and "Key ID" (with a question mark icon). There is also a "View SSL cert fingerprints" link and a "Submit" button. At the bottom of the page, a footer contains the text: "The Conga Cluster and Storage Management System is Copyright © 2000-2008 Red Hat, Inc. Distributed under the GNU GPL license."

Done

station202:8084



The other dom0 cluster node should be listed. Add the root password for the second node as well and press *Submit* again:

Luci — homebase — Add a running cluster to be managed by Luci - Mozilla Firefox

File Edit View History Bookmarks Tools Help

https://station202:8084/luci/homebase?pagetype=7

Google

redhat CLUSTER AND STORAGE SYSTEMS

homebase cluster storage help log out

admin

- Add a System
- Add an Existing Cluster
- Add a User

### Add an Existing Cluster

Enter one node from the cluster you wish to add to the Luci management interface.

**Cluster Name: dom0\_cluster1**

System Hostname	Root Password	Key ID
station200.example.com	*****	
station201.example.com	*****	

Check if system passwords are identical.

[View SSL cert fingerprints](#)

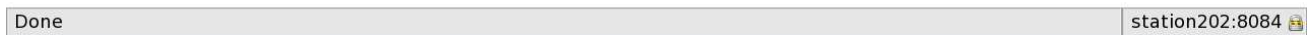
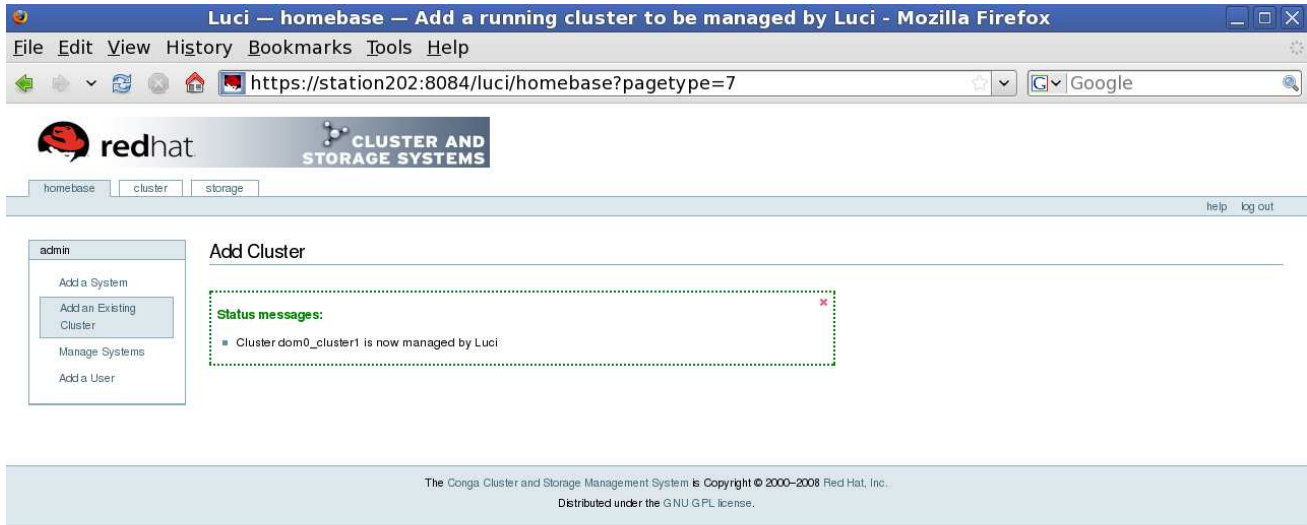
The Conga Cluster and Storage Management System is Copyright © 2000-2008 Red Hat, Inc.  
Distributed under the GNU GPL license.

Done

station202:8084



luci should respond with a successful message regarding that the cluster has now been added to luci management:



The user will now be able to browse through luci's web front end to explore the cluster and managed resources.