

# SUSE Linux Enterprise High Availability Extension

11 SP4

[www.suse.com](http://www.suse.com)

Jun 03 2015

High Availability Guide



# High Availability Guide

**List of Authors:** Tanja Roth, Thomas Schraitle

Copyright © 2006–2015 SUSE LLC and contributors. All rights reserved.

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or (at your option) version 1.3; with the Invariant Section being this copyright notice and license. A copy of the license version 1.2 is included in the section entitled “GNU Free Documentation License”.

For SUSE or Novell trademarks, see the Novell Trademark and Service Mark list <http://www.novell.com/company/legal/trademarks/tmlist.html>. All other third party trademarks are the property of their respective owners. A trademark symbol (®, <sup>TM</sup> etc.) denotes a SUSE or Novell trademark; an asterisk (\*) denotes a third party trademark.

All information found in this book has been compiled with utmost attention to detail. However, this does not guarantee complete accuracy. Neither SUSE LLC, its affiliates, the authors nor the translators shall be held liable for possible errors or the consequences thereof.

# Contents

## About This Guide xi

1 Feedback .....	xii
2 Documentation Conventions .....	xii
3 About the Making of This Manual .....	xiii

## I Installation and Setup 1

### 1 Product Overview 3

1.1 Availability as Add-On/Extension .....	4
1.2 Key Features .....	4
1.3 Benefits .....	8
1.4 Cluster Configurations: Storage .....	11
1.5 Architecture .....	14

### 2 System Requirements and Recommendations 19

2.1 Hardware Requirements .....	19
2.2 Software Requirements .....	20
2.3 Storage Requirements .....	20
2.4 Other Requirements and Recommendations .....	21

### 3 Installation and Basic Setup 25

3.1 Definition of Terms .....	25
3.2 Overview .....	28

3.3 Installation as Add-on .....	29
3.4 Automatic Cluster Setup (sleha-bootstrap) .....	30
3.5 Manual Cluster Setup (YaST) .....	35
3.6 Mass Deployment with AutoYaST .....	49

## **II Configuration and Administration 51**

### **4 Configuration and Administration Basics 53**

4.1 Global Cluster Options .....	53
4.2 Cluster Resources .....	56
4.3 Resource Monitoring .....	72
4.4 Resource Constraints .....	73
4.5 Managing Services on Remote Hosts .....	84
4.6 Monitoring System Health .....	87
4.7 Maintenance Mode .....	88
4.8 For More Information .....	89

### **5 Configuring and Managing Cluster Resources (Web Interface) 91**

5.1 Hawk—Overview .....	91
5.2 Configuring Global Cluster Options .....	98
5.3 Configuring Cluster Resources .....	99
5.4 Managing Cluster Resources .....	118
5.5 Monitoring Multiple Clusters .....	131
5.6 Hawk for Geo Clusters .....	134
5.7 Troubleshooting .....	135

### **6 Configuring and Managing Cluster Resources (GUI) 137**

6.1 Pacemaker GUI—Overview .....	138
----------------------------------	-----

6.2 Configuring Global Cluster Options .....	141
6.3 Configuring Cluster Resources .....	142
6.4 Managing Cluster Resources .....	163

## **7 Configuring and Managing Cluster Resources (Command Line) 169**

7.1 crmsh—Overview .....	170
7.2 Configuring Global Cluster Options .....	177
7.3 Configuring Cluster Resources .....	178
7.4 Managing Cluster Resources .....	190
7.5 Setting Passwords Independent of <code>cib.xml</code> .....	194
7.6 Retrieving History Information .....	194
7.7 For More Information .....	196

## **8 Adding or Modifying Resource Agents 197**

8.1 STONITH Agents .....	197
8.2 Writing OCF Resource Agents .....	198
8.3 OCF Return Codes and Failure Recovery .....	199

## **9 Fencing and STONITH 203**

9.1 Classes of Fencing .....	203
9.2 Node Level Fencing .....	204
9.3 STONITH Resources and Configuration .....	206
9.4 Monitoring Fencing Devices .....	209
9.5 Special Fencing Devices .....	210
9.6 Basic Recommendations .....	211
9.7 For More Information .....	212

## **10 Access Control Lists 215**

10.1 Requirements and Prerequisites .....	216
---	-----

10.2 Enabling Use of ACLs In Your Cluster .....	216
10.3 The Basics of ACLs .....	217
10.4 Configuring ACLs with the Pacemaker GUI .....	220
10.5 Configuring ACLs with Hawk .....	222
10.6 Configuring ACLs with crmsh .....	223
<b>11 Network Device Bonding</b>	<b>225</b>
11.1 Configuring Bonding Devices with YaST .....	225
11.2 Hotplugging of Bonding Slaves .....	228
11.3 For More Information .....	230
<b>12 Load Balancing with Linux Virtual Server</b>	<b>231</b>
12.1 Conceptual Overview .....	231
12.2 Configuring IP Load Balancing with YaST .....	234
12.3 Further Setup .....	239
12.4 For More Information .....	239
<b>13 Geo Clusters (Multi-Site Clusters)</b>	<b>241</b>
<b>III Storage and Data Replication</b>	<b>243</b>
<b>14 OCFS2</b>	<b>245</b>
14.1 Features and Benefits .....	245
14.2 OCFS2 Packages and Management Utilities .....	246
14.3 Configuring OCFS2 Services and a STONITH Resource .....	247
14.4 Creating OCFS2 Volumes .....	249
14.5 Mounting OCFS2 Volumes .....	252
14.6 Using Quotas on OCFS2 File Systems .....	253
14.7 For More Information .....	254

## **15 DRBD 255**

15.1 Conceptual Overview .....	255
15.2 Installing DRBD Services .....	257
15.3 Configuring the DRBD Service .....	258
15.4 Testing the DRBD Service .....	264
15.5 Tuning DRBD .....	266
15.6 Troubleshooting DRBD .....	267
15.7 For More Information .....	268

## **16 Cluster Logical Volume Manager (cLVM) 271**

16.1 Conceptual Overview .....	271
16.2 Configuration of cLVM .....	272
16.3 Configuring Eligible LVM2 Devices Explicitly .....	282
16.4 For More Information .....	283

## **17 Storage Protection 285**

17.1 Storage-based Fencing .....	286
17.2 Ensuring Exclusive Storage Activation .....	294
17.3 For More Information .....	296

## **18 Samba Clustering 297**

18.1 Conceptual Overview .....	297
18.2 Basic Configuration .....	299
18.3 Joining an Active Directory Domain .....	301
18.4 Debugging and Testing Clustered Samba .....	303
18.5 For More Information .....	304

## **19 Disaster Recovery with Rear 307**

19.1 Conceptual Overview .....	307
19.2 Setting Up Rear and Your Backup Solution .....	311

19.3 Creating the Recovery Installation System .....	313
19.4 Testing the Recovery Process .....	313
19.5 Recovering from Disaster .....	314
19.6 For More Information .....	315

## **IV Appendix 317**

### **20 Troubleshooting 319**

20.1 Installation and First Steps .....	319
20.2 Logging .....	320
20.3 Resources .....	323
20.4 STONITH and Fencing .....	324
20.5 Miscellaneous .....	325
20.6 For More Information .....	326

### **A Naming Conventions 327**

### **B Example of Setting Up a Simple Testing Resource 329**

B.1 Configuring a Resource with the GUI .....	329
B.2 Configuring a Resource Manually .....	331

### **C Example Configuration for OCFS2 and cLVM 333**

### **D Cluster Management Tools 335**

### **E Upgrading Your Cluster and Updating Software Packages 339**

E.1 Terminology .....	339
E.2 Upgrading your Cluster to the Latest Product Version .....	340
E.3 Updating Software Packages on Cluster Nodes .....	348



E.4 For More Information .....	349
--------------------------------	-----

<b>F What's New?</b>	<b>351</b>
----------------------	------------

F.1 Version 10 SP3 to Version 11 .....	351
--	-----

F.2 Version 11 to Version 11 SP1 .....	354
--	-----

F.3 Version 11 SP1 to Version 11 SP2 .....	356
--	-----

F.4 Version 11 SP2 to Version 11 SP3 .....	358
--	-----

F.5 Version 11 SP3 to Version 11 SP4 .....	360
--	-----

<b>Terminology .....</b>	<b>365</b>
--------------------------	------------

<b>G GNU Licenses</b>	<b>373</b>
-----------------------	------------

G.1 GNU Free Documentation License .....	373
--	-----



# About This Guide

SUSE® Linux Enterprise High Availability Extension is an integrated suite of open source clustering technologies that enables you to implement highly available physical and virtual Linux clusters. For quick and efficient configuration and administration, the High Availability Extension includes both a graphical user interface (GUI) and a command line interface (CLI). Additionally, it comes with the HA Web Konsole (Hawk), allowing you to administer your Linux cluster also via a Web interface.

This guide is intended for administrators who need to set up, configure, and maintain High Availability (HA) clusters. Both approaches (GUI and CLI) are covered in detail to help the administrators choose the appropriate tool that matches their needs for performing the key tasks.

This guide is divided into the following parts:

## Installation and Setup

Before starting to install and configure your cluster, make yourself familiar with cluster fundamentals and architecture, get an overview of the key features and benefits. Learn which hardware and software requirements must be met and what preparations to take before executing the next steps. Perform the installation and basic setup of your HA cluster using YaST.

## Configuration and Administration

Add, configure and manage cluster resources, using either the graphical user interface (Pacemaker GUI), the Web interface (HA Web Konsole), or the command line interface (`crm` shell). To avoid unauthorized access to the cluster configuration, define roles and assign them to certain users for fine-grained control. Learn how to make use of load balancing and fencing. In case you consider writing your own resource agents or modifying existing ones, get some background information on how to create different types of resource agents.

## Storage and Data Replication

SUSE Linux Enterprise High Availability Extension ships with a cluster-aware file system and volume manager: OCFS2 and the clustered Logical Volume Manager (cLVM). For replication of your data, use DRBD\* to mirror the data of a High Availability service from the active node of a cluster to its standby node. Furthermore, a clustered Samba server also provides a High Availability solution for heterogeneous environments.

## Appendix

Lists the new features and behavior changes of the High Availability Extension since the last release. Learn how to migrate your cluster to the most recent release version and find an example of setting up a simple testing resource.

Many chapters in this manual contain links to additional documentation resources. These include additional documentation that is available on the system as well as documentation available on the Internet.

For an overview of the documentation available for your product and the latest documentation updates, refer to <http://www.suse.com/doc/>.

# 1 Feedback

Several feedback channels are available:

## Bugs and Enhancement Requests

For services and support options available for your product, refer to <http://www.suse.com/support/>.

To report bugs for a product component, log into the Novell Customer Center from <http://www.suse.com/support/> and select *My Support > Service Request*.

## User Comments

We want to hear your comments about and suggestions for this manual and the other documentation included with this product. Use the User Comments feature at the bottom of each page in the online documentation or go to <http://www.suse.com/doc/feedback.html> and enter your comments there.

## Mail

For feedback on the documentation of this product, you can also send a mail to [doc-team@suse.de](mailto:doc-team@suse.de). Make sure to include the document title, the product version, and the publication date of the documentation. To report errors or suggest enhancements, provide a concise description of the problem and refer to the respective section number and page (or URL).

# 2 Documentation Conventions

The following typographical conventions are used in this manual:

- `/etc/passwd`: directory names and filenames
- *placeholder*: replace *placeholder* with the actual value
- `PATH`: the environment variable `PATH`
- `ls, --help`: commands, options, and parameters
- `user`: users or groups
- `Alt, Alt + F1`: a key to press or a key combination; keys are shown in uppercase as on a keyboard
- *File, File > Save As*: menu items, buttons
- ¶**amd64 em64t**¶: This paragraph is only relevant for the architectures `amd64`, `em64t`, and `ipf`. The arrows mark the beginning and the end of the text block. ¶
- *Dancing Penguins* (Chapter *Penguins*, ↑Another Manual): This is a reference to a chapter in another manual.

## 3 About the Making of This Manual

This book is written in Novdoc, a subset of DocBook (see <http://www.docbook.org>). The XML source files were validated by `xmllint`, processed by `xsltproc`, and converted into XSL-FO using a customized version of Norman Walsh's stylesheets. The final PDF is formatted through XEP from RenderX.



# **Part I. Installation and Setup**





# Product Overview

SUSE® Linux Enterprise High Availability Extension is an integrated suite of open source clustering technologies that enables you to implement highly available physical and virtual Linux clusters, and to eliminate single points of failure. It ensures the high availability and manageability of critical network resources including data, applications, and services. Thus, it helps you maintain business continuity, protect data integrity, and reduce unplanned downtime for your mission-critical Linux workloads.

It ships with essential monitoring, messaging, and cluster resource management functionality (supporting failover, failback, and migration (load balancing) of individually managed cluster resources).

This chapter introduces the main product features and benefits of the High Availability Extension. Inside you will find several example clusters and learn about the components making up a cluster. The last section provides an overview of the architecture, describing the individual architecture layers and processes within the cluster.

For explanations of some common terms used in the context of High Availability clusters, refer to Terminology (page 365).

# 1.1 Availability as Add-On/ Extension

The High Availability Extension is available as an add-on to SUSE Linux Enterprise Server 11 SP4. Support for geographically dispersed clusters (Geo clusters) is available as a separate extension to the High Availability Extension, called Geo Clustering for SUSE Linux Enterprise High Availability Extension.

## 1.2 Key Features

SUSE® Linux Enterprise High Availability Extension helps you ensure and manage the availability of your network resources. The following sections highlight some of the key features:

### 1.2.1 Wide Range of Clustering Scenarios

The High Availability Extension supports the following scenarios:

- Active/active configurations
- Active/passive configurations: N+1, N+M, N to 1, N to M
- Hybrid physical and virtual clusters, allowing virtual servers to be clustered with physical servers. This improves service availability and resource usage.
- Local clusters
- Metro clusters (“stretched” local clusters)
- Geo clusters (geographically dispersed clusters)

Your cluster can contain up to 32 Linux servers. Any server in the cluster can restart resources (applications, services, IP addresses, and file systems) from a failed server in the cluster.

## 1.2.2 Flexibility

The High Availability Extension ships with Corosync/OpenAIS messaging and membership layer and Pacemaker Cluster Resource Manager. Using Pacemaker, administrators can continually monitor the health and status of their resources, manage dependencies, and automatically stop and start services based on highly configurable rules and policies. The High Availability Extension allows you to tailor a cluster to the specific applications and hardware infrastructure that fit your organization. Time-dependent configuration enables services to automatically migrate back to repaired nodes at specified times.

## 1.2.3 Storage and Data Replication

With the High Availability Extension you can dynamically assign and reassign server storage as needed. It supports Fibre Channel or iSCSI storage area networks (SANs). Shared disk systems are also supported, but they are not a requirement. SUSE Linux Enterprise High Availability Extension also comes with a cluster-aware file system and volume manager (OCFS2) and the clustered Logical Volume Manager (cLVM). For replication of your data, you can use DRBD\* to mirror the data of a High Availability service from the active node of a cluster to its standby node. Furthermore, SUSE Linux Enterprise High Availability Extension also supports CTDB (Clustered Trivial Database), a technology for Samba clustering.

## 1.2.4 Support for Virtualized Environments

SUSE Linux Enterprise High Availability Extension supports the mixed clustering of both physical and virtual Linux servers. SUSE Linux Enterprise Server 11 SP4 ships with Xen, an open source virtualization hypervisor and with KVM (Kernel-based Virtual Machine), a virtualization software for Linux which is based on hardware virtualization extensions. The cluster resource manager in the High Availability Extension can recognize, monitor, and manage services running within virtual servers, as well as services running in physical servers. Guest systems can be managed as services by the cluster.

## 1.2.5 Support of Local, Metro, and Geo Clusters

SUSE Linux Enterprise High Availability Extension has been extended to support different geographical scenarios. Support for geographically dispersed clusters (Geo clusters) is available as a separate extension to High Availability Extension, called Geo Clustering for SUSE Linux Enterprise High Availability Extension.

### Local Clusters

A single cluster in one location (for example, all nodes are located in one data center). The cluster uses multicast or unicast for communication between the nodes and manages failover internally. Network latency can be neglected. Storage is typically accessed synchronously by all nodes.

### Metro Clusters

A single cluster that can stretch over multiple buildings or data centers, with all sites connected by fibre channel. The cluster uses multicast or unicast for communication between the nodes and manages failover internally. Network latency is usually low (<5 ms for distances of approximately 20 miles). Storage is frequently replicated (mirroring or synchronous replication).

### Geo Clusters (Multi-Site Clusters)

Multiple, geographically dispersed sites with a local cluster each. The sites communicate via IP. Failover across the sites is coordinated by a higher-level entity. Geo clusters need to cope with limited network bandwidth and high latency. Storage is replicated asynchronously.

The greater the geographical distance between individual cluster nodes, the more factors may potentially disturb the high availability of services the cluster provides. Network latency, limited bandwidth and access to storage are the main challenges for long-distance clusters.

## 1.2.6 Resource Agents

SUSE Linux Enterprise High Availability Extension includes a huge number of resource agents to manage resources such as Apache, IPv4, IPv6 and many more. It also ships with resource agents for popular third party applications such as IBM WebSphere Application Server. For an overview of Open Cluster Framework

(OCF) resource agents included with your product, use the `crm ra` command as described in Section 7.1.3, “Displaying Information about OCF Resource Agents” (page 172).

## 1.2.7 User-friendly Administration Tools

The High Availability Extension ships with a set of powerful tools for basic installation and setup of your cluster as well as effective configuration and administration:

### YaST

A graphical user interface for general system installation and administration. Use it to install the High Availability Extension on top of SUSE Linux Enterprise Server as described in Section 3.3, “Installation as Add-on” (page 29). YaST also provides the following modules in the High Availability category to help configure your cluster or individual components:

- **Cluster: Basic cluster setup.** For details, refer to Section 3.5, “Manual Cluster Setup (YaST)” (page 35).
- **DRBD: Configuration of a Distributed Replicated Block Device.**
- **IP Load Balancing: Configuration of load balancing with Linux Virtual Server.** For details, refer to Chapter 12, *Load Balancing with Linux Virtual Server* (page 231).

### Pacemaker GUI

Installable graphical user interface for easy configuration and administration of your cluster. Guides you through the creation and configuration of resources and lets you execute management tasks like starting, stopping or migrating resources. For details, refer to Chapter 6, *Configuring and Managing Cluster Resources (GUI)* (page 137).

### HA Web Konsole (Hawk)

A Web-based user interface with which you can administer your Linux cluster from non-Linux machines. It is also an ideal solution in case your system does not provide a graphical user interface. It guides you through the creation and configuration of resources and lets you execute management tasks like starting, stopping or migrating resources. For details, refer to Chapter 5, *Configuring and Managing Cluster Resources (Web Interface)* (page 91).

## crm Shell

A powerful unified command line interface to configure resources and execute all monitoring or administration tasks. For details, refer to Chapter 7, *Configuring and Managing Cluster Resources (Command Line)* (page 169).

# 1.3 Benefits

The High Availability Extension allows you to configure up to 32 Linux servers into a high-availability cluster (HA cluster), where resources can be dynamically switched or moved to any server in the cluster. Resources can be configured to automatically migrate if a server fails, or they can be moved manually to troubleshoot hardware or balance the workload.

The High Availability Extension provides high availability from commodity components. Lower costs are obtained through the consolidation of applications and operations onto a cluster. The High Availability Extension also allows you to centrally manage the complete cluster and to adjust resources to meet changing workload requirements (thus, manually “load balance” the cluster). Allowing clusters of more than two nodes also provides savings by allowing several nodes to share a “hot spare”.

An equally important benefit is the potential reduction of unplanned service outages as well as planned outages for software and hardware maintenance and upgrades.

Reasons that you would want to implement a cluster include:

- Increased availability
- Improved performance
- Low cost of operation
- Scalability
- Disaster recovery
- Data protection
- Server consolidation

- Storage consolidation

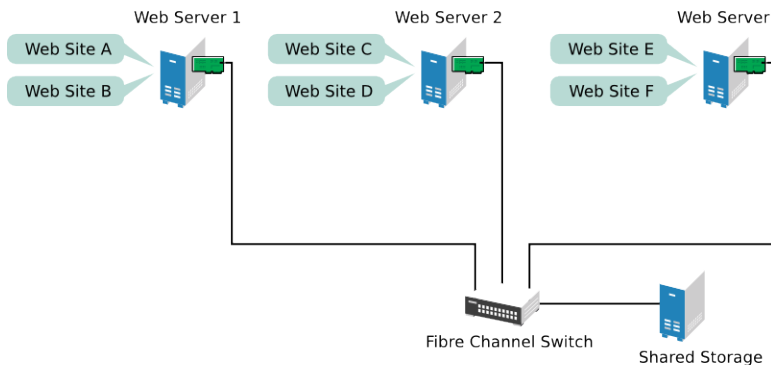
Shared disk fault tolerance can be obtained by implementing RAID on the shared disk subsystem.

The following scenario illustrates some benefits the High Availability Extension can provide.

## Example Cluster Scenario

Suppose you have configured a three-server cluster, with a Web server installed on each of the three servers in the cluster. Each of the servers in the cluster hosts two Web sites. All the data, graphics, and Web page content for each Web site are stored on a shared disk subsystem connected to each of the servers in the cluster. The following figure depicts how this setup might look.

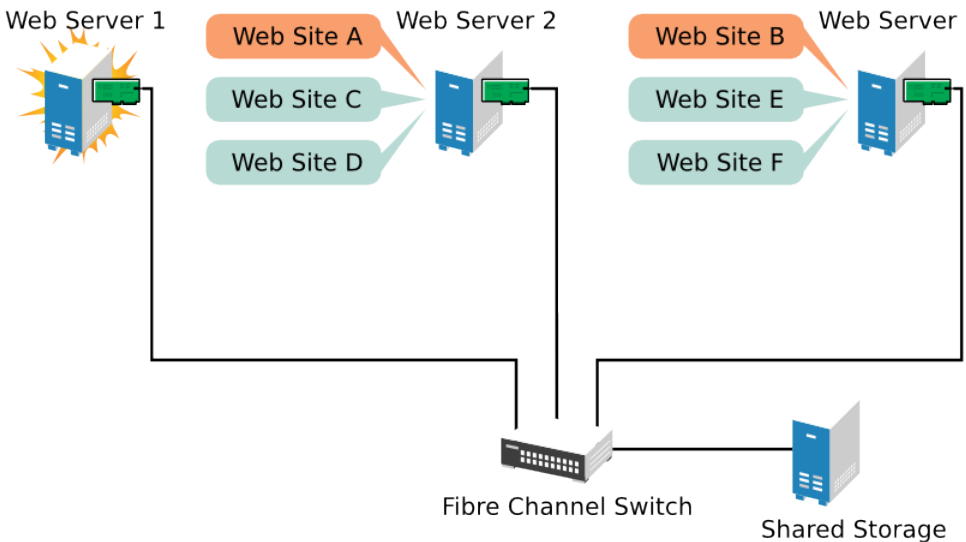
**Figure 1.1:** *Three-Server Cluster*



During normal cluster operation, each server is in constant communication with the other servers in the cluster and performs periodic polling of all registered resources to detect failure.

Suppose Web Server 1 experiences hardware or software problems and the users depending on Web Server 1 for Internet access, e-mail, and information lose their connections. The following figure shows how resources are moved when Web Server 1 fails.

**Figure 1.2:** *Three-Server Cluster after One Server Fails*



Web Site A moves to Web Server 2 and Web Site B moves to Web Server 3. IP addresses and certificates also move to Web Server 2 and Web Server 3.

When you configured the cluster, you decided where the Web sites hosted on each Web server would go should a failure occur. In the previous example, you configured Web Site A to move to Web Server 2 and Web Site B to move to Web Server 3. This way, the workload once handled by Web Server 1 continues to be available and is evenly distributed between any surviving cluster members.

When Web Server 1 failed, the High Availability Extension software did the following:

- Detected a failure and verified with STONITH that Web Server 1 was really dead. STONITH is an acronym for “Shoot The Other Node In The Head” and is a means of bringing down misbehaving nodes to prevent them from causing trouble in the cluster.
- Remounted the shared data directories that were formerly mounted on Web server 1 on Web Server 2 and Web Server 3.
- Restarted applications that were running on Web Server 1 on Web Server 2 and Web Server 3.



- Transferred IP addresses to Web Server 2 and Web Server 3.

In this example, the failover process happened quickly and users regained access to Web site information within seconds, usually without needing to log in again.

Now suppose the problems with Web Server 1 are resolved, and Web Server 1 is returned to a normal operating state. Web Site A and Web Site B can either automatically fail back (move back) to Web Server 1, or they can stay where they are. This depends on how you configured the resources for them. Migrating the services back to Web Server 1 will incur some down-time, so the High Availability Extension also allows you to defer the migration until a period when it will cause little or no service interruption. There are advantages and disadvantages to both alternatives.

The High Availability Extension also provides resource migration capabilities. You can move applications, Web sites, etc. to other servers in your cluster as required for system management.

For example, you could have manually moved Web Site A or Web Site B from Web Server 1 to either of the other servers in the cluster. Use cases for this are upgrading or performing scheduled maintenance on Web Server 1, or increasing performance or accessibility of the Web sites.

## 1.4 Cluster Configurations: Storage

Cluster configurations with the High Availability Extension might or might not include a shared disk subsystem. The shared disk subsystem can be connected via high-speed Fibre Channel cards, cables, and switches, or it can be configured to use iSCSI. If a server fails, another designated server in the cluster automatically mounts the shared disk directories that were previously mounted on the failed server. This gives network users continuous access to the directories on the shared disk subsystem.

---

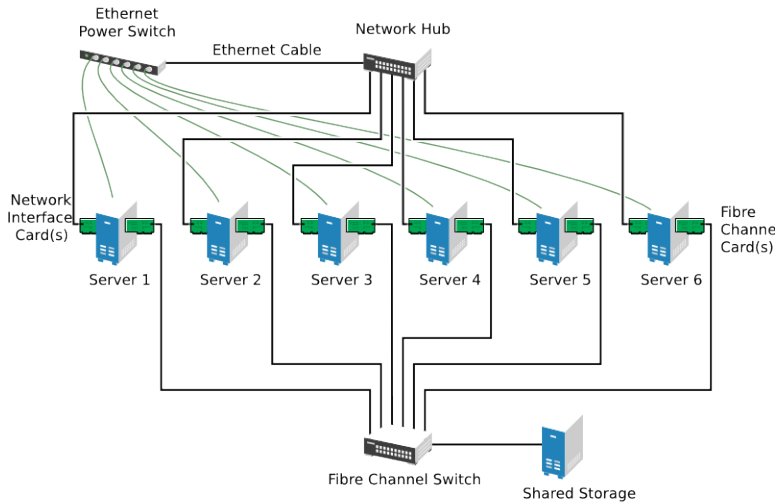
### **IMPORTANT: Shared Disk Subsystem with cLVM**

When using a shared disk subsystem with cLVM, that subsystem must be connected to all servers in the cluster from which it needs to be accessed.

---

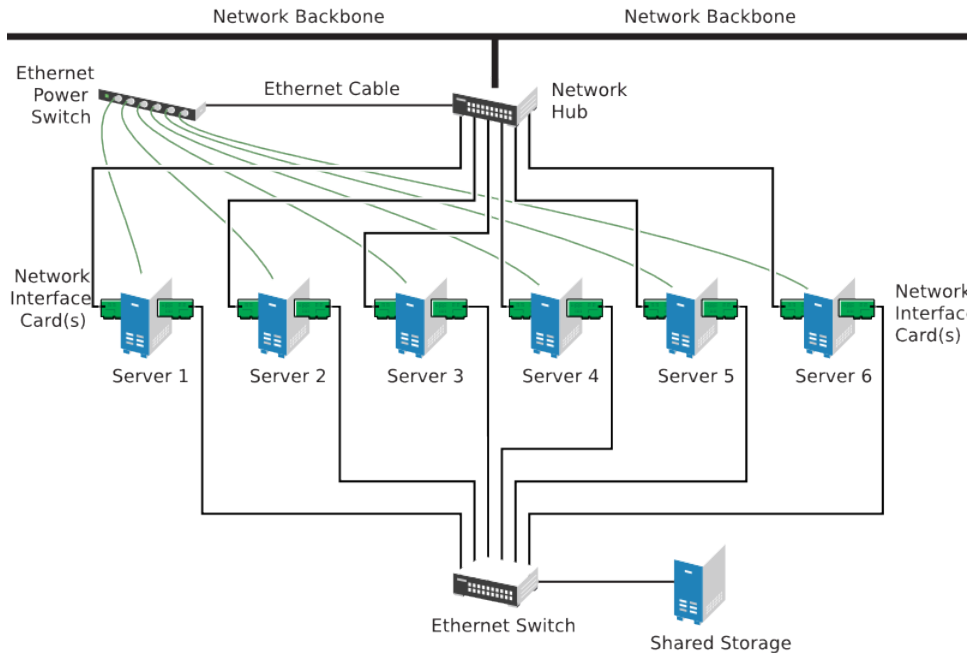
Typical resources might include data, applications, and services. The following figure shows how a typical Fibre Channel cluster configuration might look.

**Figure 1.3:** *Typical Fibre Channel Cluster Configuration*



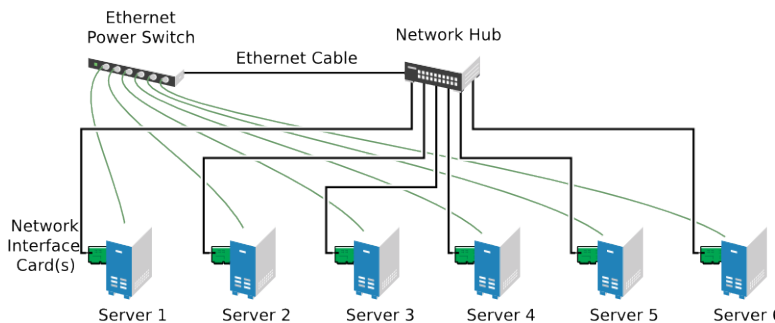
Although Fibre Channel provides the best performance, you can also configure your cluster to use iSCSI. iSCSI is an alternative to Fibre Channel that can be used to create a low-cost Storage Area Network (SAN). The following figure shows how a typical iSCSI cluster configuration might look.

**Figure 1.4:** *Typical iSCSI Cluster Configuration*



Although most clusters include a shared disk subsystem, it is also possible to create a cluster without a shared disk subsystem. The following figure shows how a cluster without a shared disk subsystem might look.

**Figure 1.5:** *Typical Cluster Configuration Without Shared Storage*



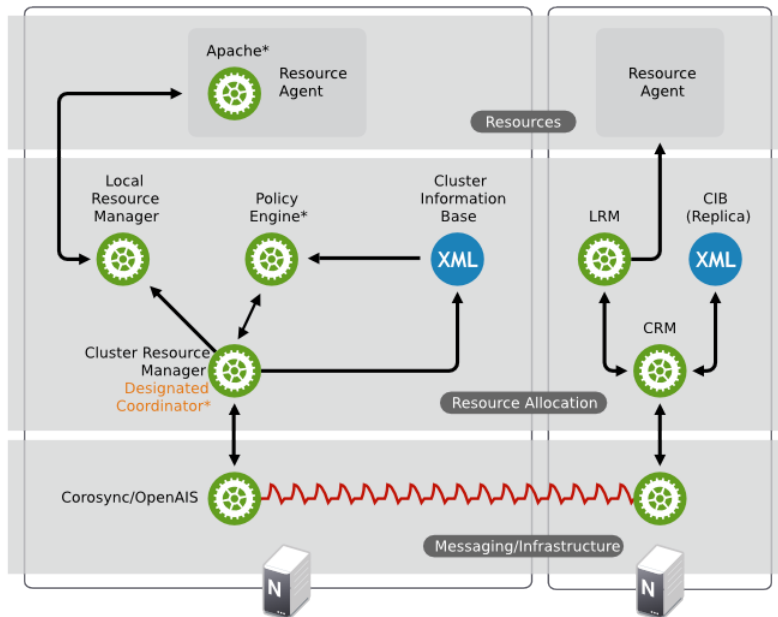
# 1.5 Architecture

This section provides a brief overview of the High Availability Extension architecture. It identifies and provides information on the architectural components, and describes how those components interoperate.

## 1.5.1 Architecture Layers

The High Availability Extension has a layered architecture. Figure 1.6, “Architecture” (page 14) illustrates the different layers and their associated components.

**Figure 1.6:** *Architecture*



## 1.5.1.1 Messaging and Infrastructure Layer

The primary or first layer is the messaging/infrastructure layer, also known as the Corosync/OpenAIS layer. This layer contains components that send out the messages containing “I’m alive” signals, as well as other information. The program of the High Availability Extension resides in the messaging/infrastructure layer.

## 1.5.1.2 Resource Allocation Layer

The next layer is the resource allocation layer. This layer is the most complex, and consists of the following components:

### Cluster Resource Manager (CRM)

Every action taken in the resource allocation layer passes through the Cluster Resource Manager. If other components of the resource allocation layer (or components which are in a higher layer) need to communicate, they do so through the local CRM. On every node, the CRM maintains the “Cluster Information Base (CIB)”.

### Cluster Information Base (CIB)

The Cluster Information Base is an in-memory XML representation of the entire cluster configuration and current status. It contains definitions of all cluster options, nodes, resources, constraints and the relationship to each other. The CIB also synchronizes updates to all cluster nodes. There is one master CIB in the cluster, maintained by the “Designated Coordinator (DC)”. All other nodes contain a CIB replica.

### Designated Coordinator (DC)

One CRM in the cluster is elected as DC. The DC is the only entity in the cluster that can decide that a cluster-wide change needs to be performed, such as fencing a node or moving resources around. The DC is also the node where the master copy of the CIB is kept. All other nodes get their configuration and resource allocation information from the current DC. The DC is elected from all nodes in the cluster after a membership change.

### Policy Engine (PE)

Whenever the Designated Coordinator needs to make a cluster-wide change (react to a new CIB), the Policy Engine calculates the next state of the cluster based on the current state and configuration. The PE also produces a transition

graph containing a list of (resource) actions and dependencies to achieve the next cluster state. The PE always runs on the DC.

#### Local Resource Manager (LRM)

The LRM calls the local Resource Agents (see Section 1.5.1.3, “Resource Layer” (page 16)) on behalf of the CRM. It can thus perform start / stop / monitor operations and report the result to the CRM. The LRM is the authoritative source for all resource-related information on its local node.

### 1.5.1.3 Resource Layer

The highest layer is the Resource Layer. The Resource Layer includes one or more Resource Agents (RA). Resource Agents are programs (usually shell scripts) that have been written to start, stop, and monitor a certain kind of service (a resource). Resource Agents are called only by the LRM. Third parties can include their own agents in a defined location in the file system and thus provide out-of-the-box cluster integration for their own software.

## 1.5.2 Process Flow

SUSE Linux Enterprise High Availability Extension uses Pacemaker as CRM. The CRM is implemented as daemon (`crmd`) that has an instance on each cluster node. Pacemaker centralizes all cluster decision-making by electing one of the `crmd` instances to act as a master. Should the elected `crmd` process (or the node it is on) fail, a new one is established.

A CIB, reflecting the cluster’s configuration and current state of all resources in the cluster is kept on each node. The contents of the CIB are automatically kept synchronous across the entire cluster.

Many actions performed in the cluster will cause a cluster-wide change. These actions can include things like adding or removing a cluster resource or changing resource constraints. It is important to understand what happens in the cluster when you perform such an action.

For example, suppose you want to add a cluster IP address resource. To do this, you can use one of the command line tools or the Web interface to modify the CIB. It is not required to perform the actions on the DC, you can use either tool on any node in the cluster and they will be relayed to the DC. The DC will then replicate the CIB change to all cluster nodes.

Based on the information in the CIB, the PE then computes the ideal state of the cluster and how it should be achieved and feeds a list of instructions to the DC. The DC sends commands via the messaging/infrastructure layer which are received by the crmd peers on other nodes. Each crmd uses its LRM (implemented as lrmd) to perform resource modifications. The lrmd is not cluster-aware and interacts directly with resource agents (scripts).

All peer nodes report the results of their operations back to the DC. After the DC concludes that all necessary operations are successfully performed in the cluster, the cluster will go back to the idle state and wait for further events. If any operation was not carried out as planned, the PE is invoked again with the new information recorded in the CIB.

In some cases, it may be necessary to power off nodes to protect shared data or complete resource recovery. For this Pacemaker comes with a fencing subsystem, stonithd. STONITH is an acronym for “Shoot The Other Node In The Head” and is usually implemented with a remote power switch. In Pacemaker, STONITH devices are modeled as resources (and configured in the CIB) to enable them to be easily monitored for failure. However, stonithd takes care of understanding the STONITH topology such that its clients simply request a node be fenced and it does the rest.





# System Requirements and Recommendations

The following section informs you about system requirements, and some prerequisites for SUSE® Linux Enterprise High Availability Extension. It also includes recommendations for cluster setup.

## 2.1 Hardware Requirements

The following list specifies hardware requirements for a cluster based on SUSE® Linux Enterprise High Availability Extension. These requirements represent the minimum hardware configuration. Additional hardware might be necessary, depending on how you intend to use your cluster.

- 1 to 32 Linux servers with software as specified in Section 2.2, “Software Requirements” (page 20). The servers do not require identical hardware (memory, disk space, etc.), but they must have the same architecture. Cross-platform clusters are not supported.
- At least two TCP/IP communication media per cluster node. Cluster nodes use multicast or unicast for communication so the network equipment must support the communication means you want to use. The communication media should support a data rate of 100 Mbit/s or higher. Preferably, the Ethernet channels should be bonded as described in Chapter 11, *Network Device Bonding* (page 225). Alternatively, use the second interface for a redundant communication channel in Corosync. See also Procedure 3.7, “Defining a Redundant Communication Channel” (page 39).

- Optional: A shared disk subsystem connected to all servers in the cluster from where it needs to be accessed. See Section 2.3, “Storage Requirements” (page 20).
- A STONITH mechanism. A STONITH device is a power switch which the cluster uses to reset nodes that are thought to be dead or behaving in a strange manner. This is the only reliable way to ensure that no data corruption is performed by nodes that hang and only appear to be dead.

## 2.2 Software Requirements

Ensure that the following software requirements are met:

- SUSE® Linux Enterprise Server 11 SP4 (with all available online updates) is installed on all nodes that will be part of the cluster.
- SUSE Linux Enterprise High Availability Extension 11 SP4 (with all available online updates) is installed on all nodes that will be part of the cluster.
- If you want to use Geo clusters, make sure that Geo Clustering for SUSE Linux Enterprise High Availability Extension 11 SP4 (with all available online updates) is installed on all nodes that will be part of the cluster.

## 2.3 Storage Requirements

To make data highly available, a shared disk system (Storage Area Network, or SAN) is recommended for your cluster. If a shared disk subsystem is used, ensure the following:

- The shared disk system is properly set up and functional according to the manufacturer’s instructions.
- The disks contained in the shared disk system should be configured to use mirroring or RAID to add fault tolerance to the shared disk system. Hardware-based RAID is recommended. Host-based software RAID is not supported for all configurations.
- If you are using iSCSI for shared disk system access, ensure that you have properly configured iSCSI initiators and targets.

- When using DRBD\* to implement a mirroring RAID system that distributes data across two machines, make sure to only access the device provided by DRBD—never the backing device. Use the same (bonded) NICs that the rest of the cluster uses to leverage the redundancy provided there.

## 2.4 Other Requirements and Recommendations

For a supported and useful High Availability setup, consider the following recommendations:

### Number of Cluster Nodes

Each cluster must consist of at least two cluster nodes.

---

#### **IMPORTANT: Odd Number of Cluster Nodes**

It is strongly recommended to use an *odd* number of cluster nodes with a *minimum* of three nodes.

A cluster needs quorum to keep services running. Therefore a three-node cluster can tolerate only failure of one node at a time, whereas a five-node cluster can tolerate failures of two nodes etc.

---

### STONITH

---

#### **IMPORTANT: No Support Without STONITH**

A cluster without STONITH is not supported.

---

For a supported High Availability setup, ensure the following:

- Each node in the High Availability cluster must have at least one STONITH device (usually a piece of hardware). We strongly recommend multiple STONITH devices per node, unless SBD is used. SBD provides a way to enable STONITH and fencing in clusters without external power switches, but it requires shared storage.

- The global cluster options `stonith-enabled` and `startup-fencing` must be set to `true`. As soon as you change them, you will lose support.

### Redundant Communication Paths

For a supported High Availability setup, it is required to set up cluster communication via two or more redundant paths. This can be done via:

- *Network Device Bonding* (page 225).
- A second communication channel in Corosync. For details, see Procedure 3.7, “Defining a Redundant Communication Channel” (page 39).

If possible, choose network device bonding.

### Time Synchronization

Cluster nodes should synchronize to an NTP server outside the cluster. For more information, see the *Administration Guide* for SUSE Linux Enterprise Server 11 SP4, available at <http://www.suse.com/doc/>. Refer to the chapter *Time Synchronization with NTP*.

If nodes are not synchronized, log files and cluster reports are very hard to analyze.

### NIC Names

Must be identical on all nodes.

### Host Name and IP Address

Configure host name resolution by editing the `/etc/hosts` file on *each* server in the cluster. To ensure that cluster communication is not slowed down or tampered with by any DNS:

- Use static IP addresses.
- List all cluster nodes in this file with their fully qualified host name and short host name. It is essential that members of the cluster can find each other by name. If the names are not available, internal cluster communication will fail.

For more information, see the *Administration Guide* for SUSE Linux Enterprise Server 11 SP4, available at <http://www.suse.com/doc>. Refer to chapter

*Basic Networking*, section *Configuring a Network Connection with YaST > Configuring Host Name and DNS*.

## Storage Requirements

Some services may require shared storage. For requirements, see Section 2.3, “Storage Requirements” (page 20). You can also use an external NFS share or DRBD. If using an external NFS share, it must be reliably accessible from all cluster nodes via redundant communication paths. See “Redundant Communication Paths” (page 22).

When using SBD as STONITH device, additional requirements apply for the shared storage. For details, see [http://linux-ha.org/wiki/SBD\\_Fencing](http://linux-ha.org/wiki/SBD_Fencing), section *Requirements*.

## SSH

All cluster nodes must be able to access each other via SSH. Tools like `hb_report` or `crm_report` (for troubleshooting) and Hawk's *History Explorer* require passwordless SSH access between the nodes, otherwise they can only collect data from the current node. In case you use a non-standard SSH port, use the `-X` option (see man page). For example, if your SSH port is 3479, invoke an `hb_report` with:

```
root # hb_report -X "-p 3479" [...]
```

---

### NOTE: Regulatory Requirements

If passwordless SSH access does not comply with regulatory requirements, you can use the following work-around for `hb_report`:

Create a user that can log in without a password (for example, using public key authentication). Configure `sudo` for this user so it does not require a `root` password. Start `hb_report` from command line with the `-u` option to specify the user's name. For more information, see the `hb_report` man page.

For the history explorer there is currently no alternative for passwordless login.

---



# Installation and Basic Setup

This chapter describes how to install and set up SUSE® Linux Enterprise High Availability Extension 11 SP4 from scratch. Choose between an automatic setup or a manual setup. The automatic setup enables you to have a cluster up and running within a few minutes (with the choice to adjust any options later on), whereas the manual setup allows you to set your individual options right at the beginning.

Refer to chapter Appendix E, *Upgrading Your Cluster and Updating Software Packages* (page 339) if you want to migrate a cluster that runs an older version of SUSE Linux Enterprise High Availability Extension, or if you want to update software packages on nodes that belong to a running cluster.

## 3.1 Definition of Terms

This chapter uses several terms that are defined below.

### Existing Cluster

The term “existing cluster” is used to refer to any cluster that consists of at least one node. Existing clusters have a basic Corosync configuration that defines the communication channels, but they do not necessarily have resource configuration yet.

### Multicast

A technology used for a one-to-many communication within a network that can be used for cluster communication. Corosync supports both multicast and

unicast. If multicast does not comply with your corporate IT policy, use unicast instead.

---

**NOTE: Switches and Multicast**

If you want to use multicast for cluster communication, make sure your switches support multicast.

---

**Multicast Address (`mcastaddr`)**

IP address to be used for multicasting by the Corosync executive. The IP address can either be IPv4 or IPv6. If IPv6 networking is used, node IDs must be specified. You can use any multicast address in your private network.

**Multicast Port (`mcastport`)**

The port to use for cluster communication. Corosync uses two ports: the specified `mcastport` for receiving multicast, and `mcastport -1` for sending multicast.

**Unicast**

A technology for sending messages to a single network destination. Corosync supports both multicast and unicast. In Corosync, unicast is implemented as UDP-unicast (UDPU).

**Bind Network Address (`bindnetaddr`)**

The network address the Corosync executive should bind to. To ease sharing configuration files across the cluster, OpenAIS uses network interface netmask to mask only the address bits that are used for routing the network. For example, if the local interface is `192.168.5.92` with netmask `255.255.255.0`, set `bindnetaddr` to `192.168.5.0`. If the local interface is `192.168.5.92` with netmask `255.255.255.192`, set `bindnetaddr` to `192.168.5.64`.

---

**NOTE: Network Address for All Nodes**

As the same Corosync configuration will be used on all nodes, make sure to use a network address as `bindnetaddr`, not the address of a specific network interface.

---

**Redundant Ring Protocol (RRP)**

Allows the use of multiple redundant local area networks for resilience against partial or total network faults. This way, cluster communication can still be kept up as long as a single network is operational. Corosync supports the



Totem Redundant Ring Protocol. A logical token-passing ring is imposed on all participating nodes to deliver messages in a reliable and sorted manner. A node is allowed to broadcast a message only if it holds the token. For more information, refer to <http://www.rcsc.de/pdf/icdcs02.pdf>.

When having defined redundant communication channels in Corosync, use RRP to tell the cluster how to use these interfaces. RRP can have three modes (`rrp_mode`):

- If set to `active`, Corosync uses both interfaces actively.
- If set to `passive`, Corosync sends messages alternatively over the available networks.
- If set to `none`, RRP is disabled.

## Csync2

A synchronization tool that can be used to replicate configuration files across all nodes in the cluster, and even across Geo clusters. Csync2 can handle any number of hosts, sorted into synchronization groups. Each synchronization group has its own list of member hosts and its include/exclude patterns that define which files should be synchronized in the synchronization group. The groups, the host names belonging to each group, and the include/exclude rules for each group are specified in the Csync2 configuration file, `/etc/csync2/csync2.cfg`.

For authentication, Csync2 uses the IP addresses and pre-shared keys within a synchronization group. You need to generate one key file for each synchronization group and copy it to all group members.

For more information about Csync2, refer to <http://oss.linbit.com/csync2/paper.pdf>

## conntrack Tools

Allow interaction with the in-kernel connection tracking system for enabling *stateful* packet inspection for iptables. Used by the High Availability Extension to synchronize the connection status between cluster nodes. For detailed information, refer to <http://conntrack-tools.netfilter.org/>.

## AutoYaST

AutoYaST is a system for installing one or more SUSE Linux Enterprise systems automatically and without user intervention. On SUSE Linux Enterprise you can

create an AutoYaST profile that contains installation and configuration data. The profile tells AutoYaST what to install and how to configure the installed system to get a ready-to-use system in the end. This profile can then be used for mass deployment in different ways (for example, to clone existing cluster nodes).

For detailed instructions on how to use AutoYaST in various scenarios, see the *SUSE Linux Enterprise 11 SP4 Deployment Guide*, available at <http://www.suse.com/doc>. Refer to chapter *Automated Installation*.

## 3.2 Overview

The following basic steps are needed for installation and initial cluster setup.

### 1 Installation as Add-on (page 29):

Install the software packages with YaST. Alternatively, you can install them from the command line with `zypper`:

```
root # zypper in -t pattern ha_sles
```

### 2 Initial Cluster Setup:

After installing the software on all nodes that will be part of your cluster, the following steps are needed to initially configure the cluster.

**2a** Defining the Communication Channels (page 36)

**2b** Optional: Defining Authentication Settings (page 41)

**2c** Transferring the Configuration to All Nodes (page 42). Whereas the configuration of `Csync2` is done on one node only, the services `Csync2` and `xinetd` need to be started on all nodes.

**2d** Optional: Synchronizing Connection Status Between Cluster Nodes (page 45)

**2e** Configuring Services (page 46)

**2f** Bringing the Cluster Online (page 48). The `OpenAIS/Corosync` service needs to be started on all nodes.

The cluster setup steps can either be executed automatically (with bootstrap scripts) or manually (with the YaST cluster module or from command line).

- If you decide for an automatic cluster setup, refer to Section 3.4, “Automatic Cluster Setup (sleha-bootstrap)” (page 30).
- For a manual setup (or for adjusting any options after the automatic setup), refer to Section 3.5, “Manual Cluster Setup (YaST)” (page 35).

You can also use a combination of both setup methods, for example: set up one node with YaST cluster and then use `sleha-join` to integrate more nodes.

Existing nodes can also be cloned for mass deployment with AutoYaST. The cloned nodes will have the same packages installed and the same system configuration. For details, refer to Section 3.6, “Mass Deployment with AutoYaST” (page 49).

## 3.3 Installation as Add-on

The packages needed for configuring and managing a cluster with the High Availability Extension are included in the `High Availability` installation pattern. This pattern is only available after SUSE Linux Enterprise High Availability Extension has been installed as an add-on to SUSE® Linux Enterprise Server. For information on how to install add-on products, see the *SUSE Linux Enterprise 11 SP4 Deployment Guide*, available at <http://www.suse.com/doc>. Refer to chapter *Installing Add-On Products*.

### **Procedure 3.1:** *Installing the High Availability Pattern*

- 1 Start YaST as `root` user and select *Software > Software Management*.

Alternatively, start the YaST module as `root` on a command line with `yast2 sw_single`.

- 2 From the *Filter* list, select *Patterns* and activate the *High Availability* pattern in the pattern list.
- 3 Click *Accept* to start installing the packages.

---

**NOTE: Installing Software Packages on All Parties**

The software packages needed for High Availability clusters are *not* automatically copied to the cluster nodes.

---

- 4 Install the High Availability pattern on *all* machines that will be part of your cluster.

If you do not want to install SUSE Linux Enterprise Server 11 SP4 and SUSE Linux Enterprise High Availability Extension 11 SP4 manually on all nodes that will be part of your cluster, use AutoYaST to clone existing nodes. For more information, refer to Section 3.6, “Mass Deployment with AutoYaST” (page 49).

## 3.4 Automatic Cluster Setup (sleha-bootstrap)

The `sleha-bootstrap` package provides everything you need to get a one-node cluster up and running, to make other nodes join, and to remove nodes from an existing cluster:

### Automatically Setting Up the First Node (page 31)

With `sleha-init`, define the basic parameters needed for cluster communication and (optionally) set up a STONITH mechanism to protect your shared storage. This leaves you with a running one-node cluster.

### Adding Nodes to an Existing Cluster (page 32)

With `sleha-join`, add more nodes to your cluster.

### Removing Nodes From An Existing Cluster (page 34)

With `sleha-remove`, remove nodes from your cluster.

All commands execute bootstrap scripts that require only a minimum of time and manual intervention. The bootstrap scripts for initialization and joining automatically open the ports in the firewall that are needed for cluster communication. The configuration is written to `/etc/sysconfig/SuSEfirewall12.d/services/cluster`. Any options set during the bootstrap process can be modified later with the YaST cluster module.

Before starting the automatic setup, make sure that the following prerequisites are fulfilled on all nodes that will participate in the cluster:

### ***Prerequisites***

- The requirements listed in Section 2.2, “Software Requirements” (page 20) and Section 2.4, “Other Requirements and Recommendations” (page 21) are fulfilled.
- The `sleha-bootstrap` package is installed.
- The network is configured according to your needs. For example, a private network is available for cluster communication and network device bonding is configured. For information on bonding, refer to Chapter 11, *Network Device Bonding* (page 225).
- If you want to use SBD for your shared storage, you need one shared block device for SBD. The block device need not be formatted. For more information, refer to Chapter 17, *Storage Protection* (page 285).
- All nodes must be able to see the shared storage via the same paths (`/dev/disk/by-path/...` or `/dev/disk/by-id/...`).

### ***Procedure 3.2: Automatically Setting Up the First Node***

The `sleha-init` command checks for configuration of NTP and guides you through configuration of the cluster communication layer (Corosync), and (optionally) through the configuration of SBD to protect your shared storage. Follow the steps below. For details, refer to the `sleha-init` man page.

- 1 Log in as `root` to the physical or virtual machine you want to use as cluster node.
- 2 Start the bootstrap script by executing

```
root # sleha-init
```

If NTP has not been configured to start at boot time, a message appears.

If you decide to continue anyway, the script will automatically generate keys for SSH access and for the Csync2 synchronization tool and start the services needed for both.

- 3 To configure the cluster communication layer (Corosync):

- 3a** Enter a network address to bind to. By default, the script will propose the network address of `eth0`. Alternatively, enter a different network address, for example the address of `bond0`.
  - 3b** Enter a multicast address. The script proposes a random address that you can use as default.
  - 3c** Enter a multicast port. The script proposes 5405 as default.
- 4** To configure SBD (optional), enter a persistent path to the partition of your block device that you want to use for SBD. The path must be consistent across all nodes in the cluster.

Finally, the script will start the OpenAIS service to bring the one-node cluster online and enable the Web management interface Hawk. The URL to use for Hawk is displayed on the screen.

- 5** For any details of the setup process, check `/var/log/sleha-bootstrap.log`.

You now have a running one-node cluster. Check the cluster status with

`crm status:`

```
root # crm status
Last updated: Thu Jul  3 11:04:10 2014
Last change: Thu Jul  3 10:58:43 2014
Current DC: alice (175704363) - partition with quorum
1 Nodes configured
0 Resources configured

Online: [ alice ]
```

---

### **IMPORTANT: Secure Password**

The bootstrap procedure creates a Linux user named `hacluster` with the password `linux`. You need it for logging in to Hawk. Replace the default password with a secure one as soon as possible:

```
root # passwd hacluster
```

---

### ***Procedure 3.3: Adding Nodes to an Existing Cluster***

If you have a cluster up and running (with one or more nodes), add more cluster nodes with the `sleha-join bootstrap` script. The script only needs access to

an existing cluster node and will complete the basic setup on the current machine automatically. Follow the steps below. For details, refer to the `sleha-join` man page.

If you have configured the existing cluster nodes with the YaST cluster module, make sure the following prerequisites are fulfilled before you run `sleha-join`:

- The `root` user on the existing nodes has SSH keys in place for passwordless login.
- Csync2 is configured on the existing nodes. For details, refer to Procedure 3.9, “Configuring Csync2 with YaST” (page 42).

If you are logged in to the first node via Hawk, you can follow the changes in cluster status and view the resources being activated in the Web interface.

- 1 Log in as `root` to the physical or virtual machine supposed to join the cluster.
- 2 Start the bootstrap script by executing:

```
root # sleha-join
```

If NTP has not been configured to start at boot time, a message appears.

- 3 If you decide to continue anyway, you will be prompted for the IP address of an existing node. Enter the IP address.
- 4 If you have not already configured a passwordless SSH access between both machines, you will also be prompted for the `root` password of the existing node.

After logging in to the specified node, the script will copy the Corosync configuration, configure SSH and Csync2, and will bring the current machine online as new cluster node. Apart from that, it will start the service needed for Hawk. If you have configured shared storage with OCFS2, it will also automatically create the mountpoint directory for the OCFS2 file system.

- 5 Repeat the steps above for all machines you want to add to the cluster.
- 6 For details of the process, check `/var/log/sleha-bootstrap.log`.

Check the cluster status with `crm status`. If you have successfully added a second node, the output will be similar to the following:

```
root # crm status
```

```
Last updated: Thu Jul 3 11:07:10 2014
Last change: Thu Jul 3 10:58:43 2014
Current DC: alice (175704363) - partition with quorum
2 Nodes configured
0 Resources configured

Online: [ alice bob ]
```

---

### **IMPORTANT: Check no-quorum-policy**

After adding all nodes, check if you need to adjust the `no-quorum-policy` in the global cluster options. This is especially important for two-node clusters. For more information, refer to Section 4.1.2, “Option `no-quorum-policy`” (page 54).

---

### ***Procedure 3.4: Removing Nodes From An Existing Cluster***

If you have a cluster up and running (with at least two nodes), you can remove single nodes from the cluster with the `sleha-remove` bootstrap script. You need to know the IP address or host name of the node you want to remove from the cluster. Follow the steps below. For details, refer to the `sleha-remove` man page.

**1** Log in as `root` to one of the cluster nodes.

**2** Start the bootstrap script by executing:

```
root # sleha-remove -c IP_ADDR_OR_HOSTNAME
```

The script enables the `sshd`, stops the OpenAIS service on the specified node, and propagates the files to synchronize with `Csync2` across the remaining nodes.

If you specified a host name and the node to remove cannot be contacted (or the host name cannot be resolved), the script will inform you and ask whether to remove the node anyway. If you specified an IP address and the node cannot be contacted, you will be asked to enter the host name and to confirm whether to remove the node anyway.

**3** To remove more nodes, repeat the step above.

**4** For details of the process, check `/var/log/sleha-bootstrap.log`.

If you need to re-add the removed node at a later point in time, add it with `sleha-join`. For details, refer to Procedure 3.3, “Adding Nodes to an Existing Cluster” (page 32).



**Procedure 3.5:** *Removing the High Availability Extension Software From a Machine*

To remove the High Availability Extension software from a machine that you no longer need as cluster node, proceed as follows.

- 1 Stop the cluster service:

```
root # rcopenais stop
```

- 2 Remove the High Availability Extension add-on:

```
root # zypper rm -t products sle-hae
```

## 3.5 Manual Cluster Setup (YaST)

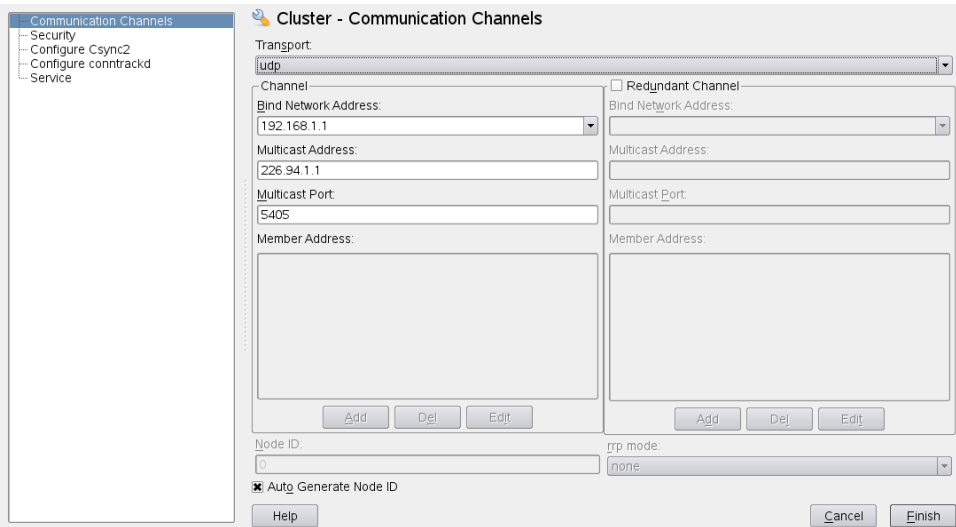
See Section 3.2, “Overview” (page 28) for an overview of all steps for initial setup.

### 3.5.1 YaST Cluster Module

The following sections guide you through each of the setup steps, using the YaST cluster module. To access it, start YaST as `root` and select *High Availability > Cluster*. Alternatively, start the module from command line with `yast2 cluster`.

If you start the cluster module for the first time, it appears as wizard, guiding you through all the steps necessary for basic setup. Otherwise, click the categories on the left panel to access the configuration options for each step.

**Figure 3.1:** YaST Cluster Module—Overview



Note that some options in the YaST cluster module apply only to the current node, whereas others may automatically be transferred to all nodes. Find detailed information about this in the following sections.

## 3.5.2 Defining the Communication Channels

For successful communication between the cluster nodes, define at least one communication channel.

---

### IMPORTANT: Redundant Communication Paths

It is highly recommended to set up cluster communication via two or more redundant paths. This can be done via:

- *Network Device Bonding* (page 225).
- A second communication channel in Corosync. For details, see Procedure 3.7, “Defining a Redundant Communication Channel” (page 39).

If possible, choose network device bonding.

---

**Procedure 3.6:** *Defining the First Communication Channel*

For communication between the cluster nodes, use either multicast (UDP) or unicast (UDPU).

**1** In the YaST cluster module, switch to the *Communication Channels* category.

**2** To use multicast:

**2a** Set the *Transport* protocol to UDP.

**2b** Define the *Bind Network Address*. Set the value to the subnet you will use for cluster multicast.

**2c** Define the *Multicast Address*.

**2d** Define the *Multicast Port*.

With the values entered above, you have now defined *one* communication channel for the cluster. In multicast mode, the same `bindnetaddr`, `mcastaddr`, and `mcastport` will be used for all cluster nodes. All nodes in the cluster will know each other by using the same multicast address. For different clusters, use different multicast addresses.

**Figure 3.2:** *YaST Cluster—Multicast Configuration*

The screenshot shows the 'Cluster - Communication Channels' window in YaST. On the left is a sidebar with a tree view containing 'Communication Channels', 'Security', 'Configure Csync2', 'Configure corosyncd', and 'Service'. The main area is titled 'Cluster - Communication Channels'. It features a 'Transport:' dropdown set to 'udp'. Below this are two columns of configuration fields. The left column, under the 'Channel' header, includes 'Bind Network Address' (a dropdown showing '192.168.1.1'), 'Multicast Address' (a text field with '226.94.1.1'), 'Multicast Port' (a text field with '5405'), and 'Member Address' (a large empty text area). The right column, under the 'Redundant Channel' header (which has an unchecked checkbox), includes identical fields. At the bottom of each column are 'Add', 'Del', and 'Edit' buttons. Below these columns are a 'Node ID' field, a 'trp mode:' dropdown set to 'none', and a checked checkbox for 'Auto Generate Node ID'. At the very bottom are 'Help', 'Cancel', and 'Finish' buttons.

**3** To use unicast:

**3a** Set the *Transport* protocol to UDP.

**3b** Define the *Bind Network Address*. Set the value to the subnet you will use for cluster unicast.

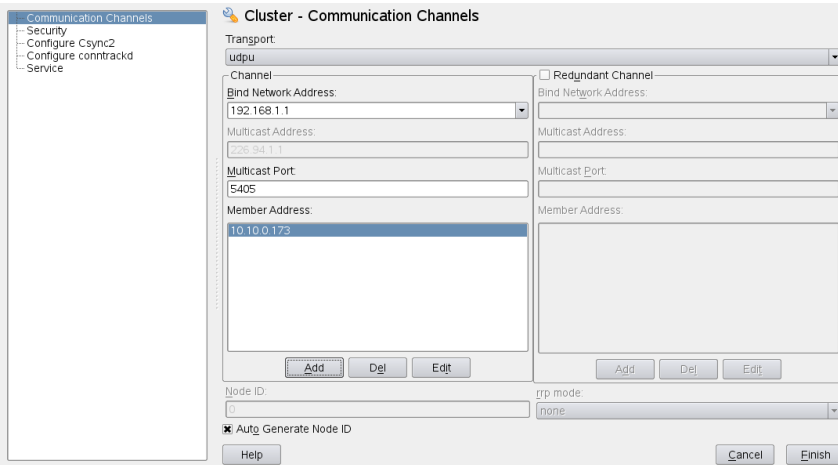
**3c** Define the *Multicast Port*.

**3d** For unicast communication, Corosync needs to know the IP addresses of all nodes in the cluster. For each node that will be part of the cluster, click *Add* and enter the following details:

- *IP Address*
- *Redundant IP Address* (only required if you use a second communication channel in Corosync)
- *Node ID* (only required if the option *Auto Generate Node ID* is disabled)

To modify or remove any addresses of cluster members, use the *Edit* or *Del* buttons.

**Figure 3.3:** YaST Cluster—Unicast Configuration



- 4 The option *Auto Generate Node ID* is enabled by default. If you are using IPv4 addresses, node IDs are optional but they are required when using IPv6 addresses. To automatically generate a unique ID for every cluster node (which is less error-prone than specifying IDs manually for each node), keep this option enabled.
- 5 If you modified any options for an existing cluster, confirm your changes and close the cluster module. YaST writes the configuration to `/etc/corosync/corosync.conf`.
- 6 If needed, define a second communication channel as described below.  
Or click *Next* and proceed with Procedure 3.8, “Enabling Secure Authentication” (page 41).

**Procedure 3.7:** *Defining a Redundant Communication Channel*

If network device bonding cannot be used for any reason, the second best choice is to define a redundant communication channel (a second ring) in Corosync. That way, two physically separate networks can be used for communication. In case one network fails, the cluster nodes can still communicate via the other network.

---

**IMPORTANT: Redundant Rings and `/etc/hosts`**

If multiple rings are configured, each node can have multiple IP addresses. This needs to be reflected in the `/etc/hosts` file of all nodes.

---

- 1 In the YaST cluster module, switch to the *Communication Channels* category.
- 2 Activate *Redundant Channel*. The redundant channel must use the same protocol as the first communication channel you defined.
- 3 If you use multicast, define the *Bind Network Address*, the *Multicast Address* and the *Multicast Port* for the redundant channel.

If you use unicast, define the *Bind Network Address*, the *Multicast Port* and enter the IP addresses of all nodes that will be part of the cluster.

Now you have defined an additional communication channel in Corosync that will form a second token-passing ring. In `/etc/corosync/corosync.conf`, the primary ring (the first channel you have configured) gets the `ringnumber 0`, the second ring (redundant channel) the `ringnumber 1`.

- 4 To tell Corosync how and when to use the different channels, select the *rrp\_mode* you want to use (`active` or `passive`). For more information about the modes, refer to “Redundant Ring Protocol (RRP)” (page 26) or click *Help*. As soon as RRP is used, the Stream Control Transmission Protocol (SCTP) is used for communication between the nodes (instead of TCP). The High Availability Extension monitors the status of the current rings and automatically re-enables redundant rings after faults. Alternatively, you can also check the ring status manually with `corosync-cfgtool`. View the available options with `-h`.

If only one communication channel is defined, *rrp\_mode* is automatically disabled (value `none`).

- 5 If you modified any options for an existing cluster, confirm your changes and close the cluster module. YaST writes the configuration to `/etc/corosync/corosync.conf`.
- 6 For further cluster configuration, click *Next* and proceed with Section 3.5.3, “Defining Authentication Settings” (page 41).

Find an example file for a UDP setup in `/etc/corosync/corosync.conf.example`. An example for UDPU setup is available in `/etc/corosync/corosync.conf.example.udpu`.

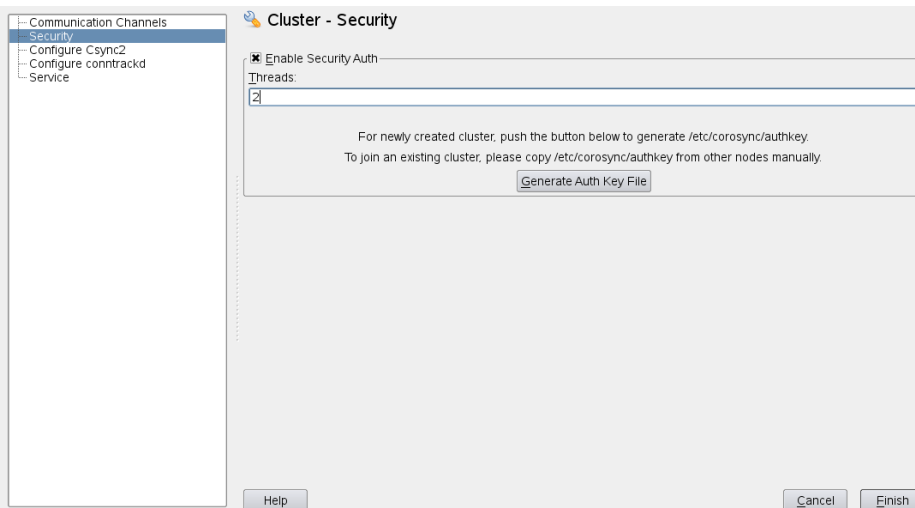
## 3.5.3 Defining Authentication Settings

The next step is to define the authentication settings for the cluster. You can use HMAC/SHA1 authentication that requires a shared secret used to protect and authenticate messages. The authentication key (password) you specify will be used on all nodes in the cluster.

### **Procedure 3.8:** *Enabling Secure Authentication*

- 1 In the YaST cluster module, switch to the *Security* category.
- 2 Activate *Enable Security Auth*.
- 3 For a newly created cluster, click *Generate Auth Key File*. An authentication key is created and written to `/etc/corosync/authkey`.

**Figure 3.4:** *YaST Cluster—Security*



If you want the current machine to join an existing cluster, do not generate a new key file. Instead, copy the `/etc/corosync/authkey` from one of the nodes to the current machine (either manually or with Csync2).

- 4 If you modified any options for an existing cluster, confirm your changes and close the cluster module. YaST writes the configuration to `/etc/corosync/corosync.conf`.

- 5 For further cluster configuration, click *Next* and proceed with Section 3.5.4, “Transferring the Configuration to All Nodes” (page 42).

## 3.5.4 Transferring the Configuration to All Nodes

Instead of copying the resulting configuration files to all nodes manually, use the `csync2` tool for replication across all nodes in the cluster.

This requires the following basic steps:

- 1 Configuring Csync2 with YaST (page 42).
- 2 Synchronizing the Configuration Files with Csync2 (page 43).

Csync2 helps you to keep track of configuration changes and to keep files synchronized across the cluster nodes:

- You can define a list of files that are important for operation.
- You can show changes of these files (against the other cluster nodes).
- You can synchronize the configured files with a single command.
- With a simple shell script in `~/ .bash_logout`, you can be reminded about unsynchronized changes before logging out of the system.

Find detailed information about Csync2 at <http://oss.linbit.com/csync2/> and <http://oss.linbit.com/csync2/paper.pdf>.

### **Procedure 3.9:** *Configuring Csync2 with YaST*

- 1 In the YaST cluster module, switch to the *Csync2* category.
- 2 To specify the synchronization group, click *Add* in the *Sync Host* group and enter the local host names of all nodes in your cluster. For each node, you must use exactly the strings that are returned by the `hostname` command.
- 3 Click *Generate Pre-Shared-Keys* to create a key file for the synchronization group. The key file is written to `/etc/csync2/key_hagroup`. After it has been created, it must be copied manually to all members of the cluster.





- The same Csync2 configuration is available on all nodes. Copy the file `/etc/csync2/csync2.cfg` manually to all nodes after you have configured it as described in Procedure 3.9, “Configuring Csync2 with YaST” (page 42). It is recommended to include this file in the list of files to be synchronized with Csync2.
- Copy the file `/etc/csync2/key_hagroup` you have generated on one node in Step 3 (page 42) to *all* nodes in the cluster. It is needed for authentication by Csync2. However, do *not* regenerate the file on the other nodes as it needs to be the same file on all nodes.
- Both Csync2 and `xinetd` must be running on *all* nodes.

---

### **NOTE: Starting Services at Boot Time**

Execute the following commands on all nodes to make both services start automatically at boot time and to start `xinetd` now:

```
root # chkconfig csync2 on
chkconfig xinetd on
rcxinetd start
```

---

- 1** On the node that you want to copy the configuration *from*, execute the following command:

```
root # csync2 -xv
```

This will synchronize all the files once by pushing them to the other nodes. If all files are synchronized successfully, Csync2 will finish with no errors.

If one or several files that are to be synchronized have been modified on other nodes (not only on the current one), Csync2 will report a conflict. You will get an output similar to the one below:

```
While syncing file /etc/corosync/corosync.conf:
ERROR from peer hex-14: File is also marked dirty here!
Finished with 1 errors.
```

- 2** If you are sure that the file version on the current node is the “best” one, you can resolve the conflict by forcing this file and resynchronizing:

```
root # csync2 -f /etc/corosync/corosync.conf
csync2 -x
```

For more information on the Csync2 options, run `csync2 -help`.

---

**NOTE: Pushing Synchronization After Any Changes**

Csync2 only pushes changes. It does *not* continuously synchronize files between the nodes.

Each time you update files that need to be synchronized, you need to push the changes to the other nodes: Run `csync2 -xv` on the node where you did the changes. If you run the command on any of the other nodes with unchanged files, nothing will happen.

---

## 3.5.5 Synchronizing Connection Status Between Cluster Nodes

To enable *stateful* packet inspection for iptables, configure and use the `conntrack` tools.

- 1 Configuring the `conntrackd` with YaST (page 45).
- 2 Configuring a resource for `conntrackd` (class: `ocf`, provider: `heartbeat`). If you use Hawk to add the resource, use the default values proposed by Hawk.

After configuring the `conntrack` tools, you can use them for *Load Balancing with Linux Virtual Server* (page 231).

### **Procedure 3.11:** *Configuring the conntrackd with YaST*

Use the YaST cluster module to configure the user-space `conntrackd`. It needs a dedicated network interface that is not used for other communication channels. The daemon can be started via a resource agent afterward.

- 1 In the YaST cluster module, switch to the *Configure conntrackd* category.
- 2 Select a *Dedicated Interface* for synchronizing the connection status. The IPv4 address of the selected interface is automatically detected and shown in YaST. It must already be configured and it must support multicast.
- 3 Define the *Multicast Address* to be used for synchronizing the connection status.

- 4 In *Group Number*, define a numeric ID for the group to synchronize the connection status to.
- 5 Click *Generate /etc/conntrackd/conntrackd.conf* to create the configuration file for conntrackd.
- 6 If you modified any options for an existing cluster, confirm your changes and close the cluster module.
- 7 For further cluster configuration, click *Next* and proceed with Section 3.5.6, “Configuring Services” (page 46).

**Figure 3.6:** YaST Cluster—conntrackd

**Cluster - Configure conntrackd**

Conntrackd is a daemon which helps to duplicate firewall status between cluster nodes. YaST can help to configure some basic aspects of conntrackd. You need to start it with the `ocf:heartbeat:conntrackd`.

Dedicated Interface:  IP: 10.11.2.137

Multicast Address:

Group Number:

## 3.5.6 Configuring Services

In the YaST cluster module define whether to start certain services on a node at boot time. You can also use the module to start and stop the services manually. To bring the cluster nodes online and start the cluster resource manager, OpenAIS must be running as a service.

### **Procedure 3.12:** *Enabling the Cluster Services*

- 1 In the YaST cluster module, switch to the *Service* category.

- 2 To start OpenAIS each time this cluster node is booted, select the respective option in the *Booting* group. If you select *Off* in the *Booting* group, you must start OpenAIS manually each time this node is booted. To start OpenAIS manually, use the `rcopenais start` command.

---

**NOTE: No-Start-on-Boot Parameter for OpenAIS**

While generally disabling the cluster service (including other start/stop scripts) at boot time might break the cluster configuration sometimes, enabling it unconditionally at boot time may also lead to unwanted effect with regards to fencing.

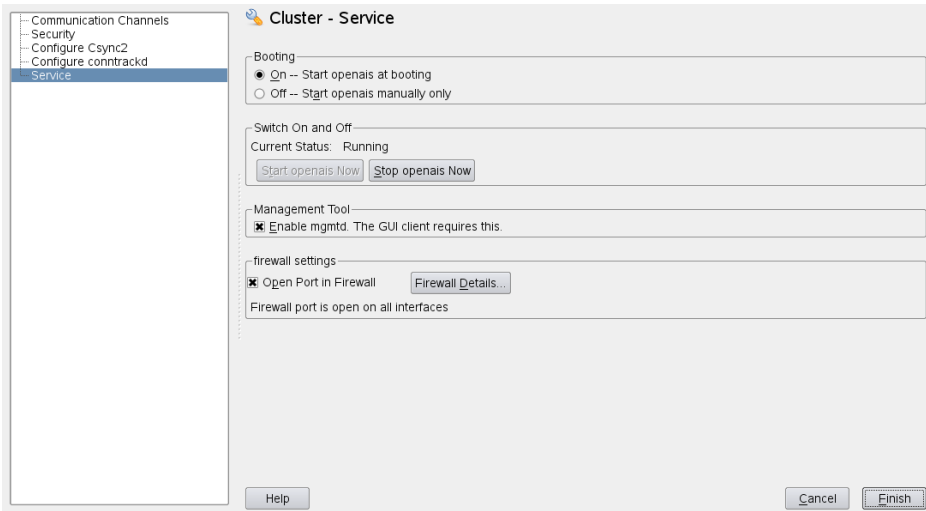
To fine-tune this, insert the `START_ON_BOOT` parameter to `/etc/sysconfig/openais`. Setting `START_ON_BOOT=No` will prevent the OpenAIS service from starting at boot time (allowing you to start it manually whenever you want to start it). The default is `START_ON_BOOT=Yes`.

---

- 3 If you want to use the Pacemaker GUI for configuring, managing and monitoring cluster resources, activate *Enable mgmtd*. This daemon is needed for the GUI.
- 4 To start or stop OpenAIS immediately, click the respective button.
- 5 To open the ports in the firewall that are needed for cluster communication on the current machine, activate *Open Port in Firewall*. The configuration is written to `/etc/sysconfig/SuSEfirewall2.d/services/cluster`.
- 6 If you modified any options for an existing cluster node, confirm your changes and close the cluster module. Note that the configuration only applies to the current machine, not to all cluster nodes.

If you have done the initial cluster setup exclusively with the YaST cluster module, you have now completed the basic configuration steps. Proceed with Section 3.5.7, “Bringing the Cluster Online” (page 48).

**Figure 3.7:** *YaST Cluster—Services*



## 3.5.7 Bringing the Cluster Online

After the initial cluster configuration is done, start the OpenAIS/Corosync service on *each* cluster node to bring the stack online:

**Procedure 3.13:** *Starting OpenAIS/Corosync and Checking the Status*

- 1 Log in to an existing node.
- 2 Check if the service is already running:  

```
root # rcopenais status
```

If not, start OpenAIS/Corosync now:

```
root # rcopenais start
```
- 3 Repeat the steps above for each of the cluster nodes.
- 4 On one of the nodes, check the cluster status with the `crm status` command. If all nodes are online, the output should be similar to the following:

```
root # crm status
```

```
Last updated: Thu Jul  3 11:07:10 2014
Last change: Thu Jul  3 10:58:43 2014
Current DC: alice (175704363) - partition with quorum
2 Nodes configured
0 Resources configured

Online: [ alice bob ]
```

This output indicates that the cluster resource manager is started and is ready to manage resources.

After the basic configuration is done and the nodes are online, you can start to configure cluster resources, using one of the cluster management tools like the crm shell, the Pacemaker GUI, or the HA Web Konsole. For more information, refer to the following chapters.

## 3.6 Mass Deployment with AutoYaST

The following procedure is suitable for deploying cluster nodes which are clones of an already existing node. The cloned nodes will have the same packages installed and the same system configuration.

**Procedure 3.14:** *Cloning a Cluster Node with AutoYaST*

---

### IMPORTANT: Identical Hardware

This scenario assumes you are rolling out SUSE Linux Enterprise High Availability Extension 11 SP4 to a set of machines with identical hardware configurations.

---

If you need to deploy cluster nodes on non-identical hardware, refer to chapter *Automated Installation*, section *Rule-Based Autoinstallation* in the *SUSE Linux Enterprise 11 SP4 Deployment Guide*, available at <http://www.suse.com/doc>.

- 1 Make sure the node you want to clone is correctly installed and configured. For details, refer to Section 3.3, “Installation as Add-on” (page 29), and Section 3.4, “Automatic Cluster Setup (sleha-bootstrap)” (page 30) or Section 3.5, “Manual Cluster Setup (YaST)” (page 35), respectively.

**2** Follow the description outlined in the *SUSE Linux Enterprise 11 SP4 Deployment Guide* for simple mass installation. This includes the following basic steps:

- 2a** Creating an AutoYaST profile. Use the AutoYaST GUI to create and modify a profile based on the existing system configuration. In AutoYaST, choose the *High Availability* module and click the *Clone* button. If needed, adjust the configuration in the other modules and save the resulting control file as XML.
- 2b** Determining the source of the AutoYaST profile and the parameter to pass to the installation routines for the other nodes.
- 2c** Determining the source of the SUSE Linux Enterprise Server and SUSE Linux Enterprise High Availability Extension installation data.
- 2d** Determining and setting up the boot scenario for autoinstallation.
- 2e** Passing the command line to the installation routines, either by adding the parameters manually or by creating an `info` file.
- 2f** Starting and monitoring the autoinstallation process.

After the clone has been successfully installed, execute the following steps to make the cloned node join the cluster:

**Procedure 3.15:** *Bringing the Cloned Node Online*

- 1** Transfer the key configuration files from the already configured nodes to the cloned node with `Csync2` as described in Section 3.5.4, “Transferring the Configuration to All Nodes” (page 42).
- 2** To bring the node online, start the OpenAIS service on the cloned node as described in Section 3.5.7, “Bringing the Cluster Online” (page 48).

The cloned node will now join the cluster because the `/etc/corosync/corosync.conf` file has been applied to the cloned node via `Csync2`. The CIB is automatically synchronized among the cluster nodes.



## **Part II. Configuration and Administration**



# Configuration and Administration Basics

The main purpose of an HA cluster is to manage user services. Typical examples of user services are an Apache Web server or a database. From the user's point of view, the services do something specific when ordered to do so. To the cluster, however, they are only resources which may be started or stopped—the nature of the service is irrelevant to the cluster.

In this chapter, we will introduce some basic concepts you need to know when configuring resources and administering your cluster. The following chapters show you how to execute the main configuration and administration tasks with each of the management tools the High Availability Extension provides.

## 4.1 Global Cluster Options

Global cluster options control how the cluster behaves when confronted with certain situations. They are grouped into sets and can be viewed and modified with the cluster management tools like Hawk and the `crm` shell.

### 4.1.1 Overview

For an overview of all global cluster options and their default values, see *Pacemaker Explained*, available from <http://www.clusterlabs.org/doc/>. Refer to section *Available Cluster Options*.

The predefined values can usually be kept. However, to make key functions of your cluster work correctly, you need to adjust the following parameters after basic cluster setup:

- Option `no-quorum-policy` (page 54)
- Option `stonith-enabled` (page 55)

Learn how to adjust those parameters with the cluster management tools of your choice:

- Hawk: Procedure 5.3, “Modifying Global Cluster Options” (page 98)
- Pacemaker GUI: Procedure 6.1, “Modifying Global Cluster Options” (page 141)
- `crm shell`: Section 7.2, “Configuring Global Cluster Options” (page 177)

## 4.1.2 Option `no-quorum-policy`

This global option defines what to do when a cluster partition does not have quorum (no majority of nodes is part of the partition).

Allowed values are:

`ignore`

The quorum state does not influence the cluster behavior; resource management is continued.

This setting is useful for the following scenarios:

- Two-node clusters: Since a single node failure would always result in a loss of majority, usually you want the cluster to carry on regardless. Resource integrity is ensured using fencing, which also prevents split brain scenarios.
- Resource-driven clusters: For local clusters with redundant communication channels, a split brain scenario only has a certain probability. Thus, a loss of communication with a node most likely indicates that the node has crashed, and that the surviving nodes should recover and start serving the resources again.

If `no-quorum-policy` is set to `ignore`, a 4-node cluster can sustain concurrent failure of three nodes before service is lost. With the other settings, it would lose quorum after concurrent failure of two nodes.

#### `freeze`

If quorum is lost, the cluster partition freezes. Resource management is continued: running resources are not stopped (but possibly restarted in response to monitor events), but no further resources are started within the affected partition.

This setting is recommended for clusters where certain resources depend on communication with other nodes (for example, OCFS2 mounts). In this case, the default setting `no-quorum-policy=stop` is not useful, as it would lead to the following scenario: Stopping those resources would not be possible while the peer nodes are unreachable. Instead, an attempt to stop them would eventually time out and cause a `stop failure`, triggering escalated recovery and fencing.

#### `stop` (default value)

If quorum is lost, all resources in the affected cluster partition are stopped in an orderly fashion.

#### `suicide`

If quorum is lost, all nodes in the affected cluster partition are fenced.

## 4.1.3 Option `stonith-enabled`

This global option defines if to apply fencing, allowing STONITH devices to shoot failed nodes and nodes with resources that cannot be stopped. By default, this global option is set to `true`, because for normal cluster operation it is necessary to use STONITH devices. According to the default value, the cluster will refuse to start any resources if no STONITH resources have been defined.

If you need to disable fencing for any reasons, set `stonith-enabled` to `false`, but be aware that this has impact on the support status for your product. Furthermore, with `stonith-enabled="false"`, resources like the Distributed Lock Manager (DLM) and all services depending on DLM (such as cLVM2, GFS2, and OCFS2) will fail to start.

---

**IMPORTANT: No Support Without STONITH**

A cluster without STONITH is not supported.

---

## 4.2 Cluster Resources

As a cluster administrator, you need to create cluster resources for every resource or application you run on servers in your cluster. Cluster resources can include Web sites, e-mail servers, databases, file systems, virtual machines, and any other server-based applications or services you want to make available to users at all times.

### 4.2.1 Resource Management

Before you can use a resource in the cluster, it must be set up. For example, if you want to use an Apache server as a cluster resource, set up the Apache server first and complete the Apache configuration before starting the respective resource in your cluster.

If a resource has specific environment requirements, make sure they are present and identical on all cluster nodes. This kind of configuration is not managed by the High Availability Extension. You must do this yourself.

---

**NOTE: Do Not Touch Services Managed by the Cluster**

When managing a resource with the High Availability Extension, the same resource must not be started or stopped otherwise (outside of the cluster, for example manually or on boot or reboot). The High Availability Extension software is responsible for all service start or stop actions.

If you need to execute testing or maintenance tasks after the services are already running under cluster control, make sure to put the resources, nodes, or the whole cluster into maintenance mode before you touch any of them manually. For details, see Section 4.7, “Maintenance Mode” (page 88).

---

After having configured the resources in the cluster, use the cluster management tools to start, stop, clean up, remove or migrate any resources manually. For details how to do so with your preferred cluster management tool:

- Hawk: Chapter 5, *Configuring and Managing Cluster Resources (Web Interface)* (page 91)
- Pacemaker GUI: Chapter 6, *Configuring and Managing Cluster Resources (GUI)* (page 137)
- crm shell: Chapter 7, *Configuring and Managing Cluster Resources (Command Line)* (page 169)

## 4.2.2 Supported Resource Agent Classes

For each cluster resource you add, you need to define the standard that the resource agent conforms to. Resource agents abstract the services they provide and present an accurate status to the cluster, which allows the cluster to be non-committal about the resources it manages. The cluster relies on the resource agent to react appropriately when given a start, stop or monitor command.

Typically, resource agents come in the form of shell scripts. The High Availability Extension supports the following classes of resource agents:

### Linux Standards Base (LSB) Scripts

LSB resource agents are generally provided by the operating system/distribution and are found in `/etc/init.d`. To be used with the cluster, they must conform to the LSB init script specification. For example, they must have several actions implemented, which are, at minimum, `start`, `stop`, `restart`, `reload`, `force-reload`, and `status`. For more information, see [http://refspecs.linuxbase.org/LSB\\_4.1.0/LSB-Core-generic/LSB-Core-generic/inisrptact.html](http://refspecs.linuxbase.org/LSB_4.1.0/LSB-Core-generic/LSB-Core-generic/inisrptact.html).

The configuration of those services is not standardized. If you intend to use an LSB script with High Availability, make sure that you understand how the relevant script is configured. Often you can find information about this in the documentation of the relevant package in `/usr/share/doc/packages/PACKAGENAME`.

### Open Cluster Framework (OCF) Resource Agents

OCF RA agents are best suited for use with High Availability, especially when you need multi-state resources or special monitoring abilities. The agents are generally located in `/usr/lib/ocf/resource.d/provider/`. Their functionality is similar to that of LSB scripts. However, the configuration

is always done with environmental variables which allow them to accept and process parameters easily. The OCF specification (as it relates to resource agents) can be found at <http://www.opencf.org/cgi-bin/viewcvs.cgi/specs/ra/resource-agent-api.txt?rev=HEAD&content-type=text/vnd.viewcvs-markup>. OCF specifications have strict definitions of which exit codes must be returned by actions, see Section 8.3, “OCF Return Codes and Failure Recovery” (page 199). The cluster follows these specifications exactly.

All OCF Resource Agents are required to have at least the actions `start`, `stop`, `status`, `monitor`, and `meta-data`. The `meta-data` action retrieves information about how to configure the agent. For example, if you want to know more about the `IPaddr` agent by the provider `heartbeat`, use the following command:

```
OCF_ROOT=/usr/lib/ocf /usr/lib/ocf/resource.d/heartbeat/IPaddr meta-data
```

The output is information in XML format, including several sections (general description, available parameters, available actions for the agent).

Alternatively, use the `crmsh` to view information on OCF resource agents. For details, see Section 7.1.3, “Displaying Information about OCF Resource Agents” (page 172).

### STONITH (Fencing) Resource Agents

This class is used exclusively for fencing related resources. For more information, see Chapter 9, *Fencing and STONITH* (page 203).

The agents supplied with the High Availability Extension are written to OCF specifications.

## 4.2.3 Types of Resources

The following types of resources can be created:

### Primitives

A primitive resource, the most basic type of resource.

Learn how to create primitive resources with your preferred cluster management tool:

- Hawk: Procedure 5.5, “Adding Primitive Resources” (page 101)



- Pacemaker GUI: Procedure 6.2, “Adding Primitive Resources” (page 142)
- crm shell: Section 7.3.1, “Creating Cluster Resources” (page 178)

### Groups

Groups contain a set of resources that need to be located together, started sequentially and stopped in the reverse order. For more information, refer to Section 4.2.5.1, “Groups” (page 60).

### Clones

Clones are resources that can be active on multiple hosts. Any resource can be cloned, provided the respective resource agent supports it. For more information, refer to Section 4.2.5.2, “Clones” (page 62).

### Multi-state Resources (formerly known as Master/Slave Resources)

Multi-state resources are a special type of clone resources that can have multiple modes. For more information, refer to Section 4.2.5.3, “Multi-state Resources” (page 63).

## 4.2.4 Resource Templates

If you want to create lots of resources with similar configurations, defining a resource template is the easiest way. After having been defined, it can be referenced in primitives—or in certain types of constraints, as described in Section 4.4.3, “Resource Templates and Constraints” (page 77).

If a template is referenced in a primitive, the primitive will inherit all operations, instance attributes (parameters), meta attributes, and utilization attributes defined in the template. Additionally, you can define specific operations or attributes for your primitive. If any of these are defined in both the template and the primitive, the values defined in the primitive will take precedence over the ones defined in the template.

Learn how to define resource templates with your preferred cluster configuration tool:

- Hawk: Section 5.3.4, “Using Resource Templates” (page 104)
- crm shell: Section 7.3.2, “Creating Resource Templates” (page 179)

## 4.2.5 Advanced Resource Types

Whereas primitives are the simplest kind of resources and therefore easy to configure, you will probably also need more advanced resource types for cluster configuration, such as groups, clones or multi-state resources.

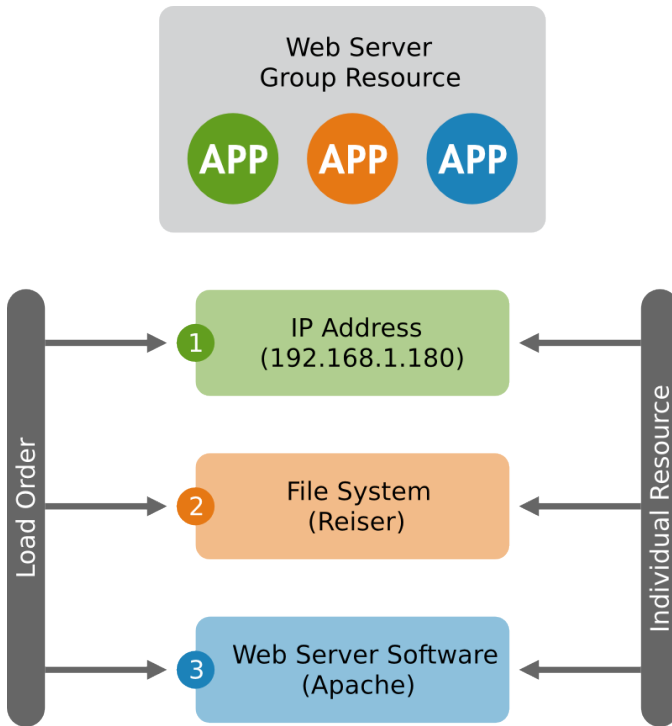
### 4.2.5.1 Groups

Some cluster resources depend on other components or resources. They require that each component or resource starts in a specific order and runs together on the same server with resources it depends on. To simplify this configuration, you can use cluster resource groups.

***Example 4.1:*** *Resource Group for a Web Server*

An example of a resource group would be a Web server that requires an IP address and a file system. In this case, each component is a separate resource that is combined into a cluster resource group. The resource group would run on one or more servers. In case of a software or hardware malfunction, the group would fail over to another server in the cluster, similar to an individual cluster resource.

**Figure 4.1:** *Group Resource*



Groups have the following properties:

#### Starting and Stopping

Resources are started in the order they appear in and stopped in the reverse order.

#### Dependency

If a resource in the group cannot run anywhere, then none of the resources located after that resource in the group is allowed to run.

#### Contents

Groups may only contain a collection of primitive cluster resources. Groups must contain at least one resource, otherwise the configuration is not valid. To refer to the child of a group resource, use the child's ID instead of the group's ID.

## Constraints

Although it is possible to reference the group's children in constraints, it is usually preferable to use the group's name instead.

## Stickiness

Stickiness is additive in groups. Every *active* member of the group will contribute its stickiness value to the group's total. So if the default `resource-stickiness` is 100 and a group has seven members (five of which are active), the group as a whole will prefer its current location with a score of 500.

## Resource Monitoring

To enable resource monitoring for a group, you must configure monitoring separately for each resource in the group that you want monitored.

Learn how to create groups with your preferred cluster management tool:

- Hawk: Procedure 5.16, “Adding a Resource Group” (page 116)
- Pacemaker GUI: Procedure 6.13, “Adding a Resource Group” (page 159)
- crm shell: Section 7.3.9, “Configuring a Cluster Resource Group” (page 188)

## 4.2.5.2 Clones

You may want certain resources to run simultaneously on multiple nodes in your cluster. To do this you must configure a resource as a clone. Examples of resources that might be configured as clones include STONITH and cluster file systems like OCFS2. You can clone any resource provided. This is supported by the resource's Resource Agent. Clone resources may even be configured differently depending on which nodes they are hosted.

There are three types of resource clones:

### Anonymous Clones

These are the simplest type of clones. They behave identically anywhere they are running. Because of this, there can only be one instance of an anonymous clone active per machine.

### Globally Unique Clones

These resources are distinct entities. An instance of the clone running on one node is not equivalent to another instance on another node; nor would any two instances on the same node be equivalent.

### Stateful Clones (Multi-state Resources)

Active instances of these resources are divided into two states, active and passive. These are also sometimes called primary and secondary, or master and slave. Stateful clones can be either anonymous or globally unique. See also Section 4.2.5.3, “Multi-state Resources” (page 63).

Clones must contain exactly one group or one regular resource.

When configuring resource monitoring or constraints, clones have different requirements than simple resources. For details, see *Pacemaker Explained*, available from <http://www.clusterlabs.org/doc/>. Refer to section *Clones - Resources That Get Active on Multiple Hosts*.

Learn how to create clones with your preferred cluster management tool:

- Hawk: Procedure 5.17, “Adding or Modifying Clones” (page 117)
- Pacemaker GUI: Procedure 6.15, “Adding or Modifying Clones” (page 162)
- crm shell: Section 7.3.10, “Configuring a Clone Resource” (page 189).

### 4.2.5.3 Multi-state Resources

Multi-state resources are a specialization of clones. They allow the instances to be in one of two operating modes (called `master` or `slave`, but can mean whatever you want them to mean). Multi-state resources must contain exactly one group or one regular resource.

When configuring resource monitoring or constraints, multi-state resources have different requirements than simple resources. For details, see *Pacemaker Explained*, available from <http://www.clusterlabs.org/doc/>. Refer to section *Multi-state - Resources That Have Multiple Modes*.

## 4.2.6 Resource Options (Meta Attributes)

For each resource you add, you can define options. Options are used by the cluster to decide how your resource should behave—they tell the CRM how to treat a specific resource. Resource options can be set with the `crm_resource --meta` command or with the Pacemaker GUI as described in Procedure 6.3, “Adding or Modifying Meta and Instance Attributes” (page 144). Alternatively, use Hawk: Procedure 5.5, “Adding Primitive Resources” (page 101).

**Table 4.1:** *Options for a Primitive Resource*

Option	Description	Default
priority	If not all resources can be active, the cluster will stop lower priority resources to keep higher priority ones active.	0
target-role	In what state should the cluster attempt to keep this resource? Allowed values: stopped, started, master.	started
is-managed	Is the cluster allowed to start and stop the resource? Allowed values: true, false. If the value is set to false, the status of the resource is still monitored and any failures are reported (which is different from setting a resource to maintenance="true").	true
maintenance	Can the resources be touched manually? Allowed values: true, false. If set to true, all resources become unmanaged: the cluster will stop monitoring them and thus be	false

Option	Description	Default
	oblivious about their status. You can stop or restart cluster resources at will, without the cluster attempting to restart them.	
resource-stickiness	How much does the resource prefer to stay where it is? Defaults to the value of default-resource-stickiness in the rsc_defaults section.	calculated
migration-threshold	How many failures should occur for this resource on a node before making the node ineligible to host this resource?	INFINITY (disabled)
multiple-active	What should the cluster do if it ever finds the resource active on more than one node? Allowed values: block (mark the resource as unmanaged), stop_only, stop_start.	stop_start
failure-timeout	How many seconds to wait before acting as if the failure had	0 (disabled)

Option	Description	Default
	not occurred (and potentially allowing the resource back to the node on which it failed)?	
<code>allow-migrate</code>	Allow resource migration for resources which support <code>migrate_to/</code> <code>migrate_from</code> actions.	<code>false</code>
<code>remote-node</code>	<p>The name of the remote node this resource defines. This both enables the resource as a remote node and defines the unique name used to identify the remote node. If no other parameters are set, this value will also be assumed as the host name to connect to at <code>remote-port</code>.</p> <hr/> <p><b>WARNING: Use Unique IDs</b></p> <p>This value must not overlap with any existing resource or node IDs.</p> <hr/>	<code>none</code> (disabled)



Option	Description	Default
<code>remote-port</code>	Custom port for the guest connection to <code>pacemaker_remote</code> .	3121
<code>remote-addr</code>	The IP address or host name to connect to if the remote node's name is not the host name of the guest.	<code>remote-node</code> (value used as host name)
<code>remote-connect-timeout</code>	How long before a pending guest connection will time out.	60s

## 4.2.7 Instance Attributes (Parameters)

The scripts of all resource classes can be given parameters which determine how they behave and which instance of a service they control. If your resource agent supports parameters, you can add them with the `crm_resource` command or with the GUI as described in Procedure 6.3, “Adding or Modifying Meta and Instance Attributes” (page 144). Alternatively, use Hawk: Procedure 5.5, “Adding Primitive Resources” (page 101). In the `crm` command line utility and in Hawk, instance attributes are called `params` or `Parameter`, respectively. The list of instance attributes supported by an OCF script can be found by executing the following command as `root`:

```
root # crm ra info [class:[provider:]]resource_agent
```

or (without the optional parts):

```
root # crm ra info resource_agent
```

The output lists all the supported attributes, their purpose and default values.

For example, the command

```
root # crm ra info IPAddr
```

## returns the following output:

Manages virtual IPv4 addresses (portable version) (ocf:heartbeat:IPAddr)

This script manages IP alias IP addresses  
It can add an IP alias, or remove one.

Parameters (\* denotes required, [] the default):

ip\* (string): IPv4 address

The IPv4 address to be configured in dotted quad notation, for example  
"192.168.1.1".

nic (string, [eth0]): Network interface

The base network interface on which the IP address will be brought  
online.

If left empty, the script will try and determine this from the  
routing table.

Do NOT specify an alias interface in the form eth0:1 or anything here;  
rather, specify the base interface only.

cidr\_netmask (string): Netmask

The netmask for the interface in CIDR format. (ie, 24), or in  
dotted quad notation 255.255.255.0).

If unspecified, the script will also try to determine this from the  
routing table.

broadcast (string): Broadcast address

Broadcast address associated with the IP. If left empty, the script will  
determine this from the netmask.

iflabel (string): Interface label

You can specify an additional label for your IP address here.

lvs\_support (boolean, [false]): Enable support for LVS DR

Enable support for LVS Direct Routing configurations. In case a IP  
address is stopped, only move it to the loopback device to allow the  
local node to continue to service requests, but no longer advertise it  
on the network.

local\_stop\_script (string):

Script called when the IP is released

local\_start\_script (string):

Script called when the IP is added

ARP\_INTERVAL\_MS (integer, [500]): milliseconds between gratuitous ARPs  
milliseconds between ARPs

ARP\_REPEAT (integer, [10]): repeat count

How many gratuitous ARPs to send out when bringing up a new address

```
ARP_BACKGROUND (boolean, [yes]): run in background
run in background (no longer any reason to do this)

ARP_NETMASK (string, [ffffffffffff]): netmask for ARP
netmask for ARP - in nonstandard hexadecimal format.

Operations' defaults (advisory minimum):

start          timeout=90
stop           timeout=100
monitor_0      interval=5s timeout=20s
```

---

**NOTE: Instance Attributes for Groups, Clones or Multi-state Resources**

Note that groups, clones and multi-state resources do not have instance attributes. However, any instance attributes set will be inherited by the group's, clone's or multi-state resource's children.

---

# 4.2.8 Resource Operations

By default, the cluster will not ensure that your resources are still healthy. To instruct the cluster to do this, you need to add a monitor operation to the resource's definition. Monitor operations can be added for all classes or resource agents. For more information, refer to Section 4.3, "Resource Monitoring" (page 72).

**Table 4.2:** *Resource Operation Properties*

Operation	Description
id	Your name for the action. Must be unique. (The ID is not shown).
name	The action to perform. Common values: monitor, start, stop.
interval	How frequently to perform the operation. Unit: seconds
timeout	How long to wait before declaring the action has failed.

Operation	Description
<code>requires</code>	What conditions need to be satisfied before this action occurs. Allowed values: <code>nothing</code> , <code>quorum</code> , <code>fencing</code> . The default depends on whether fencing is enabled and if the resource's class is <code>stonith</code> . For <code>STONITH</code> resources, the default is <code>nothing</code> .
<code>on-fail</code>	<p>The action to take if this action ever fails. Allowed values:</p> <ul style="list-style-type: none"> <li>• <code>ignore</code>: Pretend the resource did not fail.</li> <li>• <code>block</code>: Do not perform any further operations on the resource.</li> <li>• <code>stop</code>: Stop the resource and do not start it elsewhere.</li> <li>• <code>restart</code>: Stop the resource and start it again (possibly on a different node).</li> <li>• <code>fence</code>: Bring down the node on which the resource failed (<code>STONITH</code>).</li> <li>• <code>standby</code>: Move <i>all</i> resources away from the node on which the resource failed.</li> </ul>
<code>enabled</code>	If <code>false</code> , the operation is treated as if it does not exist. Allowed values: <code>true</code> , <code>false</code> .

Operation	Description
<code>role</code>	Run the operation only if the resource has this role.
<code>record-pending</code>	Can be set either globally or for individual resources. Makes the CIB reflect the state of “in-flight” operations on resources.
<code>description</code>	Description of the operation.

## 4.2.9 Timeout Values

Timeouts values for resources can be influenced by the following parameters:

- `op_defaults` (global timeout for operations),
- a specific timeout value defined in a resource template,
- a specific timeout value defined for a resource.

---

### NOTE: Priority of Values

If a *specific* value is defined for a resource, it takes precedence over the global default. A specific value for a resource also takes precedence over a value that is defined in a resource template.

---

Getting timeout values right is very important. Setting them too low will result in a lot of (unnecessary) fencing operations for the following reasons:

1. If a resource runs into a timeout, it fails and the cluster will try to stop it.
2. If stopping the resource also fails (for example, because the timeout for stopping is set too low), the cluster will fence the node. It considers the node where this happens to be out of control.

The CRM executes an initial monitoring for each resource on every node, the so-called *probe*. A probe is also executed after the cleanup of a resource. If no

specific timeout is configured for the resource's monitoring operation, the CRM will automatically check for any other monitoring operations. If multiple monitoring operations are defined for a resource, the CRM will select the one with the smallest interval and will use its timeout value as default timeout for probing. If no monitor operation is configured, the cluster-wide default applies. The default is 20 seconds (if not specified otherwise by configuring the `op_defaults` parameter). If you do not want to rely on the automatic calculation or the `op_defaults` value, define a specific timeout for this monitoring. Do so by adding a monitoring operation to the respective resource, with the `interval` set to 0, for example:

```
crm(live)configure# primitive rscl ocf:pacemaker:Dummy \  
  op monitor interval="0" timeout="60"
```

The probe of `rscl` will time out in 60s, independent of the global timeout defined in `op_defaults`, or any other operation timeouts configured.

You can adjust the global default for operations and set any specific timeout values with both `crmsh` and Hawk. The best practice for determining and setting timeout values is as follows:

#### ***Procedure 4.1: Determining Timeout Values***

- 1 Check how long it takes your resources to start and stop (under load).
- 2 If needed, add the `op_defaults` parameter and set the (default) timeout value accordingly:
  - 2a For example, set `op_defaults` to 60 seconds:

```
crm(live)configure# op_defaults timeout=60
```
  - 2b For resources that need longer periods of time, define individual timeout values.
- 3 When configuring operations for a resource, add separate `start` and `stop` operations. When configuring operations with Hawk, it will provide useful timeout proposals for those operations.

## **4.3 Resource Monitoring**

If you want to ensure that a resource is running, you must configure resource monitoring for it.

If the resource monitor detects a failure, the following takes place:

- Log file messages are generated, according to the configuration specified in the logging section of `/etc/corosync/corosync.conf`. By default, the logs are written to syslog, usually `/var/log/messages`.
- The failure is reflected in the cluster management tools (Pacemaker GUI, Hawk, `crm_mon`), and in the CIB status section.
- The cluster initiates noticeable recovery actions which may include stopping the resource to repair the failed state and restarting the resource locally or on another node. The resource also may not be restarted, depending on the configuration and state of the cluster.

If you do not configure resource monitoring, resource failures after a successful start will not be communicated, and the cluster will always show the resource as healthy.

Usually, resources are only monitored by the cluster as long as they are running. However, to detect concurrency violations, also configure monitoring for resources which are stopped. For example:

```
primitive dummy1 ocf:heartbeat:Dummy \  
    op monitor interval="300s" role="Stopped" timeout="10s" \  
    op monitor interval="30s" timeout="10s"
```

This configuration triggers a monitoring operation every 300 seconds for the resource `dummy1` as soon as it is in `role="Stopped"`. When running, it will be monitored every 30 seconds.

Learn how to add monitor operations to resources with your preferred cluster management tool:

- Hawk: Procedure 5.15, “Adding or Modifying Monitor Operations” (page 114)
- Pacemaker GUI: Procedure 6.12, “Adding or Modifying Monitor Operations” (page 157)
- `crm` shell: Section 7.3.8, “Configuring Resource Monitoring” (page 188)

## 4.4 Resource Constraints

Having all the resources configured is only part of the job. Even if the cluster knows all needed resources, it might still not be able to handle them correctly. Resource

constraints let you specify which cluster nodes resources can run on, what order resources will load, and what other resources a specific resource is dependent on.

## 4.4.1 Types of Constraints

There are three different kinds of constraints available:

### Resource Location

Locational constraints that define on which nodes a resource may be run, may not be run or is preferred to be run.

### Resource Colocation

Colocational constraints that tell the cluster which resources may or may not run together on a node.

### Resource Order

Ordering constraints to define the sequence of actions.

### 4.4.1.1 Resource Sets

#### Using Resource Sets for Defining Constraints

As an alternative format for defining location, colocation or ordering constraints, you can use `resource sets`, where primitives are grouped together in one set. Previously this was possible either by defining a resource group (which could not always accurately express the design), or by defining each relationship as an individual constraint. The latter caused a constraint explosion as the number of resources and combinations grew. The configuration via resource sets is not necessarily less verbose, but is easier to understand and maintain, as the following examples show.

#### **Example 4.2:** *A Resource Set for Location Constraints*

For example, you can use the following configuration of a resource set (`loc-  
alice`) in the `crmsh` to place two virtual IPs (`vip1` and `vip2`) on the same node, `alice`:

```
crm(live)configure# primitive vip1 ocf:heartbeat:IPaddr2 params  
ip=192.168.1.5
```



```
crm(live)configure# primitive vip1 ocf:heartbeat:IPaddr2 params
ip=192.168.1.6
crm(live)configure# location loc-alice { vip1 vip2 } inf: alice
```

If you want to use resource sets to replace a configuration of colocation constraints, consider the following two examples:

**Example 4.3:** *A Chain of Colocated Resources*

```
<constraints>
  <rsc_colocation id="coloc-1" rsc="B" with-rsc="A" score="INFINITY"/>
  <rsc_colocation id="coloc-2" rsc="C" with-rsc="B" score="INFINITY"/>
  <rsc_colocation id="coloc-3" rsc="D" with-rsc="C" score="INFINITY"/>
</constraints>
```

The same configuration expressed by a resource set:

```
<constraints>
  <rsc_colocation id="coloc-1" score="INFINITY" >
    <resource_set id="colocated-set-example" sequential="true">
      <resource_ref id="A"/>
      <resource_ref id="B"/>
      <resource_ref id="C"/>
      <resource_ref id="D"/>
    </resource_set>
  </rsc_colocation>
</constraints>
```

If you want to use resource sets to replace a configuration of ordering constraints, consider the following two examples:

**Example 4.4:** *A Chain of Ordered Resources*

```
<constraints>
  <rsc_order id="order-1" first="A" then="B" />
  <rsc_order id="order-2" first="B" then="C" />
  <rsc_order id="order-3" first="C" then="D" />
</constraints>
```

The same purpose can be achieved by using a resource set with ordered resources:

**Example 4.5:** *A Chain of Ordered Resources Expressed as Resource Set*

```
<constraints>
  <rsc_order id="order-1">
    <resource_set id="ordered-set-example" sequential="true">
      <resource_ref id="A"/>
      <resource_ref id="B"/>
      <resource_ref id="C"/>
      <resource_ref id="D"/>
    </resource_set>
  </rsc_order>
</constraints>
```

```
        </resource_set>
        </rsc_order>
</constraints>
```

Sets can be either ordered (`sequential=true`) or unordered (`sequential=false`). Furthermore, the `require-all` attribute can be used to switch between AND and OR logic.

## Resource Sets for Colocation Constraints Without Dependencies

Sometimes it is useful to place a group of resources on the same node (defining a colocation constraint), but without having hard dependencies between the resources. For example, you want two resources to be placed on the same node, but you do *not* want the cluster to restart the other one if one of them fails. This can be achieved on the `crm` shell by using the `weak bond` command.

Learn how to set these “weak bonds” with your preferred cluster management tool:

- `crms`: Section 7.3.4.3, “Collocating Sets for Resources Without Dependency” (page 183)

### 4.4.1.2 For More Information

Learn how to add the various kinds of constraints with your preferred cluster management tool:

- Hawk: Section 5.3.5, “Configuring Resource Constraints” (page 106)
- `crms`: Section 7.3.4, “Configuring Resource Constraints” (page 181)

For more information on configuring constraints and detailed background information about the basic concepts of ordering and colocation, refer to the following documents. They are available at <http://www.clusterlabs.org/doc/>:

- *Pacemaker Explained*, chapter *Resource Constraints*
- *Colocation Explained*
- *Ordering Explained*

## 4.4.2 Scores and Infinity

When defining constraints, you also need to deal with scores. Scores of all kinds are integral to how the cluster works. Practically everything from migrating a resource to deciding which resource to stop in a degraded cluster is achieved by manipulating scores in some way. Scores are calculated on a per-resource basis and any node with a negative score for a resource cannot run that resource. After calculating the scores for a resource, the cluster then chooses the node with the highest score.

INFINITY is currently defined as 1,000,000. Additions or subtractions with it stick to the following three basic rules:

- Any value + INFINITY = INFINITY
- Any value - INFINITY = -INFINITY
- INFINITY - INFINITY = -INFINITY

When defining resource constraints, you specify a score for each constraint. The score indicates the value you are assigning to this resource constraint. Constraints with higher scores are applied before those with lower scores. By creating additional location constraints with different scores for a given resource, you can specify an order for the nodes that a resource will fail over to.

## 4.4.3 Resource Templates and Constraints

If you have defined a resource template (see Section 4.2.4, “Resource Templates” (page 59)), it can be referenced in the following types of constraints:

- order constraints,
- colocation constraints,
- rsc\_ticket constraints (for Geo clusters).

However, colocation constraints must not contain more than one reference to a template. Resource sets must not contain a reference to a template.

Resource templates referenced in constraints stand for all primitives which are derived from that template. This means, the constraint applies to all primitive

resources referencing the resource template. Referencing resource templates in constraints is an alternative to resource sets and can simplify the cluster configuration considerably. For details about resource sets, refer to Procedure 5.11, “Using Resource Sets for Colocation or Order Constraints” (page 110).

## 4.4.4 Failover Nodes

A resource will be automatically restarted if it fails. If that cannot be achieved on the current node, or it fails `N` times on the current node, it will try to fail over to another node. Each time the resource fails, its failcount is raised. You can define a number of failures for resources (a `migration-threshold`), after which they will migrate to a new node. If you have more than two nodes in your cluster, the node a particular resource fails over to is chosen by the High Availability software.

However, you can specify the node a resource will fail over to by configuring one or several location constraints and a `migration-threshold` for that resource.

Learn how to specify failover nodes with your preferred cluster management tool:

- Hawk: Section 5.3.6, “Specifying Resource Failover Nodes” (page 111)
- Pacemaker GUI: Section 6.3.5, “Specifying Resource Failover Nodes” (page 151)
- `crm shell`: Section 7.3.5, “Specifying Resource Failover Nodes” (page 184)

### **Example 4.6:** *Migration Threshold—Process Flow*

For example, let us assume you have configured a location constraint for resource `rscl` to preferably run on `alice`. If it fails there, `migration-threshold` is checked and compared to the failcount. If `failcount >= migration-threshold` then the resource is migrated to the node with the next best preference.

After the threshold has been reached, the node will no longer be allowed to run the failed resource until the resource's failcount is reset. This can be done manually by the cluster administrator or by setting a `failure-timeout` option for the resource.

For example, a setting of `migration-threshold=2` and `failure-timeout=60s` would cause the resource to migrate to a new node after two

failures. It would be allowed to move back (depending on the stickiness and constraint scores) after one minute.

There are two exceptions to the migration threshold concept, occurring when a resource either fails to start or fails to stop:

- Start failures set the failcount to `INFINITY` and thus always cause an immediate migration.
- Stop failures cause fencing (when `stonith-enabled` is set to `true` which is the default).

In case there is no `STONITH` resource defined (or `stonith-enabled` is set to `false`), the resource will not migrate.

For details on using migration thresholds and resetting failcounts with your preferred cluster management tool:

- Hawk: Section 5.3.6, “Specifying Resource Failover Nodes” (page 111)
- Pacemaker GUI: Section 6.3.5, “Specifying Resource Failover Nodes” (page 151)
- `crm shell`: Section 7.3.5, “Specifying Resource Failover Nodes” (page 184)

## 4.4.5 Failback Nodes

A resource might fail back to its original node when that node is back online and in the cluster. If you want to prevent a resource from failing back to the node that it was running on, or if you want to specify a different node for the resource to fail back to, change its `resource stickiness` value. You can either specify resource stickiness when you are creating a resource, or afterwards.

Consider the following implications when specifying resource stickiness values:

Value is 0:

This is the default. The resource will be placed optimally in the system. This may mean that it is moved when a “better” or less loaded node becomes available. This option is almost equivalent to automatic failback, except that the resource may be moved to a node that is not the one it was previously active on.

Value is greater than 0:

The resource will prefer to remain in its current location, but may be moved if a more suitable node is available. Higher values indicate a stronger preference for a resource to stay where it is.

Value is less than 0:

The resource prefers to move away from its current location. Higher absolute values indicate a stronger preference for a resource to be moved.

Value is `INFINITY`:

The resource will always remain in its current location unless forced off because the node is no longer eligible to run the resource (node shutdown, node standby, reaching the `migration-threshold`, or configuration change). This option is almost equivalent to completely disabling automatic failback.

Value is `-INFINITY`:

The resource will always move away from its current location.

## 4.4.6 Placing Resources Based on Their Load Impact

Not all resources are equal. Some, such as Xen guests, require that the node hosting them meets their capacity requirements. If resources are placed such that their combined need exceed the provided capacity, the resources diminish in performance (or even fail).

To take this into account, the High Availability Extension allows you to specify the following parameters:

1. The capacity a certain node *provides*.
2. The capacity a certain resource *requires*.
3. An overall strategy for placement of resources.

Learn how to configure these settings with your preferred cluster management tool:

- Pacemaker GUI: Section 6.3.7, “Configuring Placement of Resources Based on Load Impact” (page 153)

- `crm shell`: Section 7.3.7, “Configuring Placement of Resources Based on Load Impact” (page 185)

A node is considered eligible for a resource if it has sufficient free capacity to satisfy the resource's requirements. The nature of the capacities is completely irrelevant for the High Availability Extension; it only makes sure that all capacity requirements of a resource are satisfied before moving a resource to a node.

To manually configure the resource's requirements and the capacity a node provides, use utilization attributes. You can name the utilization attributes according to your preferences and define as many name/value pairs as your configuration needs. However, the attribute's values must be integers.

If multiple resources with utilization attributes are grouped or have colocation constraints, the High Availability Extension takes that into account. If possible, the resources will be placed on a node that can fulfill *all* capacity requirements.

---

**NOTE: Utilization Attributes for Groups**

It is impossible to set utilization attributes directly for a resource group. However, to simplify the configuration for a group, you can add a utilization attribute with the total capacity needed to any of the resources in the group.

---

The High Availability Extension also provides means to detect and configure both node capacity and resource requirements automatically:

The `NodeUtilization` resource agent checks the capacity of a node (regarding CPU and RAM). To configure automatic detection, create a clone resource of the following class, provider, and type: `ocf:pacemaker:NodeUtilization`. One instance of the clone should be running on each node. After the instance has started, a utilization section will be added to the node's configuration in CIB.

For automatic detection of a resource's minimal requirements (regarding RAM and CPU) the `Xen` resource agent has been improved. Upon start of a `Xen` resource, it will reflect the consumption of RAM and CPU. Utilization attributes will automatically be added to the resource configuration.

Apart from detecting the minimal requirements, the High Availability Extension also allows to monitor the current utilization via the `VirtualDomain` resource agent. It detects CPU and RAM use of the virtual machine. To use this feature, configure a resource of the following class, provider

and type: `ocf:heartbeat:VirtualDomain`. The following instance attributes are available: `autoset_utilization_cpu` and `autoset_utilization_hv_memory`. Both default to `true`. This updates the utilization values in the CIB during each monitoring cycle.

Independent of manually or automatically configuring capacity and requirements, the placement strategy must be specified with the `placement-strategy` property (in the global cluster options). The following values are available:

`default` (default value)

Utilization values are not considered. Resources are allocated according to location scoring. If scores are equal, resources are evenly distributed across nodes.

`utilization`

Utilization values are considered when deciding if a node has enough free capacity to satisfy a resource's requirements. However, load-balancing is still done based on the number of resources allocated to a node.

`minimal`

Utilization values are considered when deciding if a node has enough free capacity to satisfy a resource's requirements. An attempt is made to concentrate the resources on as few nodes as possible (to achieve power savings on the remaining nodes).

`balanced`

Utilization values are considered when deciding if a node has enough free capacity to satisfy a resource's requirements. An attempt is made to distribute the resources evenly, thus optimizing resource performance.

---

### **NOTE: Configuring Resource Priorities**

The available placement strategies are best-effort—they do not yet use complex heuristic solvers to always reach optimum allocation results. Ensure that resource priorities are properly set so that your most important resources are scheduled first.

---

### **Example 4.7:** *Example Configuration for Load-Balanced Placing*

The following example demonstrates a three-node cluster of equal nodes, with four virtual machines.



```

node alice utilization memory="4000"
node bob utilization memory="4000"
node charly utilization memory="4000"
primitive xenA ocf:heartbeat:Xen utilization hv_memory="3500" \
    params xmfile="/etc/xen/shared-vm/vm1"
    meta priority="10"
primitive xenB ocf:heartbeat:Xen utilization hv_memory="2000" \
    params xmfile="/etc/xen/shared-vm/vm2"
    meta priority="1"
primitive xenC ocf:heartbeat:Xen utilization hv_memory="2000" \
    params xmfile="/etc/xen/shared-vm/vm3"
    meta priority="1"
primitive xenD ocf:heartbeat:Xen utilization hv_memory="1000" \
    params xmfile="/etc/xen/shared-vm/vm4"
    meta priority="5"
property placement-strategy="minimal"

```

With all three nodes up, resource `xenA` will be placed onto a node first, followed by `xenD`. `xenB` and `xenC` would either be allocated together or one of them with `xenD`.

If one node failed, too little total memory would be available to host them all. `xenA` would be ensured to be allocated, as would `xenD`. However, only one of the remaining resources `xenB` or `xenC` could still be placed. Since their priority is equal, the result would still be open. To resolve this ambiguity as well, you would need to set a higher priority for either one.

## 4.4.7 Grouping Resources by Using Tags

Tags are a new feature that has been added to Pacemaker recently. Tags are a way to refer to multiple resources at once, without creating any colocation or ordering relationship between them. This can be useful for grouping conceptually related resources. For example, if you have several resources related to a database, create a tag called `databases` and add all resources related to the database to this tag. This allows you to stop or start them all with a single command.

Tags can also be used in constraints. For example, the following location constraint `loc-db-prefer` applies to the set of resources tagged with `databases`:

```
location loc-db-prefer databases 100: alice
```

Learn how to create tags with your preferred cluster management tool:

- `crmsh`: Section 7.4.5, “Grouping/Tagging Resources” (page 192)

## 4.5 Managing Services on Remote Hosts

The possibilities for monitoring and managing services on remote hosts has become increasingly important during the last few years. SUSE Linux Enterprise High Availability Extension 11 SP3 offered fine-grained monitoring of services on remote hosts via monitoring plug-ins. The recent addition of the `pacemaker_remote` service now allows SUSE Linux Enterprise High Availability Extension 11 SP4 to fully manage and monitor resources on remote hosts just as if they were a real cluster node—without the need to install the cluster stack on the remote machines.

### 4.5.1 Monitoring Services on Remote Hosts with Nagios Plug-ins

Monitoring of virtual machines can be done with the VM agent (which only checks if the guest shows up in the hypervisor), or by external scripts called from the VirtualDomain or Xen agent. Up to now, more fine-grained monitoring was only possible with a full setup of the High Availability stack within the virtual machines.

By providing support for Nagios plug-ins, the High Availability Extension now also allows you to monitor services on remote hosts. You can collect external statuses on the guests without modifying the guest image. For example, VM guests might run Web services or simple network resources that need to be accessible. With the Nagios resource agents, you can now monitor the Web service or the network resource on the guest. In case these services are not reachable anymore, the High Availability Extension will trigger a restart or migration of the respective guest.

If your guests depend on a service (for example, an NFS server to be used by the guest), the service can either be an ordinary resource, managed by the cluster, or an external service that is monitored with Nagios resources instead.

To configure the Nagios resources, the following packages must be installed on the host:

- `nagios-plugins`
- `nagios-plugins-metadata`

YaST or Zypper will resolve any dependencies on further packages, if required.

A typical use case is to configure the Nagios plug-ins as resources belonging to a resource container, which usually is a VM. The container will be restarted if any of its resources has failed. Refer to Example 4.8, “Configuring Resources for Nagios Plug-ins” (page 85) for a configuration example. Alternatively, Nagios resource agents can also be configured as ordinary resources if you want to use them for monitoring hosts or services via the network.

#### **Example 4.8:** *Configuring Resources for Nagios Plug-ins*

```
primitive vml ocf:heartbeat:VirtualDomain \  
  params hypervisor="qemu:///system" config="/etc/libvirt/qemu/vml.xml" \  
    op start interval="0" timeout="90" \  
    op stop interval="0" timeout="90" \  
    op monitor interval="10" timeout="30" \  
primitive vml-sshd nagios:check_tcp \  
  params hostname="vml" port="22" \❶ \  
    op start interval="0" timeout="120" \❷ \  
    op monitor interval="10" \  
group g-vml-and-services vml vml-sshd \  
  meta container="vml"❸
```

- ❶ The supported parameters are same as the long options of a Nagios plug-in. Nagios plug-ins connect to services with the parameter `hostname`. Therefore the attribute's value must be a resolvable host name or an IP address.
- ❷ As it takes some time to get the guest operating system up and its services running, the start timeout of the Nagios resource must be long enough.
- ❸ A cluster resource container of type `ocf:heartbeat:Xen`, `ocf:heartbeat:VirtualDomain` or `ocf:heartbeat:lxc`. It can either be a VM or a Linux Container.

The example above contains only one Nagios resource for the `check_tcp` plug-in, but multiple Nagios resources for different plug-in types can be configured (for example, `check_http` or `check_udp`).

If the host names of the services are the same, the `hostname` parameter can also be specified for the group, instead of adding it to the individual primitives. For example:

```
group g-vml-and-services vml vml-sshd vml-httpd \  
  meta container="vml" \  
  params hostname="vml"
```

If any of the services monitored by the Nagios plug-ins fail within the VM, the cluster will detect that and restart the container resource (the VM). Which action

to take in this case can be configured by specifying the `on-fail` attribute for the service's monitoring operation. It defaults to `restart-container`.

Failure counts of services will be taken into account when considering the VM's `migration-threshold`.

## 4.5.2 Managing Services on Remote Nodes with `pacemaker_remote`

With the `pacemaker_remote` service, High Availability clusters can be extended to virtual nodes or remote bare-metal machines. They do not need to run the cluster stack to become members of the cluster.

The High Availability Extension can now launch virtual environments (KVM and LXC), plus the resources that live within those virtual environments without requiring the virtual environments to run Pacemaker or Corosync.

For the use case of managing both virtual machines as cluster resources plus the resources that live within the VMs, you can now use the following setup:

- The “normal” (bare-metal) cluster nodes run the High Availability Extension.
- The virtual machines run the `pacemaker_remote` service (almost no configuration required on the VM's side).
- The cluster stack on the “normal” cluster nodes launches the VMs and connects to the `pacemaker_remote` service running on the VMs to integrate them as remote nodes into the cluster.

As the remote nodes do not have the cluster stack installed, this has the following implications:

- Remote nodes do not take part in quorum.
- Remote nodes cannot become the DC.
- Remote nodes are not bound by the scalability limits (Corosync has a member limit of 16 nodes).

Find more information about the `remote_pacemaker` service, including multiple use cases with detailed setup instructions in *Pacemaker Remote—Extending High*

Availability into Virtual Nodes, available at <http://www.clusterlabs.org/doc/>.

## 4.6 Monitoring System Health

To prevent a node from running out of disk space and thus being unable to manage any resources that have been assigned to it, the High Availability Extension provides a resource agent, `ocf:pacemaker:SysInfo`. Use it to monitor a node's health with regard to disk partitions. The SysInfo RA creates a node attribute named `#health_disk` which will be set to `red` if any of the monitored disks' free space is below a specified limit.

To define how the CRM should react in case a node's health reaches a critical state, use the global cluster option `node-health-strategy`.

### **Procedure 4.2:** *Configuring System Health Monitoring*

To automatically move resources away from a node in case the node runs out of disk space, proceed as follows:

#### **1** Configure an `ocf:pacemaker:SysInfo` resource:

```
primitive sysinfo ocf:pacemaker:SysInfo \  
  params disks="/tmp /var"❶ min_disk_free="100M"❷ disk_unit="M"❸ \  
  op monitor interval="15s"
```

- ❶** Which disk partitions to monitor. For example, `/tmp`, `/usr`, `/var`, and `/dev`. To specify multiple partitions as attribute values, separate them with a blank.

---

#### **NOTE: / File System Always Monitored**

You do not need to specify the root partition (`/`) in `disks`. It is always monitored by default.

- ❷** The minimum free disk space required for those partitions. Optionally, you can specify the unit to use for measurement (in the example above, `M` for megabytes is used). If not specified, `min_disk_free` defaults to the unit defined in the `disk_unit` parameter.
- ❸** The unit in which to report the disk space.

**2** To complete the resource configuration, create a clone of `ocf:pacemaker:SysInfo` and start it on each cluster node.

**3** Set the `node-health-strategy` to `migrate-on-red`:

```
property node-health-strategy="migrate-on-red"
```

In case of a `#health_disk` attribute set to `red`, the policy engine adds `-INF` to the resources' score for that node. This will cause any resources to move away from this node. The STONITH resource will be the last one to be stopped but even if the STONITH resource is not running anymore, the node can still be fenced.

Fencing has direct access to the CIB and will continue to work.

After a node's health status has turned to `red`, solve the issue that led to the problem. Then clear the `red` status to make the node eligible again for running resources. Log in to the cluster node and use one of the following methods:

- Execute the following command:

```
root # crm node status-attr NODE delete #health_disk
```

- Restart Pacemaker and Corosync on that node.
- Reboot the node.

The node will be returned to service and can run resources again.

## 4.7 Maintenance Mode

Every now and then, you need to perform testing or maintenance tasks on individual cluster components or the whole cluster—be it changing the cluster configuration, updating software packages for individual nodes, or upgrading the cluster to a higher product version.

With regard to that, High Availability Extension provides `maintenance` options on several levels:

- for resources
- for nodes
- for the whole cluster

---

## **WARNING: Risk of Data Loss**

If you need to execute any testing or maintenance tasks while services are running under cluster control, make sure to follow this outline:

1. Before you start, set the individual resource, node or the whole cluster to maintenance mode. This helps to avoid unwanted side effects like resources not starting in an orderly fashion, the risk of unsynchronized CIBs across the cluster nodes or data loss.
2. Execute your maintenance task or tests.
3. After you have finished, remove the maintenance mode to start normal cluster operation.

---

In maintenance mode, you can stop or restart cluster resources at will—the High Availability Extension will not attempt to restart them. All resources automatically become unmanaged: The High Availability Extension will cease monitoring them and thus be oblivious to their status. You can even stop all Pacemaker services on a node, and all daemons and processes originally started as Pacemaker-managed cluster resources will continue to run. If you attempt to start Pacemaker services on a node while the cluster is in maintenance mode, Pacemaker will initiate a single one-shot monitor operation (a “probe”) for every resource to evaluate which resources are currently running on that node. However, it will take no further action other than determining the resources' status.

For details on setting or unsetting maintenance mode with your preferred cluster management tool:

- Hawk: Section 5.4.5, “Using Maintenance Mode” (page 123)
- crmsh: Section 7.4.6, “Using Maintenance Mode” (page 193)

## **4.8 For More Information**

<http://crmsh.github.io/>

Home page of the crm shell (crmsh), the advanced command line interface for High Availability cluster management.

<http://crmsh.github.io/documentation>

Holds several documents about the crm shell, including a *Getting Started* tutorial for basic cluster setup with crmsh and the comprehensive *Manual* for the crm shell. The latter is available at <http://crmsh.github.io/man-2.0/>. Find the tutorial at <http://crmsh.github.io/start-guide/>.

<http://clusterlabs.org/>

Home page of Pacemaker, the cluster resource manager shipped with the High Availability Extension.

<http://www.clusterlabs.org/doc/>

Holds several comprehensive manuals and some shorter documents explaining general concepts. For example:

- *Pacemaker Explained*: Contains comprehensive and very detailed information for reference.
- *Configuring Fencing with crmsh*: How to configure and use STONITH devices.
- *Colocation Explained*
- *Ordering Explained*

<http://linux-ha.org>

Home page of the The High Availability Linux Project.



# Configuring and Managing Cluster Resources (Web Interface)

In addition to the `crm` command line tool and the Pacemaker GUI, the High Availability Extension also comes with the HA Web Konsole (Hawk), a Web-based user interface for management tasks. It allows you to monitor and administer your Linux cluster from non-Linux machines as well. Furthermore, it is the ideal solution in case your system only provides a minimal graphical user interface.

This chapter introduces Hawk and covers basic tasks for configuring and managing cluster resources: modifying global cluster options, creating basic and advanced types of resources (groups and clones), configuring constraints, specifying failover nodes and failback nodes, configuring resource monitoring, starting, cleaning up or removing resources, and migrating resources manually. For detailed analysis of the cluster status, Hawk generates a cluster report (`hb_report`). You can view the cluster history or explore potential failure scenarios with the simulator.

## 5.1 Hawk—Overview

Hawk's Web interface allows you to monitor and administer your Linux cluster from non-Linux machines as well. Furthermore, it is the ideal solution in case your system only provides a minimal graphical user interface.

### 5.1.1 Starting Hawk and Logging In

The Web interface is included in the `hawk` package. It must be installed on all cluster nodes you want to connect to with Hawk. On the machine from which you want to

access a cluster node using Hawk, you only need a (graphical) Web browser with JavaScript and cookies enabled to establish the connection.

To use Hawk, the respective Web service must be started on the node that you want to connect to via the Web interface.

If you have set up your cluster with the scripts from the `sleha-bootstrap` package, the Hawk service is already started. In that case, skip Procedure 5.1, “Starting Hawk Services” (page 92) and proceed with Procedure 5.2, “Logging In to the Hawk Web Interface” (page 92).

### ***Procedure 5.1: Starting Hawk Services***

**1** On the node you want to connect to, open a shell and log in as `root`.

**2** Check the status of the service by entering

```
root # rchawk status
```

**3** If the service is not running, start it with

```
root # rchawk start
```

If you want Hawk to start automatically at boot time, execute the following command:

```
root # chkconfig hawk on
```

---

### **NOTE: User Authentication**

Hawk users must be members of the `haclient` group. The installation creates a Linux user named `hacluster`, who is added to the `haclient` group. When using the `ha-cluster-init` script for setup, a default password is set for the `hacluster` user.

Before starting Hawk, set or change the password for the `hacluster` user. Alternatively, create a new user which is a member of the `haclient` group.

Do this on every node you will connect to with Hawk.

---

### ***Procedure 5.2: Logging In to the Hawk Web Interface***

The Hawk Web interface uses the HTTPS protocol and port 7630.

---

**NOTE: Accessing Hawk Via a Virtual IP**

To access Hawk even in case the cluster node you usually connect to is down, a virtual IP address (`IPaddr` or `IPaddr2`) can be configured for Hawk as cluster resource. It does not need any special configuration.

---

- 1 On any machine, start a Web browser and make sure that JavaScript and cookies are enabled.
- 2 As URL, enter the IP address or host name of any cluster node running the Hawk Web service. Alternatively, enter the address of any virtual IP address that the cluster operator may have configured as resource:

`https://HOSTNAME_OR_IP_ADDRESS:7630/`

---

**NOTE: Certificate Warning**

If a certificate warning appears when you try to access the URL for the first time, a self-signed certificate is in use. Self-signed certificates are not considered trustworthy by default.

Ask your cluster operator for the certificate details to verify the certificate.

To proceed anyway, you can add an exception in the browser to bypass the warning.

For information on how to replace the self-signed certificate with a certificate signed by an official Certificate Authority, refer to “Replacing the Self-Signed Certificate” (page 135).

---

- 3 On the Hawk login screen, enter the *Username* and *Password* of the `hacluster` user (or of any other user that is a member of the `haclient` group).
- 4 Click *Log In*.

The *Cluster Status* screen appears, displaying the status of your cluster nodes and resources. The information shown is similar to the output of `crm status` in the `crm` shell.

## 5.1.2 Main Screen: Cluster Status

After logging in, Hawk displays the *Cluster Status* screen. It shows a summary with the most important global cluster parameters and the status of your cluster nodes and resources. The following color code is used for status display of nodes and resources:

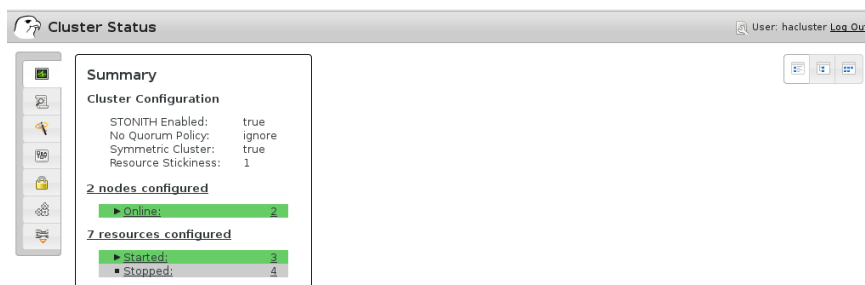
### *Hawk Color Code*

- Green: OK. For example, the resource is running or the node is online.
- Red: Bad, unclear. For example, the resource has failed or the node was not shut down cleanly.
- Yellow: In transition. For example, the node is currently being shut down or a resource is currently being started or stopped. If you click a pending resource to view its details, Hawk also displays the state to which the resource is currently changing (Starting, Stopping, Moving, Promoting, or Demoting).
- Gray: Not running, but the cluster expects it to be running. For example, nodes that the administrator has stopped or put into `standby` mode. Also nodes that are offline are displayed in gray (if they have been shut down cleanly).

In addition to the color code, Hawk also displays icons for the state of nodes, resources, tickets and for error messages in all views of the *Cluster Status* screen.

If a resource has failed, an error message with the details is shown in red at the top of the screen. To analyze the causes for the failure, click the error message. This automatically takes you to Hawk's *History Explorer* and triggers the collection of data for a time span of 20 minutes (10 minutes before and 10 minutes after the failure occurred). For more details, refer to Procedure 5.27, “Viewing Transitions with the History Explorer” (page 126).

**Figure 5.1:** *Hawk—Cluster Status (Summary View)*



The *Cluster Status* screen refreshes itself in near real-time. Choose between the following views, which you can access with the three icons in the upper right corner:

## ***Hawk Cluster Status Views***

### **Summary View**

Shows the most important global cluster parameters and the status of your cluster nodes and resources at the same time. If your setup includes Geo clusters (multi-site clusters), the summary view also shows tickets. To view details about all elements belonging to a certain category (tickets, nodes, or resources), click the category title, which is marked as a link. Otherwise click the individual elements for details.

### **Tree View**

Presents an expandable view of the most important global cluster parameters and the status of your cluster nodes and resources. If your setup includes Geo clusters (multi-site clusters), the tree view also shows tickets. Click the arrows to expand or collapse the elements belonging to the respective category. In contrast to the *Summary View* this view not only shows the IDs and status of resources but also the type (for example, primitive, clone, or group).

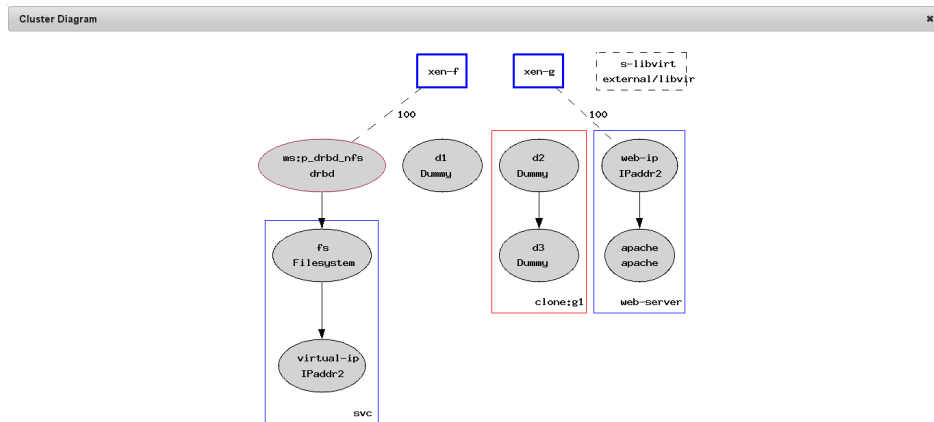
### **Table View**

This view is especially useful for larger clusters, because it shows in a concise way which resources are currently running on which node. Inactive nodes or resources are also displayed.

The top-level row of the main screen shows the user name with which you are logged in. It also allows you to *Log Out* of the Web interface, and to access the following *Tools* from the wrench icon next to the user name:

- *Simulator*. For details, refer to Section 5.4.7, “Exploring Potential Failure Scenarios” (page 128).
- *Cluster Diagram*. Select this entry for a graphical representation of the nodes and the resources configured in the CIB. The diagram also shows the ordering and colocation between resources and node assignments (scores).

**Figure 5.2:** *Hawk—Cluster Diagram*



- *Cluster Report* (`hb_report`). For details, refer to Section 5.4.8, “Generating a Cluster Report” (page 131).

To perform basic operator tasks on nodes and resources (like starting or stopping resources, bringing nodes online, or viewing details), click the wrench icon next to the node or resource. This will display a context menu. For any clone, group or multi-state child resource on any of the status screens, select the *Parent* menu item from the context menu. Clicking this will let you start, stop, etc. the top-level clone or group to which that primitive belongs.

For more complex tasks like configuring resources, constraints, or global cluster options, use the navigation bar on the left hand side. From there, you can access the following screens:

- *Cluster Status*: See Section 5.1.2, “Main Screen: Cluster Status” (page 94) for details.
- *History Explorer*: See Procedure 5.27, “Viewing Transitions with the History Explorer” (page 126) for details.
- *Setup Wizard*: See Section 5.3.1, “Configuring Resources with the Setup Wizard” (page 100) for details.
- *Cluster Configuration*: See Section 5.2, “Configuring Global Cluster Options” (page 98) for details.
- *Access Control Lists*: See Chapter 10, *Access Control Lists* (page 215) for details.
- *Resources*: See Section 5.3, “Configuring Cluster Resources” (page 99) for details.
- *Constraints*: See Section 5.3, “Configuring Cluster Resources” (page 99) for details.

---

**NOTE: Available Functions in Hawk**

By default, users logged in as `root` or `hacluster` have full read-write access to all cluster configuration tasks. However, *Access Control Lists* (page 215) can be used to define fine-grained access permissions.

If ACLs are enabled in the CRM, the available functions in Hawk depend on the user role and access permissions assigned to you. In addition, the following functions in Hawk can only be executed by the user `hacluster`:

- Generating an `hb_report`.
- Using the *History Explorer*.
- Viewing recent events for nodes or resources.

## 5.2 Configuring Global Cluster Options

Global cluster options control how the cluster behaves when confronted with certain situations. They are grouped into sets and can be viewed and modified with cluster management tools like Pacemaker GUI, Hawk, and `crm` shell. The predefined values can be kept in most cases. However, to make key functions of your cluster work correctly, you need to adjust the following parameters after basic cluster setup:

- Option `no-quorum-policy` (page 54)
- Option `stonith-enabled` (page 55)

### **Procedure 5.3:** *Modifying Global Cluster Options*

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2 In the left navigation bar, select *Cluster Properties* to view the global cluster options and their current values. Hawk displays the most important parameters with regard to *CRM Configuration*, *Resource Defaults*, and *Operation Defaults*.

**Figure 5.3:** *Hawk—Cluster Configuration*

The screenshot shows the Hawk Cluster Configuration web interface. The title bar reads "Cluster Configuration" and "User: hacluster Log Out". On the left is a navigation sidebar with icons for Home, Settings, Cluster Properties, and others. The main content area is divided into three sections:

- CRM Configuration**
  - enable-acl: ☐
  - no-quorum-policy:
  - placement-strategy:
  - stonith-enabled: ☐
- Resource Defaults**
  - migration-threshold:
  - resource-stickiness:
- Operation Defaults**
  - record-pending: ☒
  - timeout:

At the bottom of the configuration area are two buttons: "Apply Changes" and "Revert Changes".

- 3 Depending on your cluster requirements, adjust the *CRM Configuration*:

**3a** Set *no-quorum-policy* to the appropriate value.



- 3b** If you need to disable fencing for any reasons, deselect *stonith-enabled*.

---

**IMPORTANT: No Support Without STONITH**

---

A cluster without STONITH is not supported.

---

- 3c** To remove a property from the CRM configuration, click the minus icon next to the property. If a property is deleted, the cluster will behave as if that property had the default value. For details of the default values, refer to Section 4.2.6, “Resource Options (Meta Attributes)” (page 63).
- 3d** To add a new property for the CRM configuration, choose one from the drop-down box and click the plus icon.
- 4** If you need to change *Resource Defaults* or *Operation Defaults*, proceed as follows:
- 4a** To change the value of defaults that are already displayed, edit the value in the respective input field.
- 4b** To add a new resource default or operation default, choose one from the empty drop-down list, click the plus icon and enter a value. If there are default values defined, Hawk proposes them automatically.
- 4c** To remove a resource or operation default, click the minus icon next to the parameter. If no values are specified for *Resource Defaults* and *Operation Defaults*, the cluster uses the default values that are documented in Section 4.2.6, “Resource Options (Meta Attributes)” (page 63) and Section 4.2.8, “Resource Operations” (page 69).
- 5** Confirm your changes.

## 5.3 Configuring Cluster Resources

As a cluster administrator, you need to create cluster resources for every resource or application you run on servers in your cluster. Cluster resources can include Web sites, mail servers, databases, file systems, virtual machines, and any other server-based applications or services you want to make available to users at all times.

For an overview of the resource types you can create, refer to Section 4.2.3, “Types of Resources” (page 58). Apart from the basic specification of a resource (ID, class, provider, and type), you can add or modify the following parameters during or after creation of a resource:

- `Instance attributes (parameters)` determine which instance of a service the resource controls. For more information, refer to Section 4.2.7, “Instance Attributes (Parameters)” (page 67).

When creating a resource, Hawk automatically shows any required parameters. Edit them to get a valid resource configuration.

- `Meta attributes` tell the CRM how to treat a specific resource. For more information, refer to Section 4.2.6, “Resource Options (Meta Attributes)” (page 63).

When creating a resource, Hawk automatically lists the important meta attributes for that resource (for example, the `target-role` attribute that defines the initial state of a resource. By default, it is set to `Stopped`, so the resource will not start immediately).

- `Operations` are needed for resource monitoring. For more information, refer to Section 4.2.8, “Resource Operations” (page 69).

When creating a resource, Hawk displays the most important resource operations (`monitor`, `start`, and `stop`).

## 5.3.1 Configuring Resources with the Setup Wizard

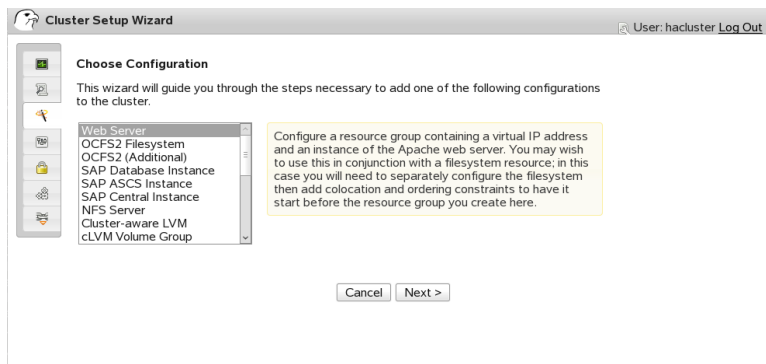
The High Availability Extension comes with a predefined set of templates for some frequently used cluster scenarios, for example, setting up a highly available NFS server. Find the predefined templates in the `hawk-templates` package. You can also define your own wizard templates. For detailed information, refer to <https://github.com/ClusterLabs/hawk/blob/master/doc/wizard.txt>.

### ***Procedure 5.4:*** *Using the Setup Wizard*

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).

- 2 In the left navigation bar, select *Setup Wizard*. The *Cluster Setup Wizard* lists the available resource templates. If you click an entry, Hawk displays a short help text about the template.
- 3 Select the resource set you want to configure and click *Next*.
- 4 Follow the instructions on the screen. If you need information about an option, click it to display a short help text in Hawk.

**Figure 5.4:** *Hawk—Setup Wizard*



## 5.3.2 Creating Simple Cluster Resources

To create the most basic type of resource, proceed as follows:

### **Procedure 5.5:** *Adding Primitive Resources*

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2 In the left navigation bar, select *Resources*. The *Resources* screen shows categories for all types of resources. It lists any resources that are already defined.
- 3 Select the *Primitive* category and click the plus icon.
- 4 Specify the resource:
  - 4a Enter a unique *Resource ID*.

- 4b** From the *Class* list, select the resource agent class you want to use for the resource: *lsb*, *ocf*, *service*, or *stonith*. For more information, see Section 4.2.2, “Supported Resource Agent Classes” (page 57).
- 4c** If you selected *ocf* as class, specify the *Provider* of your OCF resource agent. The OCF specification allows multiple vendors to supply the same resource agent.
- 4d** From the *Type* list, select the resource agent you want to use (for example, *IPaddr* or *Filesystem*). A short description for this resource agent is displayed.

The selection you get in the *Type* list depends on the *Class* (and for OCF resources also on the *Provider*) you have chosen.

- 5** Hawk automatically shows any required parameters for the resource plus an empty drop-down list that you can use to specify an additional parameter.

To define *Parameters* (instance attributes) for the resource:

- 5a** Enter values for each required parameter. A short help text is displayed as soon as you click the text box next to a parameter.
  - 5b** To completely remove a parameter, click the minus icon next to the parameter.
  - 5c** To add another parameter, click the empty drop-down list, select a parameter and enter a value for it.
- 6** Hawk automatically shows the most important resource *Operations* and proposes default values. If you do not modify any settings here, Hawk will add the proposed operations and their default values as soon as you confirm your changes.

For details on how to modify, add or remove operations, refer to Procedure 5.15, “Adding or Modifying Monitor Operations” (page 114).

- 7** Hawk automatically lists the most important meta attributes for the resource, for example `target-role`.

To modify or add *Meta Attributes*:

- 7a** To set a (different) value for an attribute, select one from the drop-down box next to the attribute or edit the value in the input field.
  - 7b** To completely remove a meta attribute, click the minus icon next to it.
  - 7c** To add another meta attribute, click the empty drop-down box and select an attribute. The default value for the attribute is displayed. If needed, change it as described above.
- 8** Click *Create Resource* to finish the configuration. A message at the top of the screen shows if the resource was successfully created or not.

**Figure 5.5:** *Hawk—Primitive Resource*

The screenshot shows the 'Create Resource' web interface. At the top, there's a header with 'Create Resource' and a user profile 'User: hackuster Log Out'. Below the header, there's a table with columns: Resource ID, Class, Provider, and Type. The values are: my\_primitive, ocf, heartbeat, and IPAddr. There's a checkbox 'Use Template' which is unchecked. Below the table, there are three sections: Parameters, Operations, and Meta Attributes. The Parameters section has a table with columns: Parameter, Value, and Action. The first row is 'ip' with value '192.268.1.1' and a minus icon. The second row is 'nic' with value 'eth0' and a plus icon. The Operations section has a table with columns: Operation, Timeout, Interval, and Action. The first row is 'monitor' with timeout '20s' and interval '5s'. The second row is 'start' with timeout '20s'. The third row is 'stop' with timeout '20s'. The Meta Attributes section has a table with columns: Attribute, Value, and Action. The first row is 'target-role' with value 'Stopped'. At the bottom right, there are two buttons: 'Create Resource' and 'Back'. On the right side of the interface, there's a text box with the following text: 'Manages virtual IPv4 addresses (portable version). This script manages IP alias IP addresses. It can add an IP alias, or remove one. nic. Network interface. The base network interface on which the IP address will be brought online. If left empty, the script will try and determine this from the routing table. Do NOT specify an alias interface in the form eth0:1 or anything here; rather, specify the base interface only. Prerequisite: There must be at least one static IP address, which is not managed by the cluster, assigned to the network interface. If you can not assign any static IP address on the interface, modify this kernel parameter: sysctl -w net.ipv4.conf.all.promote\_secondaries= (or per device)'.

## 5.3.3 Creating STONITH Resources

### IMPORTANT: No Support Without STONITH

A cluster without STONITH is not supported.

By default, the global cluster option `stonith-enabled` is set to `true`: If no STONITH resources have been defined, the cluster will refuse to start any resources. Configure one or more STONITH resources to complete the STONITH setup. While they are configured similar to other resources, the behavior of STONITH resources is

different in some respects. For details refer to Section 9.3, “STONITH Resources and Configuration” (page 206).

**Procedure 5.6:** *Adding a STONITH Resource*

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2 In the left navigation bar, select *Resources*. The *Resources* screen shows categories for all types of resources and lists all defined resources.
- 3 Select the *Primitive* category and click the plus icon.
- 4 Specify the resource:
  - 4a Enter a unique *Resource ID*.
  - 4b From the *Class* list, select the resource agent class *stonith*.
  - 4c From the *Type* list, select the STONITH plug-in for controlling your STONITH device. A short description for this plug-in is displayed.
- 5 Hawk automatically shows the required *Parameters* for the resource. Enter values for each parameter.
- 6 Hawk displays the most important resource *Operations* and proposes default values. If you do not modify any settings here, Hawk will add the proposed operations and their default values as soon as you confirm.
- 7 Adopt the default *Meta Attributes* settings if there is no reason to change them.
- 8 Confirm your changes to create the STONITH resource.

To complete your fencing configuration, add constraints, use clones or both. For more details, refer to Chapter 9, *Fencing and STONITH* (page 203).

## 5.3.4 Using Resource Templates

If you want to create lots of resources with similar configurations, defining a resource template is the easiest way. After being defined, it can be referenced in

primitives or in certain types of constraints. For detailed information about function and use of resource templates, refer to Section 4.4.3, “Resource Templates and Constraints” (page 77).

### **Procedure 5.7:** *Creating Resource Templates*

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2 In the left navigation bar, select *Resources*. The *Resources* screen shows categories for all types of resources plus a *Template* category.
- 3 Select the *Template* category and click the plus icon.
- 4 Enter a *Template ID*.
- 5 Specify the resource template as you would specify a primitive. Follow Procedure 5.5: Adding Primitive Resources, starting with Step 4b (page 102).
- 6 Click *Create Resource* to finish the configuration. A message at the top of the screen shows if the resource template was successfully created.

**Figure 5.6:** *Hawk—Resource Template*

**Create Template** User: hacluster [Log Out](#)

Template ID	Class	Provider	Type
<input type="text" value="dummy_template"/>	<input type="text" value="ocf"/>	<input type="text" value="heartbeat"/>	<input type="text" value="Dummy"/>

Operations

monitor	timeout: 20 interval: 10	<input type="button" value="edit"/> <input type="button" value="delete"/>
start	timeout: 20	<input type="button" value="edit"/> <input type="button" value="delete"/>
stop	timeout: 20	<input type="button" value="edit"/> <input type="button" value="delete"/>
<input type="button" value="Add"/>		<input type="button" value="Add"/>

Example stateless resource agent

This is a Dummy Resource Agent. It does absolutely nothing except keep track of whether its running or not. Its purpose in life is for testing and to serve as a template for RA writers. NB: Please pay attention to the timeouts specified in the actions section below. They should be meaningful for the kind of resource the agent manages. They should be the minimum advised timeouts, but they shouldn't/cannot cover all possible resource instances. So, try to be neither overly generous nor too stingy, but moderate. The minimum timeouts should never be below 10 seconds.

**Procedure 5.8:** *Referencing Resource Templates*

- 1** Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2** To reference the newly created resource template in a primitive, follow these steps:
  - 2a** In the left navigation bar, select *Resources*. The *Resources* screen shows categories for all types of resources. It lists all defined resources.
  - 2b** Select the *Primitive* category and click the plus icon.
  - 2c** Enter a unique *Resource ID*.
  - 2d** Activate *Use Template* and, from the drop-down list, select the template to reference.
  - 2e** If needed, specify further *Parameters*, *Operations*, or *Meta Attributes* as described in Procedure 5.5, “Adding Primitive Resources” (page 101).
- 3** To reference the newly created resource template in colocal or order constraints, proceed as described in Procedure 5.10, “Adding or Modifying Colocal or Order Constraints” (page 107).

## 5.3.5 Configuring Resource Constraints

After you have configured all resources, specify how the cluster should handle them correctly. Resource constraints let you specify on which cluster nodes resources can run, in which order resources will be loaded, and what other resources a specific resource depends on.

For an overview of available types of constraints, refer to Section 4.4.1, “Types of Constraints” (page 74). When defining constraints, you also need to specify scores. For more information on scores and their implications in the cluster, see Section 4.4.2, “Scores and Infinity” (page 77).

Learn how to create the different types of constraints in the following procedures.

**Procedure 5.9:** *Adding or Modifying Location Constraints*

For location constraints, specify a constraint ID, resource, score and node:



- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2 In the left navigation bar, select *Constraints*. The *Constraints* screen shows categories for all types of constraints. It lists all defined constraints.
- 3 To add a new *Location* constraint, click the plus icon in the respective category.  
  
To modify an existing constraint, click the wrench icon next to the constraint and select *Edit Constraint*.
- 4 Enter a unique *Constraint ID*. When modifying existing constraints, the ID is already defined.
- 5 Select the *Resource* for which to define the constraint. The list shows the IDs of all resources that have been configured for the cluster.
- 6 Set the *Score* for the constraint. Positive values indicate the resource can run on the *Node* you specify in the next step. Negative values mean it should not run on that node. Setting the score to `INFINITY` forces the resource to run on the node. Setting it to `-INFINITY` means the resources must not run on the node.
- 7 Select the *Node* for the constraint.
- 8 Click *Create Constraint* to finish the configuration. A message at the top of the screen shows if the constraint was successfully created.

**Figure 5.7:** Hawk—Location Constraint

The screenshot shows a web interface titled "Create Location Constraint". At the top right, it says "User: hacuster Log Out". On the left is a vertical navigation bar with icons. The main form area contains the following elements:

- Constraint ID:** A text input field containing "my\_loc\_constraint".
- Resource:** A dropdown menu showing "my\_primitive".
- Show Rule Editor:** An unchecked checkbox.
- Score:** A text input field containing "INFINITY".
- Node:** A dropdown menu showing "barett-2".
- Buttons:** "Create Constraint" and "Back" buttons at the bottom right.

**Procedure 5.10:** Adding or Modifying Colocational or Order Constraints

For both types of constraints specify a constraint ID and a score, then add resources to a dependency chain:

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2 In the left navigation bar, select *Constraints*. The *Constraints* screen shows categories for all types of constraints and lists all defined constraints.
- 3 To add a new *Colocation* or *Order* constraint, click the plus icon in the respective category.

To modify an existing constraint, click the wrench icon next to the constraint and select *Edit Constraint*.

- 4 Enter a unique *Constraint ID*. When modifying existing constraints, the ID is already defined.
- 5 Define a *Score*.

For colocation constraints, the score determines the location relationship between the resources. Positive values indicate the resources should run on the same node. Negative values indicate the resources should not run on the same node. Setting the score to `INFINITY` forces the resources to run on the same node. Setting it to `-INFINITY` means the resources must not run on the same node. The score will be combined with other factors to decide where to put the resource.

For order constraints, the constraint is mandatory if the score is greater than zero, otherwise it is only a suggestion. The default value is `INFINITY`.

- 6 For order constraints, you can usually keep the option *Symmetrical* enabled. This specifies that resources are stopped in reverse order.
- 7 To define the resources for the constraint, follow these steps:
  - 7a Select a resource from the list *Add resource to constraint*. The list shows the IDs of all resources and all resource templates configured for the cluster.
  - 7b To add the selected resource, click the plus icon next to the list. A new list appears beneath. Select the next resource from the list. As both colocation and order constraints define a dependency between resources, you need at least two resources.
  - 7c Select one of the remaining resources from the list *Add resource to constraint*. Click the plus icon to add the resource.

Now you have two resources in a dependency chain.

If you have defined an order constraint, the topmost resource will start first, then the second etc. Usually the resources will be stopped in reverse order.

However, if you have defined a colocation constraint, the arrow icons between the resources reflect their dependency, but *not* their start order. As the topmost resource depends on the next resource and so on, the cluster will first decide where to put the last resource, then place the depending ones based on that decision. If the constraint cannot be satisfied, the cluster may decide not to allow the dependent resource to run.

- 7d** Add as many resources as needed for your colocation or order constraint.
- 7e** If you want to swap the order of two resources, click the double arrow at the right hand side of the resources to swap the resources in the dependency chain.
- 8** If needed, specify further parameters for each resource, like the role (Master, Slave, Started, or Stopped).
- 9** Click *Create Constraint* to finish the configuration. A message at the top of the screen shows if the constraint was successfully created.

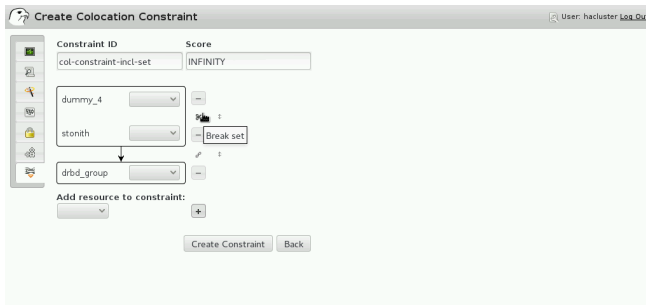
**Figure 5.8:** *Hawk—Colocation Constraint*

The screenshot shows the 'Create Colocation Constraint' web interface. At the top, there's a header with a user icon and 'User: hacluster Log Out'. The main form has two input fields: 'Constraint ID' with the value 'col-r0-on-drbd' and 'Score' with the value 'INFINITY'. Below these, there's a section for resources. It shows a dropdown menu for 'drbd\_group' and a dropdown menu for 'ms\_drbd' with the role 'Master' selected. An arrow points from 'drbd\_group' to 'ms\_drbd'. Below this, there's an 'Add resource to constraint:' section with a plus button. At the bottom right, there are two buttons: 'Create Constraint' and 'Back'.

As an alternative format for defining colocation or ordering constraints, you can use resource sets. They have the same ordering semantics as groups.

**Procedure 5.11:** *Using Resource Sets for Colocation or Order Constraints*

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2 Define colocation or order constraints as described in Procedure 5.10, “Adding or Modifying Colocational or Order Constraints” (page 107).
- 3 When you have added the resources to the dependency chain, you can put them into a resource set by clicking the chain icon at the right hand side. A resource set is visualized by a frame around the resources belonging to a set.
- 4 You can also add multiple resources to a resource set or create multiple resource sets.
- 5 To extract a resource from a resource set, click the scissors icon above the respective resource.



The resource will be removed from the set and put back into the dependency chain at its original place.

- 6 Confirm your changes to finish the constraint configuration.

For more information on configuring constraints and detailed background information about the basic concepts of ordering and colocation, refer to the documentation available at <http://www.clusterlabs.org/doc/>:

- *Pacemaker Explained*, chapter *Resource Constraints*
- *Colocation Explained*
- *Ordering Explained*

### **Procedure 5.12:** *Removing Constraints*

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2 In the left navigation bar, select *Constraints*. The *Constraints* screen shows categories for all types of constraints and lists all defined constraints.
- 3 Click the wrench icon next to a constraint and select *Remove Constraint*.

## **5.3.6 Specifying Resource Failover Nodes**

A resource will be automatically restarted if it fails. If that cannot be achieved on the current node, or it fails N times on the current node, it will try to fail over to another node. You can define a number of failures for resources (a `migration-threshold`), after which they will migrate to a new node. If you have more than two nodes in your cluster, the node to which a particular resource fails over is chosen by the High Availability software.

You can specify a specific node to which a resource will fail over by proceeding as follows:

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2 Configure a location constraint for the resource as described in Procedure 5.9, “Adding or Modifying Location Constraints” (page 106).
- 3 Add the `migration-threshold` meta attribute to the resource as described in Procedure 5.5: Adding Primitive Resources, Step 7 (page 102) and enter a *Value* for the migration-threshold. The value should be positive and less than INFINITY.
- 4 If you want to automatically expire the failcount for a resource, add the `failure-timeout` meta attribute to the resource as described in Procedure 5.5: Adding Primitive Resources, Step 7 (page 102) and enter a *Value* for the failure-timeout.
- 5 If you want to specify additional failover nodes with preferences for a resource, create additional location constraints.

The process flow regarding migration thresholds and failcounts is demonstrated in Example 4.6, “Migration Threshold—Process Flow” (page 78).

Instead of letting the failcount for a resource expire automatically, you can also clean up failcounts for a resource manually at any time. Refer to Section 5.4.2, “Cleaning Up Resources” (page 120) for details.

## 5.3.7 Specifying Resource Failback Nodes (Resource Stickiness)

A resource may fail back to its original node when that node is back online and in the cluster. To prevent this or to specify a different node for the resource to fail back to, change the stickiness value of the resource. You can either specify the resource stickiness when creating it or afterwards.

For the implications of different resource stickiness values, refer to Section 4.4.5, “Failback Nodes” (page 79).

### ***Procedure 5.13:*** *Specifying Resource Stickiness*

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2 Add the `resource-stickiness` meta attribute to the resource as described in Procedure 5.5: Adding Primitive Resources, Step 7 (page 102).
- 3 Specify a value between `-INFINITY` and `INFINITY` for the resource-stickiness.

## 5.3.8 Configuring Placement of Resources Based on Load Impact

Not all resources are equal. Some, such as Xen guests, require that the node hosting them meets their capacity requirements. If resources are placed so that their combined needs exceed the provided capacity, the performance of the resources diminishes or they fail.

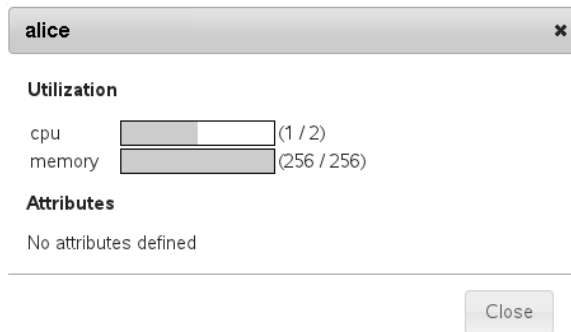
To take this into account, the High Availability Extension allows you to specify the following parameters:

1. The capacity a certain node *provides*.
2. The capacity a certain resource *requires*.
3. An overall strategy for placement of resources.

Utilization attributes are used to configure both the resource's requirements and the capacity a node provides. The High Availability Extension now also provides means to detect and configure both node capacity and resource requirements automatically. For more details and a configuration example, refer to Section 4.4.6, “Placing Resources Based on Their Load Impact” (page 80).

To display a node's capacity values (defined via utilization attributes) and the capacity currently consumed by resources running on the node, switch to the *Cluster Status* screen in Hawk. Select the node you are interested in, click the wrench icon next to the node and select *Show Details*.

**Figure 5.9:** *Hawk—Viewing a Node's Capacity Values*



After you have configured the capacities your nodes provide and the capacities your resources require, you need to set the placement strategy in the global cluster options. Otherwise the capacity configurations have no effect. Several strategies are available to schedule the load: for example, you can concentrate it on as few nodes as possible, or balance it evenly over all available nodes. For more information, refer to Section 4.4.6, “Placing Resources Based on Their Load Impact” (page 80).

**Procedure 5.14:** *Setting the Placement Strategy*

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).

- 2 In the left navigation bar, select *Cluster Properties* to view the global cluster options and their current values.
- 3 From the *Add new property* drop-down list, choose `placement-strategy`.
- 4 Depending on your requirements, set *Placement Strategy* to the appropriate value.
- 5 Click the plus icon to add the new cluster property including its value.
- 6 Confirm your changes.

## 5.3.9 Configuring Resource Monitoring

The High Availability Extension can not only detect a node failure, but also when an individual resource on a node has failed. If you want to ensure that a resource is running, configure resource monitoring for it. For resource monitoring, specify a timeout and/or start delay value, and an interval. The interval tells the CRM how often it should check the resource status. You can also set particular parameters, such as `Timeout` for `start` or `stop` operations.

### ***Procedure 5.15:*** *Adding or Modifying Monitor Operations*

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2 In the left navigation bar, select *Resources*. The *Resources* screen shows categories for all types of resources and lists all defined resources.
- 3 Select the resource to modify, click the wrench icon next to it and select *Edit Resource*. The resource definition is displayed. Hawk automatically shows the most important resource operations (`monitor`, `start`, `stop`) and proposes default values.
- 4 To change the values for an operation:
  - 4a Click the pen icon next to the operation.
  - 4b In the dialog that opens, specify the following values:
    - Enter a `timeout` value in seconds. After the specified timeout period, the operation will be treated as `failed`. The PE will decide what to



do or execute what you specified in the *On Fail* field of the monitor operation.

- For monitoring operations, define the monitoring *interval* in seconds.

If needed, use the empty drop-down box at the bottom of the *monitor* dialog to add more parameters, like *On Fail* (what to do if this action fails?) or *Requires* (what conditions need to be fulfilled before this action occurs?).

The screenshot shows a dialog box titled "monitor" with a close button (x) in the top right corner. Inside the dialog, there are four input fields with corresponding minus icons to their right:

- interval: 120
- timeout: 15
- start-delay: 15
- on-fail: stop (dropdown menu)

Below the input fields, there is a plus icon to the right of the "on-fail" dropdown. At the bottom of the dialog, there are "OK" and "Cancel" buttons.

**4c** Confirm your changes to close the dialog and to return to the *Edit Resource* screen.

- 5** To completely remove an operation, click the minus icon next to it.
- 6** To add another operation, click the empty drop-down box and select an operation. A default value for the operation is displayed. If needed, change it by clicking the pen icon.
- 7** Click *Apply Changes* to finish the configuration. A message at the top of the screen shows if the resource was successfully updated or not.

For the processes which take place if the resource monitor detects a failure, refer to Section 4.3, “Resource Monitoring” (page 72).

To view resource failures, switch to the *Cluster Status* screen in Hawk and select the resource you are interested in. Click the wrench icon next to the resource and select *Show Details*.

## 5.3.10 Configuring a Cluster Resource Group

Some cluster resources depend on other components or resources and require that each component or resource starts in a specific order and runs on the same server. To simplify this configuration we support the concept of groups.

For an example of a resource group and more information about groups and their properties, refer to Section 4.2.5.1, “Groups” (page 60).

---

### **NOTE: Empty Groups**

Groups must contain at least one resource, otherwise the configuration is not valid. In Hawk, primitives cannot be created or modified while creating a group. Before adding a group, create primitives and configure them as desired. For details, refer to Procedure 5.5, “Adding Primitive Resources” (page 101).

---

### **Procedure 5.16:** *Adding a Resource Group*

- 1** Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2** In the left navigation bar, select *Resources*. The *Resources* screen shows categories for all types of resources and lists all defined resources.
- 3** Select the *Group* category and click the plus icon.
- 4** Enter a unique *Group ID*.
- 5** To define the group members, select one or multiple entries in the list of *Available Primitives* and click the < icon to add them to the *Group Children* list. Any new group members are added to the bottom of the list. To define the order of the group members, you currently need to add and remove them in the order you desire.
- 6** If needed, modify or add *Meta Attributes* as described in Adding Primitive Resources, Step 7 (page 102).
- 7** Click *Create Group* to finish the configuration. A message at the top of the screen shows if the group was successfully created.

**Figure 5.10:** *Hawk—Resource Group*

The screenshot shows the 'Create Group' interface in the Hawk web console. At the top, there's a header with the Hawk logo and 'Create Group' title, and a user status 'User: hacluster Log Out'. The main area is divided into several sections: 'Group ID' with a text input containing 'g-webserver'; 'Group Children' with a list box containing 'my\_filesystem', 'my\_ipaddress', and 'my\_webserver'; 'Available Primitives' with a list box containing 'stonith'; and 'Meta Attributes' with a dropdown for 'target-role' set to 'Stopped'. At the bottom, there are 'Create Group' and 'Back' buttons.

## 5.3.11 Configuring a Clone Resource

If you want certain resources to run simultaneously on multiple nodes in your cluster, configure these resources as a clones. For example, cloning makes sense for resources like STONITH and cluster file systems like OCFS2. You can clone any resource provided. Cloning is supported by the resource’s Resource Agent. Clone resources may be configured differently depending on which nodes they are running on.

For an overview of the available types of resource clones, refer to Section 4.2.5.2, “Clones” (page 62).

---

### **NOTE:** Sub-resources for Clones

Clones can either contain a primitive or a group as sub-resources. In Hawk, sub-resources cannot be created or modified while creating a clone. Before adding a clone, create sub-resources and configure them as desired. For details, refer to Procedure 5.5, “Adding Primitive Resources” (page 101) or Procedure 5.16, “Adding a Resource Group” (page 116).

---

### **Procedure 5.17:** *Adding or Modifying Clones*

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).

- 2 In the left navigation bar, select *Resources*. The *Resources* screen shows categories for all types of resources and lists all defined resources.
- 3 Select the *Clone* category and click the plus icon.
- 4 Enter a unique *Clone ID*.
- 5 From the *Child Resource* list, select the primitive or group to use as a sub-resource for the clone.
- 6 If needed, modify or add *Meta Attributes* as described in Procedure 5.5: Adding Primitive Resources, Step 7 (page 102).
- 7 Click *Create Clone* to finish the configuration. A message at the top of the screen shows if the clone was successfully created.

**Figure 5.11:** Hawk—Clone Resource

The screenshot shows the 'Create Clone' window in the Hawk interface. On the left is a vertical navigation bar with icons for Home, Resources, Clones, and other functions. The main area contains the following fields and controls:

- Clone ID:** A text input field containing 'cl-dlm'.
- Child Resource:** A dropdown menu with 'dlm' selected.
- Meta Attributes:** A section with a minus sign icon and a '+' icon. It contains:
  - target-role:** A dropdown menu with 'Started' selected.
  - interleave:** A dropdown menu with 'interleave' selected.
  - A checkbox that is checked.
- Buttons:** 'Create Clone' and 'Back' buttons at the bottom right.

In the top right corner, there is a user status bar showing 'User: hacluster' and a 'Log Out' link.

## 5.4 Managing Cluster Resources

In addition to configuring your cluster resources, Hawk allows you to manage existing resources from the *Cluster Status* screen. For a general overview of the screen, its different views and the color code used for status information, refer to Section 5.1.2, “Main Screen: Cluster Status” (page 94).

Basic resource operations can be executed from any cluster status view. Both *Tree View* and *Table View* let you access the individual resources directly. However,

in the *Summary View* you need to click the links in the resources category first to display the resource details. The detailed view also shows any attributes set for that resource. For primitive resources (regular primitives, children of groups, clones, or multi-state resources), the following information will be shown additionally:

- the resource's failcount
- the last failure time stamp (if the failcount is > 0)
- operation history and timings (call id, operation, last run time stamp, execution time, queue time, return code and last change time stamp)

**Figure 5.12:** *Viewing a Resource's Details*

stonith

Attributes

target-role Started

barett-1

Fail Count 1000000

Last Failure 2014-08-04 17:00:56

Call ID	Operation	Last Run	Exec	Queue	RC	Last Change
43	start	2014-08-04 17:00:52	3628ms	0ms	1	2014-08-04 17:00:52
48	stop	2014-08-04 17:00:56	1ms	0ms	0	2014-08-04 17:00:56

barett-2

Fail Count 1000000

Last Failure 2014-08-04 17:01:00

Call ID	Operation	Last Run	Exec	Queue	RC	Last Change
48	start	2014-08-04 17:00:56	3425ms	0ms	1	2014-08-04 17:00:56
49	stop	2014-08-04 17:01:01	0ms	0ms	0	2014-08-04 17:01:01

Close

## 5.4.1 Starting Resources

Before you start a cluster resource, make sure it is set up correctly. For example, if you want to use an Apache server as a cluster resource, set up the Apache server first. Complete the Apache configuration before starting the respective resource in your cluster.

---

### NOTE: Do Not Touch Services Managed by the Cluster

When managing a resource via the High Availability Extension, the same resource must not be started or stopped otherwise (outside of the cluster,

for example manually or on boot or reboot). The High Availability Extension software is responsible for all service start or stop actions.

However, if you want to check if the service is configured properly, start it manually, but make sure that it is stopped again before High Availability takes over.

For interventions in resources that are currently managed by the cluster, set the resource to `maintenance mode` first as described in Procedure 5.23, “Applying Maintenance Mode to Resources” (page 124).

---

When creating a resource with Hawk, you can set its initial state with the `target-role` meta attribute. If you set its value to `stopped`, the resource does not start automatically after being created.

**Procedure 5.18:** *Starting A New Resource*

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2 In the left navigation bar, select *Cluster Status*.
- 3 In one of the individual resource views, click the wrench icon next to the resource and select *Start*. To continue, confirm the message that appears. As soon as the resource has started, Hawk changes the resource's color to green and shows on which node it is running.

## 5.4.2 Cleaning Up Resources

A resource will be automatically restarted if it fails, but each failure increases the resource's failcount.

If a `migration-threshold` has been set for the resource, the node will no longer run the resource when the number of failures reaches the migration threshold.

A resource's failcount can either be reset automatically (by setting a `failure-timeout` option for the resource) or you can reset it manually as described below.

**Procedure 5.19:** *Cleaning Up A Resource*

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).

- 2 In the left navigation bar, select *Cluster Status*.
- 3 In one of the individual resource views, click the wrench icon next to the failed resource and select *Clean Up*. To continue, confirm the message that appears.

This executes the commands `crm_resource -C` and `crm_failcount -D` for the specified resource on the specified node.

For more information, see the man pages of `crm_resource` and `crm_failcount`.

## 5.4.3 Removing Cluster Resources

If you need to remove a resource from the cluster, follow the procedure below to avoid configuration errors:

### ***Procedure 5.20: Removing a Cluster Resource***

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2 In the left navigation bar, select *Cluster Status*.
- 3 Clean up the resource on all nodes as described in Procedure 5.19, “Cleaning Up A Resource” (page 120).
- 4 In one of the individual resource views, click the wrench icon next to the resource and select *Stop*. To continue, confirm the message that appears.
- 5 If the resource is stopped, click the wrench icon next to it and select *Delete Resource*.

## 5.4.4 Migrating Cluster Resources

As mentioned in Section 5.3.6, “Specifying Resource Failover Nodes” (page 111), the cluster will fail over (migrate) resources automatically in case of software or hardware failures—according to certain parameters you can define (for example, migration threshold or resource stickiness). Apart from that, you can manually

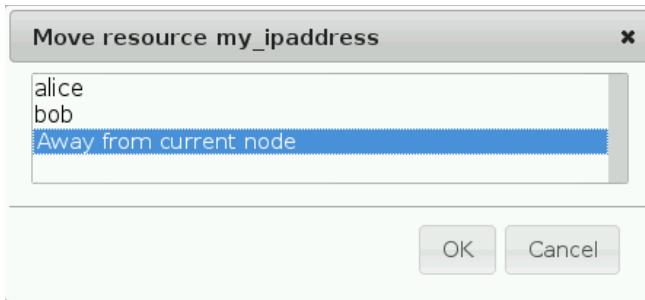
migrate a resource to another node in the cluster. Or you decide to move it away from the current node and leave the decision about where to put it to the cluster.

**Procedure 5.21:** *Manually Migrating a Resource*

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2 In the left navigation bar, select *Cluster Status*.
- 3 In one of the individual resource views, click the wrench icon next to the resource and select *Move*.
- 4 In the new window, select the node to which to move the resource.

This creates a location constraint with an INFINITY score for the destination node.

- 5 Alternatively, select to move the resource *Away from current node*.



This creates a location constraint with a -INFINITY score for the current node.

- 6 Click *OK* to confirm the migration.

To allow a resource to move back again, proceed as follows:

**Procedure 5.22:** *Clearing a Migration Constraint*

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2 In the left navigation bar, select *Cluster Status*.



- 3 In one of the individual resource views, click the wrench icon next to the resource and select *Drop Relocation Rule*. To continue, confirm the message that appears.

This uses the `crm_resource -U` command. The resource *can* move back to its original location or it may stay where it is (depending on resource stickiness).

For more information, see the `crm_resource` man page or *Pacemaker Explained*, available from <http://www.clusterlabs.org/doc/>. Refer to section *Resource Migration*.

## 5.4.5 Using Maintenance Mode

Every now and then, you need to perform testing or maintenance tasks on individual cluster components or the whole cluster—be it changing the cluster configuration, updating software packages for individual nodes, or upgrading the cluster to a higher product version.

With regard to that, High Availability Extension provides `maintenance` options on several levels:

- Applying Maintenance Mode to Resources (page 124)
- Applying Maintenance Mode to Nodes (page 124)
- Applying Maintenance Mode to the Cluster (page 125)

---

### **WARNING: Risk of Data Loss**

If you need to execute any testing or maintenance tasks while services are running under cluster control, make sure to follow this outline:

1. Before you start, set the individual resource, node or the whole cluster to maintenance mode. This helps to avoid unwanted side effects like resources not starting in an orderly fashion, the risk of unsynchronized CIBs across the cluster nodes or data loss.
  2. Execute your maintenance task or tests.
  3. After you have finished, remove the maintenance mode to start normal cluster operation.
-

For more details on what happens to the resources and the cluster while in maintenance mode, see Section 4.7, “Maintenance Mode” (page 88).

**Procedure 5.23:** *Applying Maintenance Mode to Resources*

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2 In the left navigation bar, select *Resources*. Select the resource you want to put in maintenance mode or unmanaged mode, click the wrench icon next to the resource and select *Edit Resource*.
- 3 Open the *Meta Attributes* category.
- 4 From the empty drop-down list, select the *maintenance* attribute and click the plus icon to add it.
- 5 Activate the check box next to `maintenance` to set the maintenance attribute to `yes`.
- 6 Confirm your changes.
- 7 After you have finished the maintenance task for that resource, deactivate the check box next to the `maintenance` attribute for that resource.

From this point on, the resource will be managed by the High Availability Extension software again.

**Procedure 5.24:** *Applying Maintenance Mode to Nodes*

Sometimes it is necessary to put single nodes into maintenance mode.

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2 In the left navigation bar, select *Cluster Status*.
- 3 In one of the individual nodes' views, click the wrench icon next to the node and select *Maintenance*.

This will add the following instance attribute to the node:  
`maintenance="true"`. The resources previously running on the maintenance-

mode node will become `unmanaged`. No new resources will be allocated to the node until it leaves the maintenance mode.

- 4 To deactivate the maintenance mode, click the wrench icon next to the node and select *Ready*.

**Procedure 5.25:** *Applying Maintenance Mode to the Cluster*

For setting or unsetting the maintenance mode for the whole cluster, proceed as follows:

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2 In the left navigation bar, select *Cluster Configuration*.
- 3 In the *CRM Configuration* group, select the *maintenance-mode* attribute from the empty drop-down box and click the plus icon to add it.
- 4 To set `maintenance-mode=true`, activate the check box next to `maintenance-mode` and confirm your changes.
- 5 After you have finished the maintenance task for the whole cluster, deactivate the check box next to the `maintenance-mode` attribute.

From this point on, High Availability Extension will take over cluster management again.

## 5.4.6 Viewing the Cluster History

Hawk provides the following possibilities to view past events on the cluster (on different levels and in varying detail).

**Procedure 5.26:** *Viewing Recent Events of Nodes or Resources*

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2 In the left navigation bar, select *Cluster Status*.
- 3 In the *Tree View* or *Table View*, click the wrench icon next to the resource or node you are interested in and select *View Recent Events*.

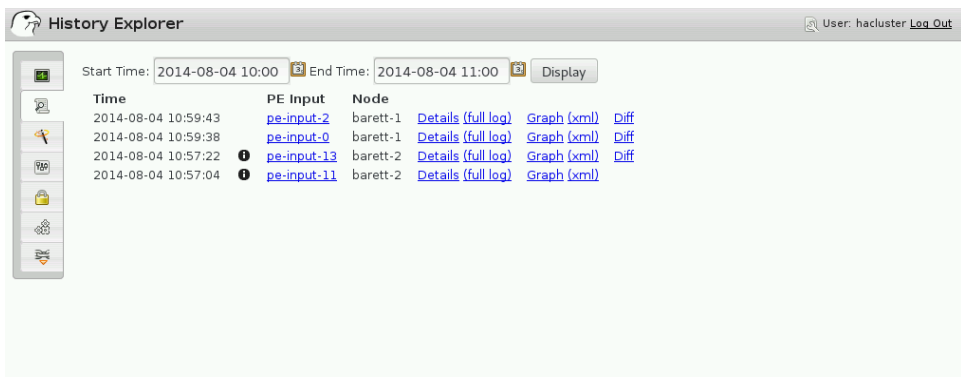
The dialog that opens shows the events of the last hour.

**Procedure 5.27:** *Viewing Transitions with the History Explorer*

The *History Explorer* provides transition information for a time frame that you can define. It also lists its previous runs and allows you to *Delete* reports that you no longer need. The *History Explorer* uses the information provided by `hb_report`.

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2 In the left navigation bar, select *History Explorer*.
- 3 By default, the period to explore is set to the last 24 hours. To modify this, set another *Start Time* and *End Time*.
- 4 Click *Display* to start collecting transition data.

**Figure 5.13:** *Hawk—History Report*



The following information is displayed:

**History Explorer Results**

*Time*

The time line of all past transitions in the cluster.

*PE Input/Node*

The `pe-input*` file for each transition and the node on which it was generated. For each transition, the cluster saves a copy of the state which is

provided to the policy engine as input. The path to this archive is logged. The `pe-input*` files are only generated on the Designated Coordinator (DC), but as the DC can change, there may be `pe-input*` files from several nodes. The files show what the Policy Engine (PE) *planned* to do.

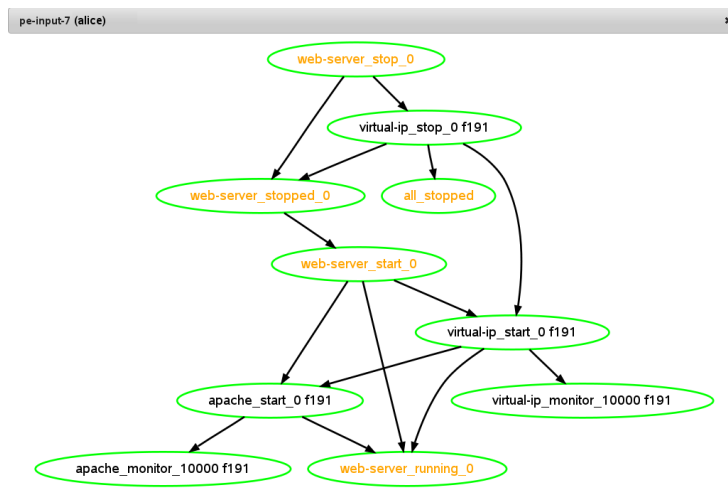
### Details/Full Log

Opens a pop-up window with snippets of logging data that belong to that particular transition. Different amounts of details are available: Clicking *Details* displays the output of `crm history transition peinput` (including the resource agents' log messages). *Full Log* also includes details from the `pengine`, `crmd`, and `lrmd` and is equivalent to `crm history transition log peinput`.

### Graph/XML

A graph and an XML representation of each transition. If you choose to show the *Graph*, the PE is reinvoked (using the `pe-input*` files), and generates a graphical visualization of the transition. Alternatively, you can view the XML representation of the graph.

**Figure 5.14:** Hawk History Report—Transition Graph



### Diff

If two or more `pe-inputs` are listed, a *Diff* link will appear to the right of each pair of `pe-inputs`. Clicking it displays the difference of configuration and status.

## 5.4.7 Exploring Potential Failure Scenarios

Hawk provides a *Simulator* that allows you to explore failure scenarios before they happen. After switching to the simulator mode, you can change the status of nodes, add or edit resources and constraints, change the cluster configuration, or execute multiple resource operations to see how the cluster would behave should these events occur. As long as the simulator mode is activated, a control dialog will be displayed in the bottom right hand corner of the *Cluster Status* screen. The simulator will collect the changes from all screens and will add them to its internal queue of events. The simulation run with the queued events will not be executed unless it is manually triggered in the control dialog. After the simulation run, you can view and analyze the details of what would have happened (log snippets, transition graph, and CIB states).

### **Procedure 5.28:** *Switching to Simulator Mode*

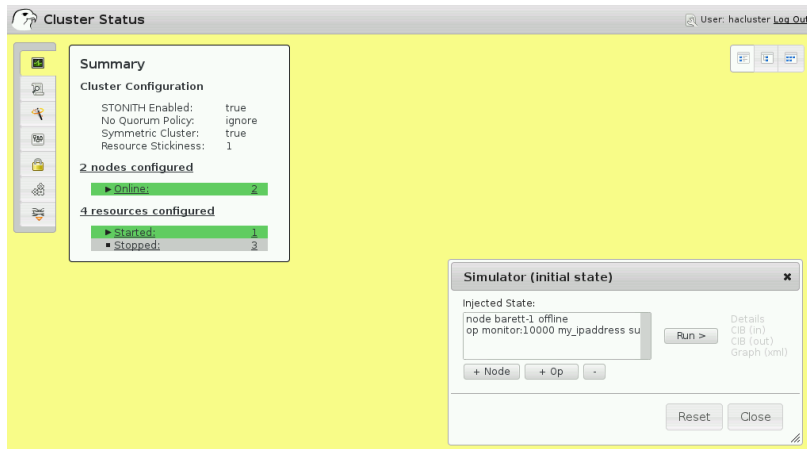
- 1** Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2** Activate the simulator mode by clicking the wrench icon in the top-level row (next to the user name), and by selecting *Simulator*.

Hawk's background changes color to indicate the simulator is active. A simulator control dialog is displayed in the bottom right hand corner of the *Cluster Status* screen. Its title *Simulator (initial state)* indicates that no simulator run has occurred yet.

- 3** Fill the simulator's event queue:
  - 3a** To simulate status change of a node: Click *+Node* in the simulator control dialog. Select the *Node* you want to manipulate and select its target *State*. Confirm your changes to add them to the queue of events listed in the controller dialog.
  - 3b** To simulate a resource operation: Click *+Op* in the simulator control dialog. Select the *Resource* to manipulate and the *Operation* to simulate. If necessary, define an *Interval*. Select the *Node* on which to run the operation and the targeted *Result*. Confirm your changes to add them to the queue of events listed in the controller dialog.

- 4 Repeat the previous steps for any other node status changes or resource operations you want to simulate.

**Figure 5.15:** *Hawk—Simulator with Injected Events*



- 5 To inject other changes that you want to simulate:

- 5a Switch to one or more of the following Hawk screens: *Cluster Status*, *Setup Wizard*, *Cluster Configuration*, *Resources*, or *Constraints*.

---

**NOTE: History Explorer and Simulator Mode**

Clicking the *History Explorer* tab will deactivate simulator mode.

---

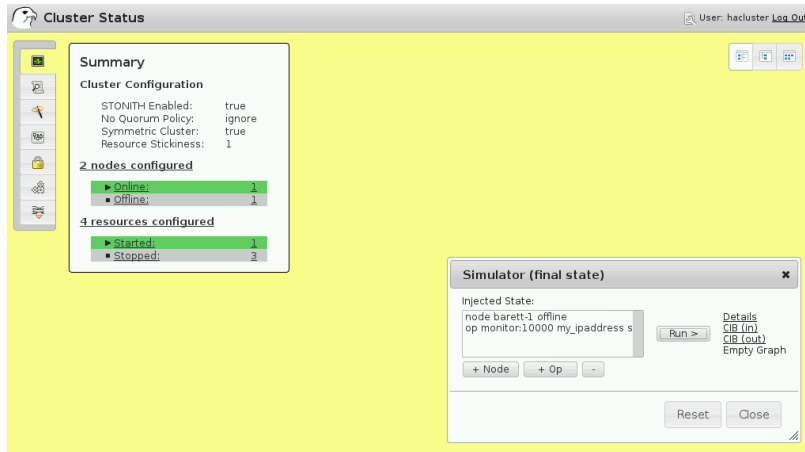
- 5b Add or modify parameters on the screens as desired.

The simulator will collect the changes from all screens and will add them to its internal queue of events.

- 5c To return to the simulator control dialog, switch to the *Cluster Status* screen or click the wrench icon in the top-level row and click *Simulator* again.
- 6 If you want to remove an event listed in *Injected State*, select the respective entry and click the minus icon beneath the list.

- 7 Start the simulation run by clicking *Run* in the simulator control dialog. The *Cluster Status* screen displays the simulated events. For example, if you marked a node as unclean, it will now be shown offline, and all its resources will be stopped. The simulator control dialog changes to *Simulator (final state)*.

**Figure 5.16:** *Hawk—Simulator in Final State*



- 8 To view more detailed information about the simulation run:
  - 8a Click the *Details* link in the simulator dialog to see log snippets of what occurred.
  - 8b Click the *Graph* link to show the transition graph.
  - 8c Click *CIB (in)* to display the initial CIB state. To see what the CIB would look like after the transition, click *CIB (out)*.
- 9 To start from scratch with a new simulation, use the *Reset* button.
- 10 To exit the simulation mode, close the simulator control dialog. The *Cluster Status* screen switches back to its normal color and displays the current cluster state.



## 5.4.8 Generating a Cluster Report

For analysis and diagnosis of problems occurring on the cluster, Hawk can generate a cluster report that collects information from all nodes in the cluster.

**Procedure 5.29:** *Generating an hb\_report*

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2 Click the wrench icon next to the user name in the top-level row, and select *Generate hb\_report*.
- 3 By default, the period to examine is the last hour. To modify this, set another *Start Time* and *End Time*.
- 4 Click *Generate*.
- 5 After the report has been created, download the \*.tar.bz2 file by clicking the respective link.

For more information about the log files that tools like `hb_report` and `crm_report` cover, refer to “How can I create a report with an analysis of all my cluster nodes?” (page 321).

## 5.5 Monitoring Multiple Clusters

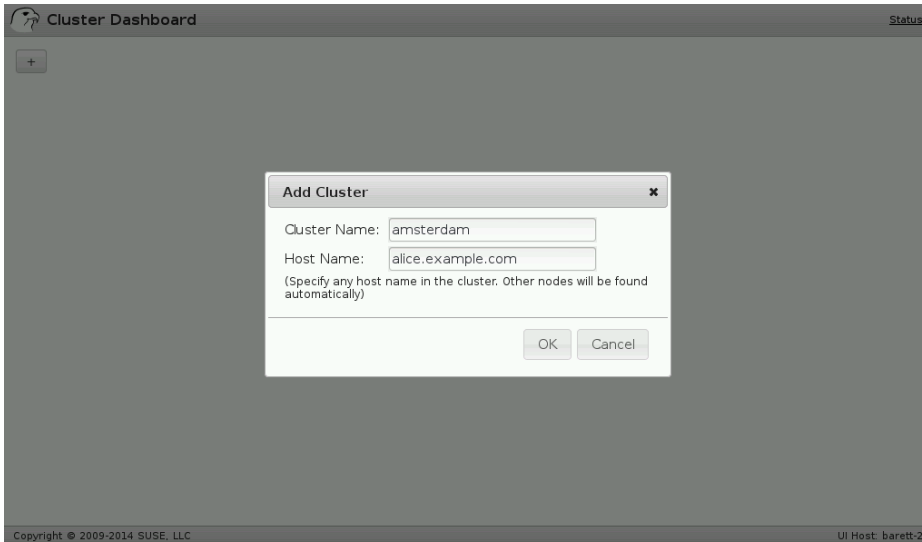
You can use Hawk as a single point of administration for monitoring multiple clusters. Hawk's *Cluster Dashboard* allows you to view a summary of multiple clusters, with each summary listing the number of nodes, resources, tickets (if you use Geo clusters), and their state. The summary also shows whether any failures have appeared in the respective cluster.

The cluster information displayed in the *Cluster Dashboard* is stored in a persistent cookie. This means you need to decide which Hawk instance you want to view the *Cluster Dashboard* on, and always use that one. The machine you are running Hawk on does not even need to be part of any cluster for that purpose—it can be a separate, unrelated system.

## **Procedure 5.30:** *Monitoring Multiple Clusters with Hawk*

### **Prerequisites**

- All clusters to be monitored from Hawk's *Cluster Dashboard* must be running SUSE Linux Enterprise High Availability Extension 11 SP4. It is not possible to monitor clusters that are running earlier versions of SUSE Linux Enterprise High Availability Extension.
  - If you did not replace the self-signed certificate for Hawk on every cluster node with your own certificate (or a certificate signed by an official Certificate Authority), log in to Hawk on *every* node in *every* cluster at least once. Verify the certificate (and add an exception in the browser to bypass the warning).
  - If you are using Mozilla Firefox, you must change its preferences to *Accept third-party cookies*. Otherwise cookies from monitored clusters will not be set, thus preventing login to the clusters you are trying to monitor.
- 1** Start the Hawk Web service on a machine you want to use for monitoring multiple clusters.
  - 2** Start a Web browser and as URL enter the IP address or host name of the machine that runs Hawk:  
  
`https://IPaddress:7630/`
  - 3** On the Hawk login screen, click the *Dashboard* link in the right upper corner.  
  
The *Add Cluster* dialog appears.

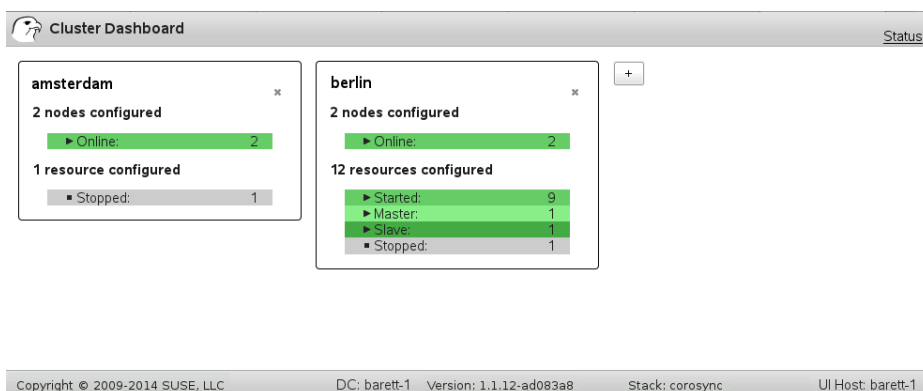


- 4 Enter a custom *Cluster Name* with which to identify the cluster in the *Cluster Dashboard*.
- 5 Enter the *Host Name* of one of the cluster nodes and confirm your changes.

The *Cluster Dashboard* opens and shows a summary of the cluster you have added.

- 6 To add more clusters to the dashboard, click the plus icon and enter the details for the next cluster.

**Figure 5.17:** *Hawk—Cluster Dashboard*



- 7 To remove a cluster from the dashboard, click the × icon next to the cluster's summary.
- 8 To view more details about a cluster, click somewhere in the cluster's box on the dashboard.

This opens a new browser window or new browser tab. If you are not currently logged in to the cluster, this takes you to the Hawk login screen. After having logged in, Hawk shows the *Cluster Status* of that cluster in the summary view. From here, you can administer the cluster with Hawk as usual.

- 9 As the *Cluster Dashboard* stays open in a separate browser window or tab, you can easily switch between the dashboard and the administration of individual clusters in Hawk.

Any status changes for nodes or resources are reflected almost immediately within the *Cluster Dashboard*.

## 5.6 Hawk for Geo Clusters

For more details on Hawk features that relate to geographically dispersed clusters (Geo clusters), see the *Quick Start Geo Clustering for SUSE Linux Enterprise High Availability Extension*.

## 5.7 Troubleshooting

### Hawk Log Files

Find the Hawk log files in `/srv/www/hawk/log`. Check these files in case you cannot access Hawk.

If you have trouble starting or stopping a resource with Hawk, check the Pacemaker log messages. By default, Pacemaker logs to `/var/log/messages`.

### Authentication Fails

If you cannot log in to Hawk with a new user that is a member of the `haclient` group (or if you experience delays until Hawk accepts logins from this user), stop the `nsd` daemon with `rcnsd stop` and try again.

### Replacing the Self-Signed Certificate

To avoid the warning about the self-signed certificate on first Hawk start-up, replace the automatically created certificate with your own certificate or a certificate that was signed by an official Certificate Authority (CA).

The certificate is stored in `/etc/lighttpd/certs/hawk-combined.pem` and contains both key and certificate.

Change the permissions to make the file only accessible by root:

```
root # chown root.root /etc/lighttpd/certs/hawk-combined.pem
      chmod 600 /etc/lighttpd/certs/hawk-combined.pem
```

After you have created or received your new key and certificate, combine them by executing the following command:

```
root # cat keyfile certificatefile > /etc/lighttpd/certs/hawk-combined.pem
```

### Login to Hawk Fails After Using History Explorer/hb\_report

Depending on the period of time you defined in the *History Explorer* or *hb\_report* and the events that took place in the cluster during this time, Hawk might collect an extensive amount of information. It is stored in log files in the `/tmp` directory. This might consume the remaining free disk space on your node. In case Hawk should not respond after using the *History Explorer* or *hb\_report*, check the hard disk of your cluster node and remove the respective log files.

### Cluster Dashboard: Unable to connect to host

If adding clusters to Hawk's dashboard fails, check the prerequisites listed in Procedure 5.30, "Monitoring Multiple Clusters with Hawk" (page 132).

### *Cluster Dashboard: Node Not Accessible*

The *Cluster Dashboard* only polls one node in each cluster for status. If the node being polled goes down, the dashboard will cycle to poll another node. In that case, Hawk briefly displays a warning message about that node being inaccessible. The message will disappear after Hawk has found another node to contact to.

# Configuring and Managing Cluster Resources (GUI)

This chapter introduces the Pacemaker GUI and covers basic tasks needed when configuring and managing cluster resources: modifying global cluster options, creating basic and advanced types of resources (groups and clones), configuring constraints, specifying failover nodes and failback nodes, configuring resource monitoring, starting, cleaning up or removing resources, and migrating resources manually.

Support for the GUI is provided by two packages: The `pacemaker-mgmt` package contains the back-end for the GUI (the `mgmtd` daemon). It must be installed on all cluster nodes you want to connect to with the GUI. On any machine where you want to run the GUI, install the `pacemaker-mgmt-client` package.

---

## NOTE: User Authentication

To log in to the cluster from the Pacemaker GUI, the respective user must be a member of the `haclient` group. The installation creates a linux user named `hacluster` and adds the user to the `haclient` group.

Before using the Pacemaker GUI, either set a password for the `hacluster` user or create a new user which is member of the `haclient` group.

Do this on every node you will connect to with the Pacemaker GUI.

---

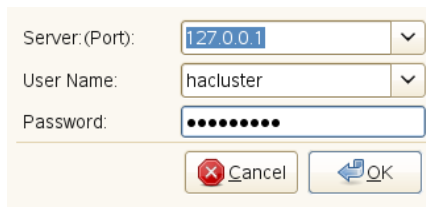
## 6.1 Pacemaker GUI—Overview

To start the Pacemaker GUI, enter `crm_gui` at the command line. To access the configuration and administration options, you need to log in to a cluster.

### 6.1.1 Logging in to a Cluster

To connect to the cluster, select *Connection > Login*. By default, the *Server* field shows the localhost's IP address and `hacluster` as *User Name*. Enter the user's password to continue.

**Figure 6.1:** *Connecting to the Cluster*

A screenshot of a login dialog box with a light beige background. It contains three input fields: 'Server:(Port):' with the value '127.0.0.1', 'User Name:' with the value 'hacluster', and 'Password:' with masked characters '.....'. Below the fields are two buttons: 'Cancel' with a red 'X' icon and 'OK' with a blue arrow icon.

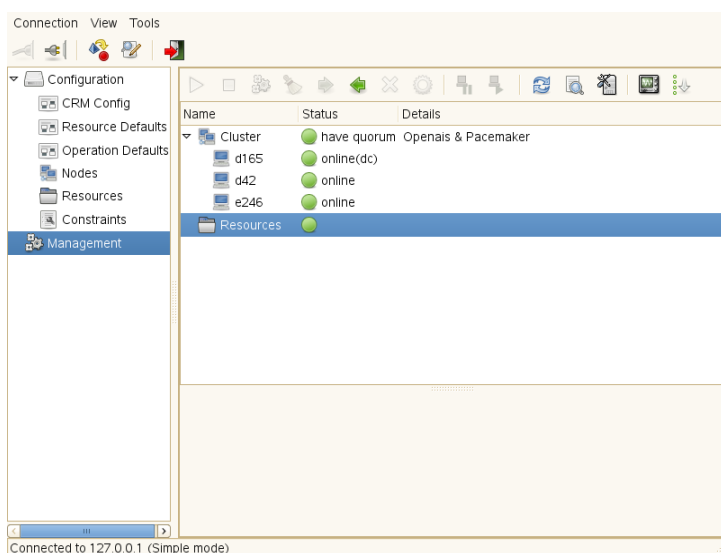
If you are running the Pacemaker GUI remotely, enter the IP address of a cluster node as *Server*. As *User Name*, you can also use any other user belonging to the `haclient` group to connect to the cluster.

### 6.1.2 Main Window

After being connected, the main window opens:



**Figure 6.2:** *Pacemaker GUI - Main Window*



---

## NOTE: Available Functions in Pacemaker GUI

By default, users logged in as `root` or `hacluster` have full read-write access to all cluster configuration tasks. However, *Access Control Lists* (page 215) can be used to define fine-grained access permissions.

If ACLs are enabled in the CRM, the available functions in the Pacemaker GUI depend on the user role and access permission assigned to you.

---

To view or modify cluster components like the CRM, resources, nodes or constraints, select the respective subentry of the *Configuration* category in the left pane and use the options that become available in the right pane. Additionally, the Pacemaker GUI lets you easily view, edit, import and export XML fragments of the CIB for the following subitems: *Resource Defaults*, *Operation Defaults*, *Nodes*, *Resources*, and *Constraints*. Select any of the *Configuration* subitems and select *Show > XML Mode* in the upper right corner of the window.

If you have already configured your resources, click the *Management* category in the left pane to show the status of your cluster and its resources. This view also allows you to set nodes to `standby` and to modify the management status of nodes (if they are currently managed by the cluster or not). To access the main functions

for resources (starting, stopping, cleaning up or migrating resources), select the resource in the right pane and use the icons in the toolbar. Alternatively, right-click the resource and select the respective menu item from the context menu.

The Pacemaker GUI also allows you to switch between different view modes, influencing the behavior of the software and hiding or showing certain aspects:

#### Simple Mode

Lets you add resources in a wizard-like mode. When creating and modifying resources, shows the frequently-used tabs for sub-objects, allowing you to directly add objects of that type via the tab.

Allows you to view and change all available global cluster options by selecting the *CRM Config* entry in the left pane. The right pane then shows the values that are currently set. If no specific value is set for an option, it shows the default values instead.

#### Expert Mode

Lets you add resources in either a wizard-like mode or via dialog windows. When creating and modifying resources, it only shows the corresponding tab if a particular type of sub-object already exists in CIB. When adding a new sub-object, you will be prompted to select the object type, thus allowing you to add all supported types of sub-objects.

When selecting the *CRM Config* entry in the left pane, it only shows the values of global cluster options that have been actually set. It hides all cluster options that will automatically use the defaults (because no values have been set). In this mode, the global cluster options can only be modified by using the individual configuration dialogs.

#### Hack Mode

Has the same functions as the expert mode. Allows you to add additional attribute sets that include specific rules to make your configuration more dynamic. For example, you can make a resource have different instance attributes depending on the node it is hosted on. Furthermore, you can add a time-based rule for a meta attribute set to determine when the attributes take effect.

The window's status bar also shows the currently active mode.

The following sections guide you through the main tasks you need to execute when configuring cluster options and resources and show you how to administer the

resources with the Pacemaker GUI. Where not stated otherwise, the step-by-step instructions reflect the procedure as executed in the simple mode.

## 6.2 Configuring Global Cluster Options

Global cluster options control how the cluster behaves when confronted with certain situations. They are grouped into sets and can be viewed and modified with the cluster management tools like Pacemaker GUI and `crm` shell. The predefined values can be kept in most cases. However, to make key functions of your cluster work correctly, you need to adjust the following parameters after basic cluster setup:

- Option `no-quorum-policy` (page 54)
- Option `stonith-enabled` (page 55)

### **Procedure 6.1:** *Modifying Global Cluster Options*

- 1 Start the Pacemaker GUI and log in to the cluster as described in Section 6.1.1, “Logging in to a Cluster” (page 138).
- 2 Select *View > Simple Mode*.
- 3 In the left pane, select *CRM Config* to view the global cluster options and their current values.
- 4 Depending on your cluster requirements, set *No Quorum Policy* to the appropriate value.
- 5 If you need to disable fencing for any reasons, deselect *stonith-enabled*.

---

#### **IMPORTANT: No Support Without STONITH**

A cluster without STONITH enabled is not supported.

---

- 6 Confirm your changes with *Apply*.

You can at any time switch back to the default values for all options by selecting *CRM Config* in the left pane and clicking *Default*.

## 6.3 Configuring Cluster Resources

As a cluster administrator, you need to create cluster resources for every resource or application you run on servers in your cluster. Cluster resources can include Web sites, e-mail servers, databases, file systems, virtual machines, and any other server-based applications or services you want to make available to users at all times.

For an overview of resource types you can create, refer to Section 4.2.3, “Types of Resources” (page 58).

### 6.3.1 Creating Simple Cluster Resources

To create the most basic type of a resource, proceed as follows:

**Procedure 6.2:** *Adding Primitive Resources*

- 1** Start the Pacemaker GUI and log in to the cluster as described in Section 6.1.1, “Logging in to a Cluster” (page 138).
- 2** In the left pane, select *Resources* and click *Add > Primitive*.
- 3** In the next dialog, set the following parameters for the resource:
  - 3a** Enter a unique *ID* for the resource.
  - 3b** From the *Class* list, select the resource agent class you want to use for that resource: *lsb*, *ocf*, *service*, or *stonith*. For more information, see Section 4.2.2, “Supported Resource Agent Classes” (page 57).
  - 3c** If you selected *ocf* as class, also specify the *Provider* of your OCF resource agent. The OCF specification allows multiple vendors to supply the same resource agent.
  - 3d** From the *Type* list, select the resource agent you want to use (for example, *IPaddr* or *Filesystem*). A short description for this resource agent is displayed below.

The selection you get in the *Type* list depends on the *Class* (and for OCF resources also on the *Provider*) you have chosen.

- 3e** Below *Options*, set the *Initial state of resource*.

- 3f** Activate *Add monitor operation* if you want the cluster to monitor if the resource is still healthy.

**Add Primitive - Basic Settings**

**Required**

ID:

Class:

Provider:

Type:

**Description**

Manages virtual IPv4 addresses.

This script manages IP alias IP addresses.  
It can add an IP alias, or remove one.

**Options**

Initial state of resource:

☒ Add monitor operation

- 4** Click *Forward*. The next window shows a summary of the parameters that you have already defined for that resource. All required *Instance Attributes* for that resource are listed. You need to edit them in order to set them to appropriate values. You may also need to add more attributes, depending on your deployment and settings. For details how to do so, refer to Procedure 6.3, “Adding or Modifying Meta and Instance Attributes” (page 144).
- 5** If all parameters are set according to your wishes, click *Apply* to finish the configuration of that resource. The configuration dialog is closed and the main window shows the newly added resource.

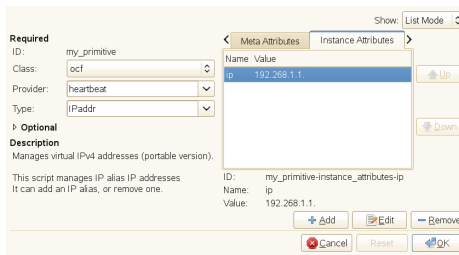
During or after creation of a resource, you can add or modify the following parameters for resources:

- *Instance attributes*—they determine which instance of a service the resource controls. For more information, refer to Section 4.2.7, “Instance Attributes (Parameters)” (page 67).
- *Meta attributes*—they tell the CRM how to treat a specific resource. For more information, refer to Section 4.2.6, “Resource Options (Meta Attributes)” (page 63).

- **Operations**—they are needed for resource monitoring. For more information, refer to Section 4.2.8, “Resource Operations” (page 69).

**Procedure 6.3:** *Adding or Modifying Meta and Instance Attributes*

- 1 In the Pacemaker GUI main window, click *Resources* in the left pane to see the resources already configured for the cluster.
- 2 In the right pane, select the resource to modify and click *Edit* (or double-click the resource). The next window shows the basic resource parameters and the *Meta Attributes*, *Instance Attributes* or *Operations* already defined for that resource.



- 3 To add a new meta attribute or instance attribute, select the respective tab and click *Add*.
- 4 Select the *Name* of the attribute you want to add. A short *Description* is displayed.
- 5 If needed, specify an attribute *Value*. Otherwise the default value of that attribute will be used.
- 6 Click *OK* to confirm your changes. The newly added or modified attribute appears on the tab.
- 7 If all parameters are set according to your wishes, click *OK* to finish the configuration of that resource. The configuration dialog is closed and the main window shows the modified resource.

---

**TIP: XML Source Code for Resources**

The Pacemaker GUI allows you to view the XML fragments that are generated from the parameters that you have defined. For individual resources, select *Show > XML Mode* in the top right corner of the resource configuration dialog.

To access the XML representation of all resources that you have configured, click *Resources* in the left pane and then select *Show > XML Mode* in the upper right corner of the main window.

The editor displaying the XML code allows you to *Import* or *Export* the XML elements or to manually edit the XML code.

---

## 6.3.2 Creating STONITH Resources

---

### **IMPORTANT: No Support Without STONITH**

A cluster without STONITH running is not supported.

---

By default, the global cluster option `stonith-enabled` is set to `true`: If no STONITH resources have been defined, the cluster will refuse to start any resources. To complete STONITH setup, you need to configure one or more STONITH resources. While they are configured similar to other resources, the behavior of STONITH resources is different in some respects. For details refer to Section 9.3, “STONITH Resources and Configuration” (page 206).

#### **Procedure 6.4:** *Adding a STONITH Resource*

- 1** Start the Pacemaker GUI and log in to the cluster as described in Section 6.1.1, “Logging in to a Cluster” (page 138).
- 2** In the left pane, select *Resources* and click *Add > Primitive*.
- 3** In the next dialog, set the following parameters for the resource:
  - 3a** Enter a unique *ID* for the resource.
  - 3b** From the *Class* list, select the resource agent class *stonith*.
  - 3c** From the *Type* list, select the STONITH plug-in for controlling your STONITH device. A short description for this plug-in is displayed below.
  - 3d** Below *Options*, set the *Initial state of resource*.
  - 3e** Activate *Add monitor operation* if you want the cluster to monitor the fencing device. For more information, refer to Section 9.4, “Monitoring Fencing Devices” (page 209).

- 4 Click *Forward*. The next window shows a summary of the parameters that you have already defined for that resource. All required *Instance Attributes* for the selected STONITH plug-in are listed. You need to edit them in order to set them to appropriate values. You may also need to add more attributes or monitor operations, depending on your deployment and settings. For details how to do so, refer to Procedure 6.3, “Adding or Modifying Meta and Instance Attributes” (page 144) and Section 6.3.8, “Configuring Resource Monitoring” (page 156).
- 5 If all parameters are set according to your wishes, click *Apply* to finish the configuration of that resource. The configuration dialog closes and the main window shows the newly added resource.

To complete your fencing configuration, add constraints or use clones or both. For more details, refer to Chapter 9, *Fencing and STONITH* (page 203).

## 6.3.3 Using Resource Templates

If you want to create several resources with similar configurations, defining a resource template is the easiest way. Once defined, it can be referenced in primitives or in certain types of constraints. For detailed information about function and use of resource templates, refer to Section 4.4.3, “Resource Templates and Constraints” (page 77).

- 1 Start the Pacemaker GUI and log in to the cluster as described in Section 6.1.1, “Logging in to a Cluster” (page 138).
- 2 In the left pane, select *Resources* and click *Add > Template*.
- 3 Enter a unique *ID* for the template.
- 4 Specify the resource template as you would specify a primitive. Follow Procedure 6.2: Adding Primitive Resources, starting with Step 3b (page 142).
- 5 If all parameters are set according to your wishes, click *Apply* to finish the configuration of the resource template. The configuration dialog is closed and the main window shows the newly added resource template.

### **Procedure 6.5:** *Referencing Resource Templates*

- 1 Start the Pacemaker GUI and log in to the cluster as described in Section 6.1.1, “Logging in to a Cluster” (page 138).



- 2** To reference the newly created resource template in a primitive, follow these steps:
  - 2a** In the left pane, select *Resources* and click *Add > Primitive*.
  - 2b** Enter a unique *ID* and specify *Class*, *Provider*, and *Type*.
  - 2c** Select the *Template* to reference.
  - 2d** If you want to set specific instance attributes, operations or meta attributes that deviate from the template, continue to configure the resource as described in Procedure 6.2, “Adding Primitive Resources” (page 142).
- 3** To reference the newly created resource template in colocalational or order constraints:
  - 3a** Configure the constraints as described in Procedure 6.7, “Adding or Modifying Colocalational Constraints” (page 149) or Procedure 6.8, “Adding or Modifying Ordering Constraints” (page 150), respectively.
  - 3b** For colocation constraints, the *Resources* drop-down list will show the IDs of all resources and resource templates that have been configured. From there, select the template to reference.
  - 3c** Likewise, for ordering constraints, the *First* and *Then* drop-down lists will show both resources and resource templates. From there, select the template to reference.

## 6.3.4 Configuring Resource Constraints

Having all the resources configured is only part of the job. Even if the cluster knows all needed resources, it might still not be able to handle them correctly. Resource constraints let you specify which cluster nodes resources can run on, what order resources will load, and what other resources a specific resource is dependent on.

For an overview which types of constraints are available, refer to Section 4.4.1, “Types of Constraints” (page 74). When defining constraints, you also need to specify scores. For more information about scores and their implications in the cluster, see Section 4.4.2, “Scores and Infinity” (page 77).

Learn how to create the different types of the constraints in the following procedures.

**Procedure 6.6:** *Adding or Modifying Locational Constraints*

- 1 Start the Pacemaker GUI and log in to the cluster as described in Section 6.1.1, “Logging in to a Cluster” (page 138).
- 2 In the Pacemaker GUI main window, click *Constraints* in the left pane to see the constraints already configured for the cluster.
- 3 In the left pane, select *Constraints* and click *Add*.
- 4 Select *Resource Location* and click *OK*.
- 5 Enter a unique *ID* for the constraint. When modifying existing constraints, the ID is already defined and is displayed in the configuration dialog.
- 6 Select the *Resource* for which to define the constraint. The list shows the IDs of all resources that have been configured for the cluster.
- 7 Set the *Score* for the constraint. Positive values indicate the resource can run on the *Node* you specify below. Negative values mean it should not run on this node. Setting the score to `INFINITY` forces the resource to run on the node. Setting it to `-INFINITY` means the resources must not run on the node.
- 8 Select the *Node* for the constraint.

The screenshot shows a configuration window titled "Required" for adding a constraint. At the top right, there is a "Show:" dropdown menu set to "List Mode". Below this, there are four labeled input fields, each with a dropdown arrow: "ID:" with the value "my\_location\_constraint", "Resource:" with the value "my\_stonith", "Score:" with the value "INFINITY", and "Node:" with the value "d42". At the bottom of the window, there are two rows of buttons. The first row contains three buttons: "+ Add", "Edit" (with a pencil icon), and "- Remove". The second row contains three buttons: "Cancel" (with a red X icon), "Reset", and "OK" (with a blue checkmark icon).

- 9 If you leave the *Node* and the *Score* field empty, you can also add rules by clicking *Add > Rule*. To add a lifetime, just click *Add > Lifetime*.

- 10 If all parameters are set according to your wishes, click *OK* to finish the configuration of the constraint. The configuration dialog is closed and the main window shows the newly added or modified constraint.

**Procedure 6.7:** *Adding or Modifying Colocational Constraints*

- 1 Start the Pacemaker GUI and log in to the cluster as described in Section 6.1.1, “Logging in to a Cluster” (page 138).
- 2 In the Pacemaker GUI main window, click *Constraints* in the left pane to see the constraints already configured for the cluster.
- 3 In the left pane, select *Constraints* and click *Add*.
- 4 Select *Resource Colocation* and click *OK*.
- 5 Enter a unique *ID* for the constraint. When modifying existing constraints, the ID is already defined and is displayed in the configuration dialog.
- 6 Select the *Resource* which is the colocation source. The list shows the IDs of all resources and resource templates that have been configured for the cluster.

If the constraint cannot be satisfied, the cluster may decide not to allow the resource to run at all.

- 7 If you leave both the *Resource* and the *With Resource* field empty, you can also add a resource set by clicking *Add > Resource Set*. To add a lifetime, just click *Add > Lifetime*.
- 8 In *With Resource*, define the colocation target. The cluster will decide where to put this resource first and then decide where to put the resource in the *Resource* field.
- 9 Define a *Score* to determine the location relationship between both resources. Positive values indicate the resources should run on the same node. Negative values indicate the resources should not run on the same node. Setting the score to `INFINITY` forces the resources to run on the same node. Setting it to `-INFINITY` means the resources must not run on the same node. The score will be combined with other factors to decide where to put the resource.
- 10 If needed, specify further parameters, like *Resource Role*.

Depending on the parameters and options you choose, a short *Description* explains the effect of the colocal constraint you are configuring.

- 11 If all parameters are set according to your wishes, click *OK* to finish the configuration of the constraint. The configuration dialog is closed and the main window shows the newly added or modified constraint.

**Procedure 6.8:** *Adding or Modifying Ordering Constraints*

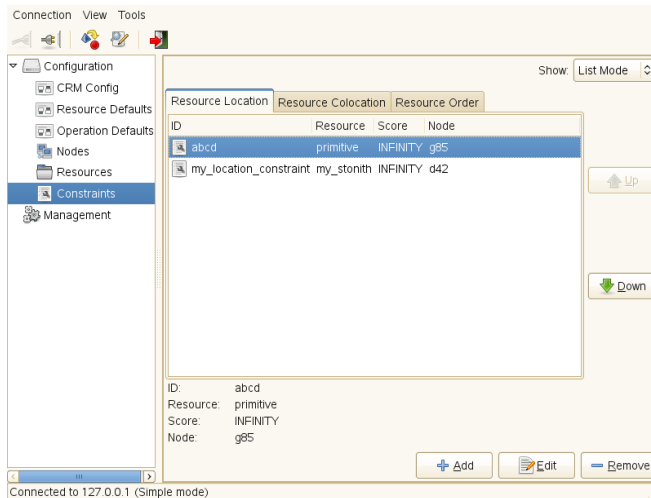
- 1 Start the Pacemaker GUI and log in to the cluster as described in Section 6.1.1, “Logging in to a Cluster” (page 138).
- 2 In the Pacemaker GUI main window, click *Constraints* in the left pane to see the constraints already configured for the cluster.
- 3 In the left pane, select *Constraints* and click *Add*.
- 4 Select *Resource Order* and click *OK*.
- 5 Enter a unique *ID* for the constraint. When modifying existing constraints, the *ID* is already defined and is displayed in the configuration dialog.
- 6 With *First*, define the resource that must be started before the resource specified with *Then* is allowed to.
- 7 With *Then* define the resource that will start after the *First* resource.

Depending on the parameters and options you choose, a short *Description* explains the effect of the ordering constraint you are configuring.

- 8 If needed, define further parameters, for example:
  - 8a Specify a *Score*. If greater than zero, the constraint is mandatory, otherwise it is only a suggestion. The default value is `INFINITY`.
  - 8b Specify a value for *Symmetrical*. If `true`, the resources are stopped in the reverse order. The default value is `true`.
- 9 If all parameters are set according to your wishes, click *OK* to finish the configuration of the constraint. The configuration dialog is closed and the main window shows the newly added or modified constraint.

You can access and modify all constraints that you have configured in the *Constraints* view of the Pacemaker GUI.

**Figure 6.3:** *Pacemaker GUI - Constraints*



## 6.3.5 Specifying Resource Failover Nodes

A resource will be automatically restarted if it fails. If that cannot be achieved on the current node, or it fails `N` times on the current node, it will try to fail over to another node. You can define a number of failures for resources (a `migration-threshold`), after which they will migrate to a new node. If you have more than two nodes in your cluster, the node a particular resource fails over to is chosen by the High Availability software.

However, you can specify the node a resource will fail over to by proceeding as follows:

- 1 Configure a location constraint for that resource as described in Procedure 6.6, “Adding or Modifying Locational Constraints” (page 148).
- 2 Add the `migration-threshold` meta attribute to that resource as described in Procedure 6.3, “Adding or Modifying Meta and Instance Attributes” (page 144) and enter a *Value* for the `migration-threshold`. The value should be positive and less than `INFINITY`.

- 3 If you want to automatically expire the failcount for a resource, add the `failure-timeout` meta attribute to that resource as described in Procedure 6.3, “Adding or Modifying Meta and Instance Attributes” (page 144) and enter a *Value* for the failure-timeout.
- 4 If you want to specify additional failover nodes with preferences for a resource, create additional location constraints.

For an example of the process flow in the cluster regarding migration thresholds and failcounts, see Example 4.6, “Migration Threshold—Process Flow” (page 78).

Instead of letting the failcount for a resource expire automatically, you can also clean up failcounts for a resource manually at any time. Refer to Section 6.4.2, “Cleaning Up Resources” (page 164) for the details.

## 6.3.6 Specifying Resource Failback Nodes (Resource Stickiness)

A resource might fail back to its original node when that node is back online and in the cluster. If you want to prevent a resource from failing back to the node it was running on prior to failover, or if you want to specify a different node for the resource to fail back to, you must change its resource stickiness value. You can either specify resource stickiness when you are creating a resource, or afterwards.

For the implications of different resource stickiness values, refer to Section 4.4.5, “Failback Nodes” (page 79).

### ***Procedure 6.9: Specifying Resource Stickiness***

- 1 Add the `resource-stickiness` meta attribute to the resource as described in Procedure 6.3, “Adding or Modifying Meta and Instance Attributes” (page 144).

**Required**




Name:  ▼

▼ **Optional**

Value:  ▼

**Description**

How much does the resource prefer to stay where it is? Defaults to the value of "default-resource-stickiness"

- 2 As *Value* for the resource-stickiness, specify a value between `-INFINITY` and `INFINITY`.

## 6.3.7 Configuring Placement of Resources Based on Load Impact

Not all resources are equal. Some, such as Xen guests, require that the node hosting them meets their capacity requirements. If resources are placed such that their combined need exceed the provided capacity, the resources diminish in performance (or even fail).

To take this into account, the High Availability Extension allows you to specify the following parameters:

1. The capacity a certain node *provides*.
2. The capacity a certain resource *requires*.
3. An overall strategy for placement of resources.

Utilization attributes are used to configure both the resource's requirements and the capacity a node provides. The High Availability Extension now also provides means to detect and configure both node capacity and resource requirements automatically. For more details and a configuration example, refer to Section 4.4.6, “Placing Resources Based on Their Load Impact” (page 80).

To manually configure the resource's requirements and the capacity a node provides, proceed as described in Procedure 6.10, “Adding Or Modifying Utilization

Attributes” (page 154). You can name the utilization attributes according to your preferences and define as many name/value pairs as your configuration needs.

**Procedure 6.10:** *Adding Or Modifying Utilization Attributes*

In the following example, we assume that you already have a basic configuration of cluster nodes and resources and now additionally want to configure the capacities a certain node provides and the capacity a certain resource requires. The procedure of adding utilization attributes is basically the same and only differs in Step 2 (page 154) and Step 3 (page 154).

- 1** Start the Pacemaker GUI and log in to the cluster as described in Section 6.1.1, “Logging in to a Cluster” (page 138).
- 2** To specify the capacity a node *provides*:
  - 2a** In the left pane, click *Node*.
  - 2b** In the right pane, select the node whose capacity you want to configure and click *Edit*.
- 3** To specify the capacity a resource *requires*:
  - 3a** In the left pane, click *Resources*.
  - 3b** In the right pane, select the resource whose capacity you want to configure and click *Edit*.
- 4** Select the *Utilization* tab and click *Add* to add an utilization attribute.
- 5** Enter a *Name* for the new attribute. You can name the utilization attributes according to your preferences.
- 6** Enter a *Value* for the attribute and click *OK*. The attribute value must be an integer.
- 7** If you need more utilization attributes, repeat Step 5 (page 154) to Step 6 (page 154).

The *Utilization* tab shows a summary of the utilization attributes that you have already defined for that node or resource.



- 8 If all parameters are set according to your wishes, click *OK* to close the configuration dialog.

Figure 6.4, “Example Configuration for Node Capacity” (page 155) shows the configuration of a node which would provide 8 CPU units and 16 GB of memory to resources running on that node:

**Figure 6.4:** *Example Configuration for Node Capacity*

Show: List Mode

**Required**

ID: bourbaki

Uname: bourbaki

Type: normal

► **Optional**

Instance Attributes Utilization

Name	Value
cpu	8
memory	16384

ID: nodes-bourbaki-cpu

Name: cpu

Value: 8

+ Add Edit - Remove

Cancel Reset OK

An example configuration for a resource requiring 4096 memory units and 4 of the CPU units of a node would look as follows:

**Figure 6.5:** Example Configuration for Resource Capacity

The screenshot shows the Pacemaker GUI configuration for a resource named 'xen1'. On the left, under the 'Required' section, the ID is 'xen1', Class is 'ocf', Provider is 'heartbeat', and Type is 'Xen'. Below this is an 'Optional' section with a description: 'Manages Xen unprivileged domains (DomUs). Resource Agent for the Xen Hypervisor. Manages Xen virtual machine instances by mapping cluster'. On the right, the 'Utilization' tab is active, showing a table with columns 'Name' and 'Value'. The table contains two rows: 'cpu' with value '4' and 'memory' with value '4096'. Below the table, a summary shows ID: 'xen1-utilization-cpu', Name: 'cpu', and Value: '4'. At the bottom right are buttons for '+ Add', 'Edit', '- Remove', 'Cancel', 'Reset', and 'OK'. A 'Show: List Mode' dropdown is at the top right.

Name	Value
cpu	4
memory	4096

ID: xen1-utilization-cpu  
Name: cpu  
Value: 4

After (manual or automatic) configuration of the capacities your nodes provide and the capacities your resources require, you need to set the placement strategy in the global cluster options, otherwise the capacity configurations have no effect. Several strategies are available to schedule the load: for example, you can concentrate it on as few nodes as possible, or balance it evenly over all available nodes. For more information, refer to Section 4.4.6, “Placing Resources Based on Their Load Impact” (page 80).

**Procedure 6.11:** Setting the Placement Strategy

- 1 Start the Pacemaker GUI and log in to the cluster as described in Section 6.1.1, “Logging in to a Cluster” (page 138).
- 2 Select *View > Simple Mode*.
- 3 In the left pane, select *CRM Config* to view the global cluster options and their current values.
- 4 Depending on your requirements, set *Placement Strategy* to the appropriate value.
- 5 If you need to disable fencing for any reasons, deselect *Stonith Enabled*.
- 6 Confirm your changes with *Apply*.

## 6.3.8 Configuring Resource Monitoring

Although the High Availability Extension can detect a node failure, it also has the ability to detect when an individual resource on a node has failed. If you want to

ensure that a resource is running, you must configure resource monitoring for it. Resource monitoring consists of specifying a timeout and/or start delay value, and an interval. The interval tells the CRM how often it should check the resource status. You can also set particular parameters, such as `Timeout` for start or stop operations.

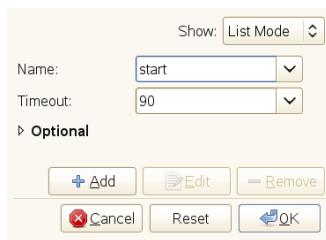
**Procedure 6.12:** *Adding or Modifying Monitor Operations*

- 1 Start the Pacemaker GUI and log in to the cluster as described in Section 6.1.1, “Logging in to a Cluster” (page 138).
- 2 In the Pacemaker GUI main window, click *Resources* in the left pane to see the resources already configured for the cluster.
- 3 In the right pane, select the resource to modify and click *Edit*. The next window shows the basic resource parameters and the meta attributes, instance attributes and operations already defined for that resource.
- 4 To add a new monitor operation, select the respective tab and click *Add*.

To modify an existing operation, select the respective entry and click *Edit*.

- 5 In *Name*, select the action to perform, for example `monitor`, `start`, or `stop`.

The parameters shown below depend on the selection you make here.



The screenshot shows a dialog box for adding a monitor operation. At the top right is a 'Show:' dropdown menu set to 'List Mode'. Below it are two dropdown menus: 'Name:' with 'start' selected and 'Timeout:' with '90' selected. Underneath these is an expandable section labeled 'Optional'. At the bottom of the dialog are four buttons: '+ Add', 'Edit', 'Remove', and 'Cancel'. Below the 'Cancel' button are 'Reset' and 'OK' buttons.

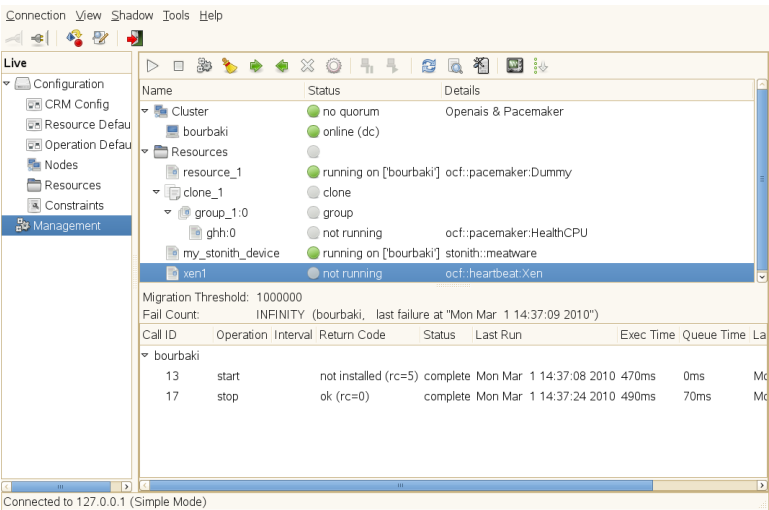
- 6 In the *Timeout* field, enter a value in seconds. After the specified timeout period, the operation will be treated as `failed`. The PE will decide what to do or execute what you specified in the *On Fail* field of the monitor operation.
- 7 If needed, expand the *Optional* section and add parameters, like *On Fail* (what to do if this action ever fails?) or *Requires* (what conditions need to be satisfied before this action occurs?).

- 8 If all parameters are set according to your wishes, click *OK* to finish the configuration of that resource. The configuration dialog is closed and the main window shows the modified resource.

For the processes which take place if the resource monitor detects a failure, refer to Section 4.3, “Resource Monitoring” (page 72).

To view resource failures in the Pacemaker GUI, click *Management* in the left pane, then select the resource whose details you want to see in the right pane. For a resource that has failed, the *Fail Count* and last failure of the resource is shown in the middle of the right pane (below the *Migration threshold* entry).

**Figure 6.6:** Viewing a Resource's Failcount



## 6.3.9 Configuring a Cluster Resource Group

Some cluster resources are dependent on other components or resources, and require that each component or resource starts in a specific order and runs together on the same server. To simplify this configuration we support the concept of groups.

For an example of a resource group and more information about groups and their properties, refer to Section 4.2.5.1, “Groups” (page 60).

---

**NOTE: Empty Groups**

Groups must contain at least one resource, otherwise the configuration is not valid.

---

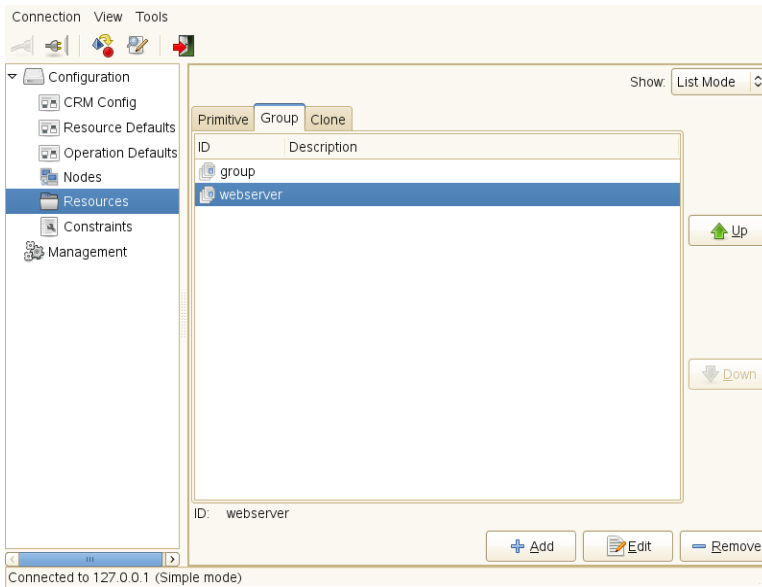
**Procedure 6.13:** *Adding a Resource Group*

- 1 Start the Pacemaker GUI and log in to the cluster as described in Section 6.1.1, “Logging in to a Cluster” (page 138).
- 2 In the left pane, select *Resources* and click *Add > Group*.
- 3 Enter a unique *ID* for the group.
- 4 Below *Options*, set the *Initial state of resource* and click *Forward*.
- 5 In the next step, you can add primitives as sub-resources for the group. These are created similar as described in Procedure 6.2, “Adding Primitive Resources” (page 142).
- 6 If all parameters are set according to your wishes, click *Apply* to finish the configuration of the primitive.
- 7 In the next window, you can continue adding sub-resources for the group by choosing *Primitive* again and clicking *OK*.

When you do not want to add more primitives to the group, click *Cancel* instead. The next window shows a summary of the parameters that you have already defined for that group. The *Meta Attributes* and *Primitives* of the group are listed. The position of the resources in the *Primitive* tab represents the order in which the resources are started in the cluster.

- 8 As the order of resources in a group is important, use the *Up* and *Down* buttons to sort the *Primitives* in the group.
- 9 If all parameters are set according to your wishes, click *OK* to finish the configuration of that group. The configuration dialog is closed and the main window shows the newly created or modified group.

**Figure 6.7:** *Pacemaker GUI - Groups*



Let us assume you already have created a resource group as explained in Procedure 6.13, “Adding a Resource Group” (page 159). The following procedure shows you how to modify the group to match Example 4.1, “Resource Group for a Web Server” (page 60).

**Procedure 6.14:** *Adding Resources to an Existing Group*

- 1 Start the Pacemaker GUI and log in to the cluster as described in Section 6.1.1, “Logging in to a Cluster” (page 138).
- 2 In the left pane, switch to the *Resources* view and in the right pane, select the group to modify and click *Edit*. The next window shows the basic group parameters and the meta attributes and primitives already defined for that resource.
- 3 Click the *Primitives* tab and click *Add*.
- 4 In the next dialog, set the following parameters to add an IP address as sub-resource of the group:

**4a** Enter a unique *ID* (for example, `my_ipaddress`).

- 4b** From the *Class* list, select *ocf* as resource agent class.
- 4c** As *Provider* of your OCF resource agent, select *heartbeat*.
- 4d** From the *Type* list, select *IPaddr* as resource agent.
- 4e** Click *Forward*.
- 4f** In the *Instance Attribute* tab, select the *IP* entry and click *Edit* (or double-click the *IP* entry).
- 4g** As *Value*, enter the desired IP address, for example, 192.168.1.180.
- 4h** Click *OK* and *Apply*. The group configuration dialog shows the newly added primitive.
- 5** Add the next sub-resources (file system and Web server) by clicking *Add* again.
- 6** Set the respective parameters for each of the sub-resources similar to steps Step 4a (page 160) to Step 4h (page 161), until you have configured all sub-resources for the group.

Show: List Mode

**Required**  
ID: group

► **Optional**

Meta Attributes Primitive

ID	Class	Provider	Type
my_ipaddress	ocf	heartbeat	IPaddr
my_filesystem	ocf	heartbeat	Filesystem
my_webserver	ocf	heartbeat	apache

Up Down

ID: my\_ipaddress  
Class: ocf  
Provider: heartbeat  
Type: IPaddr

+ Add Edit - Remove

Cancel Reset OK

As we configured the sub-resources already in the order in that they need to be started in the cluster, the order on the *Primitives* tab is already correct.

- 7 In case you need to change the resource order for a group, use the *Up* and *Down* buttons to sort the resources on the *Primitive* tab.
- 8 To remove a resource from the group, select the resource on the *Primitives* tab and click *Remove*.
- 9 Click *OK* to finish the configuration of that group. The configuration dialog is closed and the main window shows the modified group.

## 6.3.10 Configuring a Clone Resource

You may want certain resources to run simultaneously on multiple nodes in your cluster. To do this you must configure a resource as a clone. Examples of resources that might be configured as clones include STONITH and cluster file systems like OCFS2. You can clone any resource provided. This is supported by the resource's Resource Agent. Clone resources may even be configured differently depending on which nodes they are hosted.

For an overview which types of resource clones are available, refer to Section 4.2.5.2, “Clones” (page 62).

### **Procedure 6.15:** *Adding or Modifying Clones*

- 1 Start the Pacemaker GUI and log in to the cluster as described in Section 6.1.1, “Logging in to a Cluster” (page 138).
- 2 In the left pane, select *Resources* and click *Add > Clone*.
- 3 Enter a unique *ID* for the clone.
- 4 Below *Options*, set the *Initial state of resource*.
- 5 Activate the respective options you want to set for your clone and click *Forward*.
- 6 In the next step, you can either add a *Primitive* or a *Group* as sub-resources for the clone. These are created similar as described in Procedure 6.2, “Adding



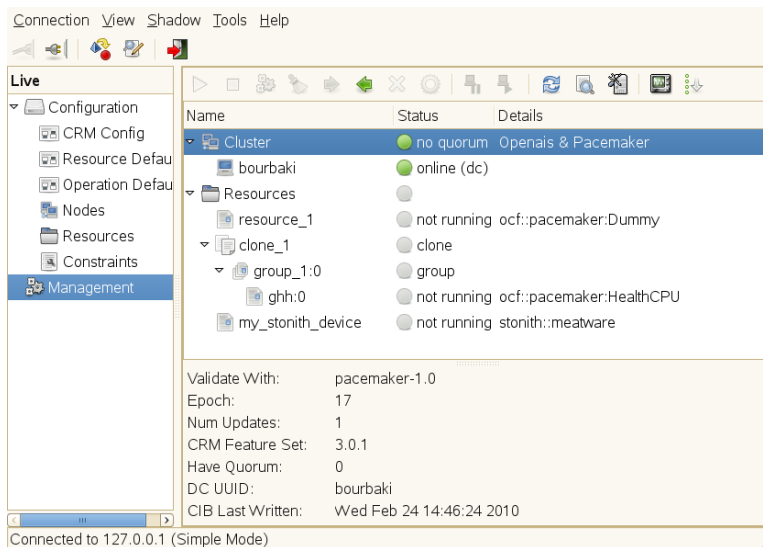
Primitive Resources” (page 142) or Procedure 6.13, “Adding a Resource Group” (page 159).

- 7 If all parameters in the clone configuration dialog are set according to your wishes, click *Apply* to finish the configuration of the clone.

## 6.4 Managing Cluster Resources

Apart from the possibility to configure your cluster resources, the Pacemaker GUI also allows you to manage existing resources. To switch to a management view and to access the available options, click *Management* in the left pane.

**Figure 6.8:** *Pacemaker GUI - Management*



### 6.4.1 Starting Resources

Before you start a cluster resource, make sure it is set up correctly. For example, if you want to use an Apache server as a cluster resource, set up the Apache server first and complete the Apache configuration before starting the respective resource in your cluster.

---

**NOTE: Do Not Touch Services Managed by the Cluster**

When managing a resource with the High Availability Extension, the same resource must not be started or stopped otherwise (outside of the cluster, for example manually or on boot or reboot). The High Availability Extension software is responsible for all service start or stop actions.

However, if you want to check if the service is configured properly, start it manually, but make sure that it is stopped again before High Availability takes over.

For interventions in resources that are currently managed by the cluster, set the resource to `unmanaged mode` first as described in Section 6.4.5, “Changing Management Mode of Resources” (page 167).

---

During creation of a resource with the Pacemaker GUI, you can set the resource's initial state with the `target-role` meta attribute. If its value has been set to `stopped`, the resource does not start automatically after being created.

**Procedure 6.16:** *Starting A New Resource*

- 1 Start the Pacemaker GUI and log in to the cluster as described in Section 6.1.1, “Logging in to a Cluster” (page 138).
- 2 Click *Management* in the left pane.
- 3 In the right pane, right-click the resource and select *Start* from the context menu (or use the *Start Resource* icon in the toolbar).

## 6.4.2 Cleaning Up Resources

A resource will be automatically restarted if it fails, but each failure raises the resource's failcount. View a resource's failcount with the Pacemaker GUI by clicking *Management* in the left pane, then selecting the resource in the right pane. If a resource has failed, its *Fail Count* is shown in the middle of the right pane (below the *Migration Threshold* entry).

If a `migration-threshold` has been set for that resource, the node will no longer be allowed to run the resource as soon as the number of failures has reached the migration threshold.

A resource's failcount can either be reset automatically (by setting a `failure-timeout` option for the resource) or you can reset it manually as described below.

**Procedure 6.17:** *Cleaning Up A Resource*

- 1 Start the Pacemaker GUI and log in to the cluster as described in Section 6.1.1, “Logging in to a Cluster” (page 138).
- 2 Click *Management* in the left pane.
- 3 In the right pane, right-click the respective resource and select *Cleanup Resource* from the context menu (or use the *Cleanup Resource* icon in the toolbar).

This executes the commands `crm_resource -C` and `crm_failcount -D` for the specified resource on the specified node.

For more information, see the man pages of `crm_resource` and `crm_failcount`.

## 6.4.3 Removing Cluster Resources

If you need to remove a resource from the cluster, follow the procedure below to avoid configuration errors:

---

**NOTE: Removing Referenced Resources**

Cluster resources cannot be removed if their ID is referenced by any constraint. If you cannot delete a resource, check where the resource ID is referenced and remove the resource from the constraint first.

---

**Procedure 6.18:** *Removing a Cluster Resource*

- 1 Start the Pacemaker GUI and log in to the cluster as described in Section 6.1.1, “Logging in to a Cluster” (page 138).
- 2 Click *Management* in the left pane.
- 3 Select the respective resource in the right pane.
- 4 Clean up the resource on all nodes as described in Procedure 6.17, “Cleaning Up A Resource” (page 165).

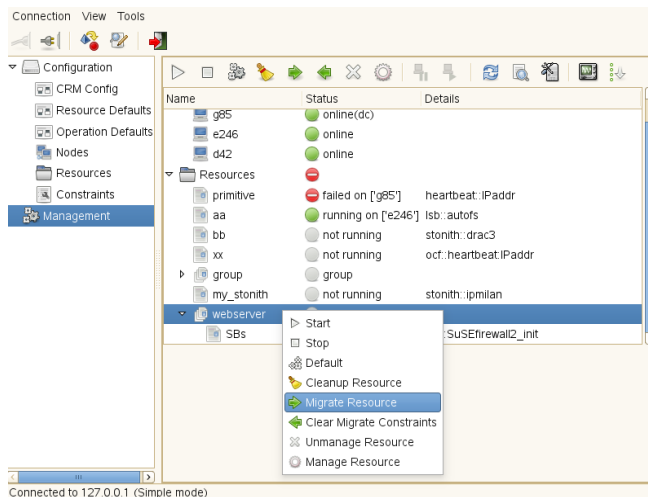
- 5 Stop the resource.
- 6 Remove all constraints that relate to the resource, otherwise removing the resource will not be possible.

## 6.4.4 Migrating Cluster Resources

As mentioned in Section 6.3.5, “Specifying Resource Failover Nodes” (page 151), the cluster will fail over (migrate) resources automatically in case of software or hardware failures—according to certain parameters you can define (for example, migration threshold or resource stickiness). Apart from that, you can also manually migrate a resource to another node in the cluster.

### **Procedure 6.19:** *Manually Migrating a Resource*

- 1 Start the Pacemaker GUI and log in to the cluster as described in Section 6.1.1, “Logging in to a Cluster” (page 138).
- 2 Click *Management* in the left pane.
- 3 Right-click the respective resource in the right pane and select *Migrate Resource*.



- 4 In the new window, select the node to which to move the resource to in *To Node*. This creates a location constraint with an `INFINITY` score for the destination node.

- 5 If you want to migrate the resource only temporarily, activate *Duration* and enter the time frame for which the resource should migrate to the new node. After the expiration of the duration, the resource *can* move back to its original location or it may stay where it is (depending on resource stickiness).
- 6 In cases where the resource cannot be migrated (if the resource's stickiness and constraint scores total more than `INFINITY` on the current node), activate the *Force* option. This forces the resource to move by creating a rule for the current location and a score of `-INFINITY`.

---

#### NOTE

This prevents the resource from running on this node until the constraint is removed with *Clear Migrate Constraints* or the duration expires.

---

- 7 Click *OK* to confirm the migration.

To allow a resource to move back again, proceed as follows:

#### **Procedure 6.20:** *Clearing a Migration Constraint*

- 1 Start the Pacemaker GUI and log in to the cluster as described in Section 6.1.1, “Logging in to a Cluster” (page 138).
- 2 Click *Management* in the left pane.
- 3 Right-click the respective resource in the right pane and select *Clear Migrate Constraints*.

This uses the `crm_resource -U` command. The resource *can* move back to its original location or it may stay where it is (depending on resource stickiness).

For more information, see the `crm_resource` man page or *Pacemaker Explained*, available from <http://www.clusterlabs.org/doc/>. Refer to section *Resource Migration*.

## 6.4.5 Changing Management Mode of Resources

When a resource is being managed by the cluster, it must not be touched otherwise (outside of the cluster). For maintenance of individual resources, you can set the

respective resources to an `unmanaged` mode that allows you to modify the resource outside of the cluster.

**Procedure 6.21:** *Changing Management Mode of Resources*

- 1** Start the Pacemaker GUI and log in to the cluster as described in Section 6.1.1, “Logging in to a Cluster” (page 138).
- 2** Click *Management* in the left pane.
- 3** Right-click the respective resource in the right pane and from the context menu, select *Unmanage Resource*.
- 4** After you have finished the maintenance task for that resource, right-click the respective resource again in the right pane and select *Manage Resource*.

From this point on, the resource will be managed by the High Availability Extension software again.

# Configuring and Managing Cluster Resources (Command Line)

To configure and manage cluster resources, either use the graphical user interface (the Pacemaker GUI) or the `crm` command line utility. For the GUI approach, refer to Chapter 6, *Configuring and Managing Cluster Resources (GUI)* (page 137).

This chapter introduces `crm`, the command line tool and covers an overview of this tool, how to use templates, and mainly configuring and managing cluster resources: creating basic and advanced types of resources (groups and clones), configuring constraints, specifying failover nodes and failback nodes, configuring resource monitoring, starting, cleaning up or removing resources, and migrating resources manually.

---

## NOTE: User Privileges

Sufficient privileges are necessary to manage a cluster. The `crm` command and its subcommands need to be run either as `root` user or as the CRM owner user (typically the user `hacluster`).

However, the `user` option allows you to run `crm` and its subcommands as a regular (unprivileged) user and to change its ID using `sudo` whenever necessary. For example, with the following command `crm` will use `hacluster` as the privileged user ID:

```
root # crm options user hacluster
```

Note that you need to set up `/etc/sudoers` so that `sudo` does not ask for a password.

---

# 7.1 crmsh—Overview

The `crm` command has several subcommands which manage resources, CIBs, nodes, resource agents, and others. It offers a thorough help system with embedded examples. All examples follow a naming convention described in Appendix A (page 327).

Help can be accessed in several ways:

---

**TIP: Distinguish Between Shell Prompt and Interactive `crm` Prompt**

To make all the code and examples more readable, this chapter uses the following notations between shell prompts and the interactive `crm` prompt:

- Shell prompt for user `root`:

```
root #
```

- Interactive `crmsh` prompt (displayed in green, if terminal supports colors):

```
crm(live)#
```

---

## 7.1.1 Getting Help

Help can be accessed in several ways:

- To output the usage of `crm` and its command line options:

```
root # crm --help
```

- To give a list of all available commands:

```
root # crm help
```

- To access other help sections, not just the command reference:

```
root # crm help topics
```

- To view the extensive help text of the `configure` subcommand:

```
root # crm configure help
```



- To print the syntax, its usage, and examples of a subcommand of `configure`:

```
root # crm configure help group
```

This is also possible:

```
root # crm help configure group
```

Almost all output of the `help` subcommand (do not mix it up with the `--help` option) opens a text viewer. This text viewer allows you to scroll up or down and read the help text more comfortably. To leave the text viewer, press the `Q` key.

---

### TIP: Use Tab Completion in Bash and Interactive Shell

The `crmsh` supports full tab completion in Bash directly, not only for the interactive shell. For example, typing `crm help config→|` will complete the word just like in the interactive shell.

---

## 7.1.2 Executing `crmsh`'s Subcommands

The `crm` command itself can be used in the following ways:

- **Directly** Concatenate all subcommands to `crm`, press `Enter` and you see the output immediately. For example, enter `crm help ra` to get information about the `ra` subcommand (resource agents).
- **As `crm` Shell Script** Use `crm` and its subcommands in a script. This can be done in two ways:

```
root # crm -f script.cli
root # crm < script.cli
```

The script can contain any command from `crm`. For example:

```
# A small script file for crm
status
node list
```

Any line starting with the hash symbol (`#`) is a comment and is ignored. If a line is too long, insert a backslash (`\`) at the end and continue in the next line. It is recommended to indent lines that belong to a certain subcommand to improve readability.

- **Interactive as Internal Shell** Type `crm` to enter the internal shell. The prompt changes to `crm(live)`. With `help` you can get an overview of the available

subcommands. As the internal shell has different levels of subcommands, you can “enter” one by typing this subcommand and press **Enter**.

For example, if you type `resource` you enter the resource management level. Your prompt changes to `crm(live) resource#`. If you want to leave the internal shell, use the commands `quit`, `bye`, or `exit`. If you need to go one level back, use `back`, `up`, `end`, or `cd`.

You can enter the level directly by typing `crm` and the respective subcommand(s) without any options and hit **Enter**.

The internal shell supports also tab completion for subcommands and resources. Type the beginning of a command, press `→|` and `crm` completes the respective object.

In addition to previously explained methods, `crmsh` also supports synchronous command execution. Use the `-w` option to activate it. If you have started `crm` without `-w`, you can enable it later with the user preference's `wait` set to `yes` (`options wait yes`). If this option is enabled, `crm` waits until the transition is finished. Whenever a transaction is started, dots are printed to indicate progress. Synchronous command execution is only applicable for commands like `resource start`.

---

**NOTE: Differentiate Between Management and Configuration Subcommands**

The `crm` tool has management capability (the subcommands `resource` and `node`) and can be used for configuration (`cib`, `configure`).

---

The following subsections give you an overview of some important aspects of the `crm` tool.

## 7.1.3 Displaying Information about OCF Resource Agents

As you need to deal with resource agents in your cluster configuration all the time, the `crm` tool contains the `ra` command. Use it to show information about resource agents and to manage them (for additional information, see also Section 4.2.2, “Supported Resource Agent Classes” (page 57)):

```
root # crm ra
crm(live)ra#
```

The command `classes` lists all classes and providers:

```
crm(live)ra# classes
lsb
ocf / heartbeat linbit lvm2 ocfs2 pacemaker
service
stonith
```

To get an overview of all available resource agents for a class (and provider) use the `list` command:

```
crm(live)ra# list ocf
AoEtarget      AudibleAlarm    CTDB             ClusterMon
Delay          Dummy           EvmsSCC          Evmsd
Filesystem     HealthCPU       HealthSMART      ICP
IPaddr         IPaddr2         IPsrcaddr        IPv6addr
LVM            LinuxSCSI       MailTo           ManageRAID
ManageVE       Pure-FTPD       Raid1            Route
SAPDatabase    SAPInstance     SendArp          ServeRAID
...
```

An overview of a resource agent can be viewed with `info`:

```
crm(live)ra# info ocf:linbit:drbd
This resource agent manages a DRBD* resource
as a master/slave resource. DRBD is a shared-nothing replicated storage
device. (ocf:linbit:drbd)
```

Master/Slave OCF Resource Agent for DRBD

Parameters (\* denotes required, [] the default):

```
drbd_resource* (string): drbd resource name
    The name of the drbd resource from the drbd.conf file.
```

```
drbdconf (string, [/etc/drbd.conf]): Path to drbd.conf
    Full path to the drbd.conf file.
```

Operations' defaults (advisory minimum):

```
start          timeout=240
promote        timeout=90
demote         timeout=90
notify         timeout=90
stop           timeout=100
monitor_Slave_0 interval=20 timeout=20 start-delay=1m
monitor_Master_0 interval=10 timeout=20 start-delay=1m
```

Leave the viewer by pressing **Q**. Find a configuration example at Appendix B, *Example of Setting Up a Simple Testing Resource* (page 329).

---

**TIP: Use `crm` Directly**

In the former example we used the internal shell of the `crm` command. However, you do not necessarily need to use it. You get the same results if you add the respective subcommands to `crm`. For example, you can list all the OCF resource agents by entering `crm ra list ocf` in your shell.

---

## 7.1.4 Using Configuration Templates

Configuration templates are ready-made cluster configurations for the `crm` shell. Do not confuse them with the *resource templates* (as described in Section 7.3.2, “Creating Resource Templates” (page 179)). Those are templates for the *cluster* and not for the `crm` shell.

Configuration templates require minimum effort to be tailored to the particular user's needs. Whenever a template creates a configuration, warning messages give hints which can be edited later for further customization.

The following procedure shows how to create a simple yet functional Apache configuration:

- 1 Log in as `root` and start the `crm` interactive shell:

```
root # crm configure
```

- 2 Create a new configuration from a configuration template:

- 2a Switch to the `template` subcommand:

```
crm(live)configure#  
template
```

- 2b List the available configuration templates:

```
crm(live)configure template# list templates  
gfs2-base    filesystem  virtual-ip  apache      clvm        ocfs2       gfs2
```

- 2c Decide which configuration template you need. As we need an Apache configuration, we choose the `apache` template and name it `g-intranet`:

```
crm(live)configure template# new g-intranet apache  
INFO: pulling in template apache
```

```
INFO: pulling in template virtual-ip
```

### 3 Define your parameters:

#### 3a List the configuration you have created:

```
crm(live)configure template# list
g-intranet
```

#### 3b Display the minimum required changes that need to be filled out by you:

```
crm(live)configure template# show
ERROR: 23: required parameter ip not set
ERROR: 61: required parameter id not set
ERROR: 65: required parameter configfile not set
```

#### 3c Invoke your preferred text editor and fill out all lines that have been displayed as errors in Step 3b (page 175):

```
crm(live)configure template#
edit
```

### 4 Show the configuration and check whether it is valid (bold text depends on the configuration you have entered in Step 3c (page 175)):

```
crm(live)configure template# show
primitive virtual-ip ocf:heartbeat:IPaddr \
  params ip="192.168.1.101"
primitive apache ocf:heartbeat:apache \
  params configfile="/etc/apache2/httpd.conf"
monitor apache 120s:60s
group g-intranet \
  apache virtual-ip
```

### 5 Apply the configuration:

```
crm(live)configure template# apply
crm(live)configure# cd ..
crm(live)configure# show
```

### 6 Submit your changes to the CIB:

```
crm(live)configure#
commit
```

It is possible to simplify the commands even more, if you know the details. The above procedure can be summarized with the following command on the shell:

```
root # crm configure template \  
    new g-intranet apache params \  
    configfile="/etc/apache2/httpd.conf" ip="192.168.1.101"
```

If you are inside your internal `crm` shell, use the following command:

```
crm(live)configure template# new intranet apache params \  
    configfile="/etc/apache2/httpd.conf" ip="192.168.1.101"
```

However, the previous command only creates its configuration from the configuration template. It does not apply nor commit it to the CIB.

## 7.1.5 Testing with Shadow Configuration

A shadow configuration is used to test different configuration scenarios. If you have created several shadow configurations, you can test them one by one to see the effects of your changes.

The usual process looks like this:

- 1 Log in as `root` and start the `crm` interactive shell:

```
root # crm configure
```

- 2 Create a new shadow configuration:

```
crm(live)configure# cib new myNewConfig  
INFO: myNewConfig shadow CIB created
```

If you omit the name of the shadow CIB, a temporary name `@tmp@` is created.

- 3 If you want to copy the current live configuration into your shadow configuration, use the following command, otherwise skip this step:

```
crm(myNewConfig)# cib reset myNewConfig
```

The previous command makes it easier to modify any existing resources later.

- 4 Make your changes as usual. After you have created the shadow configuration, all changes go there. To save all your changes, use the following command:

```
crm(myNewConfig)# commit
```

- 5 If you need the live cluster configuration again, switch back with the following command:

```
crm(myNewConfig)configure# cib use live  
crm(live)#
```

## 7.1.6 Debugging Your Configuration Changes

Before loading your configuration changes back into the cluster, it is recommended to review your changes with `pctest`. The `pctest` command can show a diagram of actions that will be induced by committing the changes. You need the `graphviz` package to display the diagrams. The following example is a transcript, adding a monitor operation:

```
root # crm configure
crm(live)configure# show fence-bob
primitive fence-bob stonith:apcsmart \
    params hostlist="bob"
crm(live)configure# monitor fence-bob 120m:60s
crm(live)configure# show changed
primitive fence-bob stonith:apcsmart \
    params hostlist="bob" \
    op monitor interval="120m" timeout="60s"
crm(live)configure# pctest
crm(live)configure# commit
```

## 7.1.7 Cluster Diagram

To output a cluster diagram as shown in Figure 5.2, “Hawk—Cluster Diagram” (page 96), use the command `crm configure graph`. It displays the current configuration on its current window, therefore requiring X11.

If you prefer Scalable Vector Graphics (SVG), use the following command:

```
root # crm configure graph dot config.svg svg
```

## 7.2 Configuring Global Cluster Options

Global cluster options control how the cluster behaves when confronted with certain situations. The predefined values can usually be kept. However, to make key functions of your cluster work correctly, you need to adjust the following parameters after basic cluster setup:

**Procedure 7.1:** *Modifying Global Cluster Options With `crm`*

**1** Log in as `root` and start the `crm` tool:

```
root # crm configure
```

**2** Use the following commands to set the options for two-node clusters only:

```
crm(live)configure# property no-quorum-policy=ignore
crm(live)configure# property stonith-enabled=true
```

---

**IMPORTANT: No Support Without STONITH**

A cluster without STONITH is not supported.

---

**3** Show your changes:

```
crm(live)configure# show
property $id="cib-bootstrap-options" \
  dc-version="1.1.1-530add2a3721a0ecccb24660a97dbfdaa3e68f51" \
  cluster-infrastructure="openais" \
  expected-quorum-votes="2" \
  no-quorum-policy="ignore" \
  stonith-enabled="true"
```

**4** Commit your changes and exit:

```
crm(live)configure# commit
crm(live)configure# exit
```

## 7.3 Configuring Cluster Resources

As a cluster administrator, you need to create cluster resources for every resource or application you run on servers in your cluster. Cluster resources can include Web sites, e-mail servers, databases, file systems, virtual machines, and any other server-based applications or services you want to make available to users at all times.

For an overview of resource types you can create, refer to Section 4.2.3, “Types of Resources” (page 58).

### 7.3.1 Creating Cluster Resources

There are three types of RAs (Resource Agents) available with the cluster (for background information, see Section 4.2.2, “Supported Resource Agent Classes” (page 57)). To add a new resource to the cluster, proceed as follows:

**1** Log in as `root` and start the `crm` tool:



```
root # crm configure
```

## 2 Configure a primitive IP address:

```
crm(live)configure# primitive myIP ocf:heartbeat:IPaddr \  
    params ip=127.0.0.99 op monitor interval=60s
```

The previous command configures a “primitive” with the name `myIP`. You need to choose a class (here `ocf`), provider (`heartbeat`), and type (`IPaddr`). Furthermore, this primitive expects other parameters like the IP address. Change the address to your setup.

## 3 Display and review the changes you have made:

```
crm(live)configure#  
show
```

## 4 Commit your changes to take effect:

```
crm(live)configure#  
commit
```

# 7.3.2 Creating Resource Templates

If you want to create several resources with similar configurations, a resource template simplifies the task. See also Section 4.4.3, “Resource Templates and Constraints” (page 77) for some basic background information. Do not confuse them with the “normal” templates from Section 7.1.4, “Using Configuration Templates” (page 174). Use the `rsc_template` command to get familiar with the syntax:

```
root # crm configure rsc_template  
usage: rsc_template <name> [<class>:[<provider>:]]<type>  
    [params <param>=<value> [<param>=<value>...]]  
    [meta <attribute>=<value> [<attribute>=<value>...]]  
    [utilization <attribute>=<value> [<attribute>=<value>...]]  
    [operations id_spec  
        [op op_type [<attribute>=<value>...] ...]]
```

For example, the following command creates a new resource template with the name `BigVM` derived from the `ocf:heartbeat:Xen` resource and some default values and operations:

```
crm(live)configure# rsc_template BigVM ocf:heartbeat:Xen \  
    params mem=1024 op monitor interval=60s
```

```

params allow_mem_management="true" \
op monitor timeout=60s interval=15s \
op stop timeout=10m \
op start timeout=10m

```

Once you defined the new resource template, you can use it in primitives or reference it in order, colocation, or rsc\_ticket constraints. To reference the resource template, use the @ sign:

```

crm(live)configure# primitive MyVM1 @BigVM \
    params xmfile="/etc/xen/shared-vm/MyVM1" name="MyVM1"

```

The new primitive MyVM1 is going to inherit everything from the BigVM resource templates. For example, the equivalent of the above two would be:

```

crm(live)configure# primitive MyVM1 ocf:heartbeat:Xen \
    params xmfile="/etc/xen/shared-vm/MyVM1" name="MyVM1" \
    params allow_mem_management="true" \
    op monitor timeout=60s interval=15s \
    op stop timeout=10m \
    op start timeout=10m

```

If you want to overwrite some options or operations, add them to your (primitive) definition. For instance, the following new primitive MyVM2 doubles the timeout for monitor operations but leaves others untouched:

```

crm(live)configure# primitive MyVM2 @BigVM \
    params xmfile="/etc/xen/shared-vm/MyVM2" name="MyVM2" \
    op monitor timeout=120s interval=30s

```

A resource template may be referenced in constraints to stand for all primitives which are derived from that template. This helps to produce a more concise and clear cluster configuration. Resource template references are allowed in all constraints except location constraints. Colocation constraints may not contain more than one template reference.

## 7.3.3 Creating a STONITH Resource

From the `crm` perspective, a STONITH device is just another resource. To create a STONITH resource, proceed as follows:

- 1 Log in as `root` and start the `crm` interactive shell:

```
root # crm configure
```

- 2 Get a list of all STONITH types with the following command:

```

crm(live)# ra list stonith
apcmaster                apcmastersnmp            apcsmart

```

baytech	bladehpi	cyclades
drac3	external/drac5	external/dracmc-
telnet		
external/hetzner	external/hmchttp	external/ibmrsa
external/ibmrsa-telnet	external/ipmi	external/
ippower9258		
external/kdumpcheck	external/libvirt	external/nut
external/rackpdu	external/riloe	external/sbd
external/vcenter	external/vmware	external/xen0
external/xen0-ha	fence_legacy	ibmhmc
ipmilan	meatware	nw_rpc100s
rcd_serial	rpsl0	suicide
wti_mpc	wti_nps	

### 3 Choose a STONITH type from the above list and view the list of possible options.

Use the following command:

```
crm(live)# ra info stonith:external/ipmi
IPMI STONITH external device (stonith:external/ipmi)
```

ipmitool based power management. Apparently, the power off method of ipmitool is intercepted by ACPI which then makes a regular shutdown. In case of a split brain on a two-node it may happen that no node survives. For two-node clusters use only the reset method.

Parameters (\* denotes required, [] the default):

```
hostname (string): Hostname
    The name of the host to be managed by this STONITH device.
...
```

### 4 Create the STONITH resource with the `stonith` class, the type you have chosen in Step 3, and the respective parameters if needed, for example:

```
crm(live)# configure
crm(live)configure# primitive my-stonith stonith:external/ipmi \
    params hostname="alice" \
    ipaddr="192.168.1.221" \
    userid="admin" passwd="secret" \
    op monitor interval=60m timeout=120s
```

## 7.3.4 Configuring Resource Constraints

Having all the resources configured is only one part of the job. Even if the cluster knows all needed resources, it might still not be able to handle them correctly. For example, try not to mount the file system on the slave node of DRBD (in fact, this would fail with DRBD). Define constraints to make this kind of information available to the cluster.

For more information about constraints, see Section 4.4, “Resource Constraints” (page 73).

### 7.3.4.1 Locational Constraints

The `location` command defines on which nodes a resource may be run, may not be run or is preferred to be run.

This type of constraint may be added multiple times for each resource. All `location` constraints are evaluated for a given resource. A simple example that expresses a preference to run the resource `fs1` on the node with the name `alice` to 100 would be the following:

```
crm(live)configure# location loc-fs1 fs1 100: alice
```

Another example is a location with pingd:

```
crm(live)configure# primitive pingd pingd \  
    params name=pingd dampen=5s multiplier=100 host_list="r1 r2"  
crm(live)configure# location loc-node_pref internal_www \  
    rule 50: #uname eq alice \  
    rule pingd: defined pingd
```

Another use case for location constraints are grouping primitives as a *resource set*. This can be useful if several resources depend on, for example, a ping attribute for network connectivity. In former times, the `-inf/ping` rules needed to be duplicated several times in the configuration, making it unnecessarily complex.

The following example creates a resource set `loc-alice`, referencing the virtual IP addresses `vip1` and `vip2`:

```
crm(live)configure# primitive vip1 ocf:heartbeat:IPaddr2 params  
    ip=192.168.1.5  
crm(live)configure# primitive vip2 ocf:heartbeat:IPaddr2 params  
    ip=192.168.1.6  
crm(live)configure# location loc-alice { vip1 vip2 } inf: alice
```

In some cases it is much more efficient and convenient to use resource patterns for your `location` command. A resource pattern is a regular expression between two slashes. For example, the above virtual IP addresses can be all matched with the following:

```
crm(live)configure# location loc-alice /vip.*/ inf: alice
```

### 7.3.4.2 Colocational Constraints

The `colocation` command is used to define what resources should run on the same or on different hosts.

It is only possible to set a score of either `+inf` or `-inf`, defining resources that must always or must never run on the same node. It is also possible to use non-infinite scores. In that case the colocation is called *advisory* and the cluster may decide not to follow them in favor of not stopping other resources if there is a conflict.

For example, to run the resources with the IDs `filesystem_resource` and `nfs_group` always on the same host, use the following constraint:

```
crm(live)configure# colocation nfs_on_filesystem inf: nfs_group
filesystem_resource
```

For a master slave configuration, it is necessary to know if the current node is a master in addition to running the resource locally.

### 7.3.4.3 Collocating Sets for Resources Without Dependency

Sometimes it is useful to be able to place a group of resources on the same node (defining a colocation constraint), but without having hard dependencies between the resources.

Use the command `weak-bond` if you want to place resources on the same node, but without any action if one of them fails.

```
root # crm configure assist weak-bond RES1 RES2
```

The implementation of `weak-bond` creates a dummy resource and a colocation constraint with the given resources automatically.

### 7.3.4.4 Ordering Constraints

The `order` command defines a sequence of action.

Sometimes it is necessary to provide an order of resource actions or operations. For example, you cannot mount a file system before the device is available to a system. Ordering constraints can be used to start or stop a service right before or after a different resource meets a special condition, such as being started, stopped, or promoted to master.

Use the following command in the `crm` shell to configure an ordering constraint:

```
crm(live)configure# order nfs_after_filesystem mandatory:
filesystem_resource nfs_group
```

## 7.3.4.5 Constraints for the Example Configuration

The example used for this section would not work without additional constraints. It is essential that all resources run on the same machine as the master of the DRBD resource. The DRBD resource must be master before any other resource starts. Trying to mount the DRBD device when it is not the master simply fails. The following constraints must be fulfilled:

- The file system must always be on the same node as the master of the DRBD resource.

```
crm(live)configure# colocation filesystem_on_master inf: \  
    filesystem_resource drbd_resource:Master
```

- The NFS server and the IP address must be on the same node as the file system.

```
crm(live)configure# colocation nfs_with_fs inf: \  
    nfs_group filesystem_resource
```

- The NFS server and the IP address start after the file system is mounted:

```
crm(live)configure# order nfs_second mandatory: \  
    filesystem_resource:start nfs_group
```

- The file system must be mounted on a node after the DRBD resource is promoted to master on this node.

```
crm(live)configure# order drbd_first inf: \  
    drbd_resource:promote filesystem_resource:start
```

## 7.3.5 Specifying Resource Failover Nodes

To determine a resource failover, use the meta attribute migration-threshold. In case failcount exceeds migration-threshold on all nodes, the resource will remain stopped. For example:

```
crm(live)configure# location rsc1-alice rsc1 100: alice
```

Normally, rsc1 prefers to run on alice. If it fails there, migration-threshold is checked and compared to the failcount. If failcount  $\geq$  migration-threshold then it is migrated to the node with the next best preference.

Start failures set the failcount to inf depend on the `start-failure-is-fatal` option. Stop failures cause fencing. If there is no STONITH defined, the resource will not migrate.

For an overview, refer to Section 4.4.4, “Failover Nodes” (page 78).

## 7.3.6 Specifying Resource Failback Nodes (Resource Stickiness)

A resource might fail back to its original node when that node is back online and in the cluster. If you want to prevent a resource from failing back to the node that it was running on, or if you want to specify a different node for the resource to fail back to, change its `resource stickiness` value. You can either specify resource stickiness when you are creating a resource, or afterwards.

For an overview, refer to Section 4.4.5, “Failback Nodes” (page 79).

## 7.3.7 Configuring Placement of Resources Based on Load Impact

Some resources may have specific capacity requirements such as minimum amount of memory. Otherwise, they may fail to start completely or run with degraded performance.

To take this into account, the High Availability Extension allows you to specify the following parameters:

1. The capacity a certain node *provides*.
2. The capacity a certain resource *requires*.
3. An overall strategy for placement of resources.

For detailed background information about the parameters and a configuration example, refer to Section 4.4.6, “Placing Resources Based on Their Load Impact” (page 80).

To configure the resource's requirements and the capacity a node provides, use utilization attributes. You can name the utilization attributes according to your preferences and define as many name/value pairs as your configuration needs. In certain cases, some agents update the utilization themselves, for example the `VirtualDomain`.

In the following example, we assume that you already have a basic configuration of cluster nodes and resources. You now additionally want to configure the capacities a certain node provides and the capacity a certain resource requires.

**Procedure 7.2:** *Adding Or Modifying Utilization Attributes With crm*

- 1 Log in as `root` and start the `crm` interactive shell:

```
root # crm configure
```

- 2 To specify the capacity a node *provides*, use the following command and replace the placeholder `NODE_1` with the name of your node:

```
crm(live)configure# node
NODE_1 utilization memory=16384 cpu=8
```

With these values, `NODE_1` would be assumed to provide 16GB of memory and 8 CPU cores to resources.

- 3 To specify the capacity a resource *requires*, use:

```
crm(live)configure# primitive xen1 ocf:heartbeat:Xen ... \
utilization memory=4096 cpu=4
```

This would make the resource consume 4096 of those memory units from nodeA, and 4 of the CPU units.

- 4 Configure the placement strategy with the `property` command:

```
crm(live)configure# property ...
```

The following values are available:

`default` (default value)

Utilization values are not considered. Resources are allocated according to location scoring. If scores are equal, resources are evenly distributed across nodes.

`utilization`

Utilization values are considered when deciding if a node has enough free capacity to satisfy a resource's requirements. However, load-balancing is still done based on the number of resources allocated to a node.

`minimal`

Utilization values are considered when deciding if a node has enough free capacity to satisfy a resource's requirements. An attempt is made to



concentrate the resources on as few nodes as possible (to achieve power savings on the remaining nodes).

balanced

Utilization values are considered when deciding if a node has enough free capacity to satisfy a resource's requirements. An attempt is made to distribute the resources evenly, thus optimizing resource performance.

---

### **NOTE: Configuring Resource Priorities**

The available placement strategies are best-effort—they do not yet use complex heuristic solvers to always reach optimum allocation results. Ensure that resource priorities are properly set so that your most important resources are scheduled first.

---

## **5 Commit your changes before leaving crmsh:**

```
crm(live) configure#  
commit
```

The following example demonstrates a three node cluster of equal nodes, with 4 virtual machines:

```
crm(live)configure# node alice utilization memory="4000"  
crm(live)configure# node bob utilization memory="4000"  
crm(live)configure# node charly utilization memory="4000"  
crm(live)configure# primitive xenA ocf:heartbeat:Xen \  
utilization hv_memory="3500" meta priority="10" \  
params xmfile="/etc/xen/shared-vm/vm1"  
crm(live)configure# primitive xenB ocf:heartbeat:Xen \  
utilization hv_memory="2000" meta priority="1" \  
params xmfile="/etc/xen/shared-vm/vm2"  
crm(live)configure# primitive xenC ocf:heartbeat:Xen \  
utilization hv_memory="2000" meta priority="1" \  
params xmfile="/etc/xen/shared-vm/vm3"  
crm(live)configure# primitive xenD ocf:heartbeat:Xen \  
utilization hv_memory="1000" meta priority="5" \  
params xmfile="/etc/xen/shared-vm/vm4"  
crm(live)configure# property placement-strategy="minimal"
```

With all three nodes up, xenA will be placed onto a node first, followed by xenD. xenB and xenC would either be allocated together or one of them with xenD.

If one node failed, too little total memory would be available to host them all. xenA would be ensured to be allocated, as would xenD. However, only one of xenB or

xenC could still be placed, and since their priority is equal, the result is not defined yet. To resolve this ambiguity as well, you would need to set a higher priority for either one.

## 7.3.8 Configuring Resource Monitoring

To monitor a resource, there are two possibilities: either define a monitor operation with the `op` keyword or use the `monitor` command. The following example configures an Apache resource and monitors it every 60 seconds with the `op` keyword:

```
crm(live)configure# primitive apache apache \  
  params ... \  
  op monitor interval=60s timeout=30s
```

The same can be done with:

```
crm(live)configure# primitive apache apache \  
  params ...  
crm(live)configure# monitor apache 60s:30s
```

For an overview, refer to Section 4.3, “Resource Monitoring” (page 72).

## 7.3.9 Configuring a Cluster Resource Group

One of the most common elements of a cluster is a set of resources that needs to be located together. Start sequentially and stop in the reverse order. To simplify this configuration we support the concept of groups. The following example creates two primitives (an IP address and an e-mail resource):

- 1 Run the `crm` command as system administrator. The prompt changes to `crm(live)`.

- 2 Configure the primitives:

```
crm(live)# configure  
crm(live)configure# primitive Public-IP ocf:IPaddr:heartbeat \  
  params ip=1.2.3.4 id=p.public-ip  
crm(live)configure# primitive Email lsb:exim \  
  params id=p.lsb-exim
```

### 3 Group the primitives with their relevant identifiers in the correct order:

```
crm(live)configure# group g-shortcut Public-IP Email
```

To change the order of a group member, use the `modgroup` command from the `configure` subcommand. Use the following commands to move the primitive `Email` before `Public-IP`. (This is just to demonstrate the feature):

```
crm(live)configure# modgroup g-shortcut add p.lsb-exim before p.public-ip
```

In case you want to remove a resource from a group (for example, `Email`), use this command:

```
crm(live)configure# modgroup g-shortcut remove p.lsb-exim
```

For an overview, refer to Section 4.2.5.1, “Groups” (page 60).

## 7.3.10 Configuring a Clone Resource

Clones were initially conceived as a convenient way to start *N* instances of an IP resource and have them distributed throughout the cluster for load balancing. They have turned out to be useful for several other purposes, including integrating with DLM, the fencing subsystem and OCFS2. You can clone any resource, provided the resource agent supports it.

Learn more about cloned resources in Section 4.2.5.2, “Clones” (page 62).

### 7.3.10.1 Creating Anonymous Clone Resources

To create an anonymous clone resource, first create a primitive resource and then refer to it with the `clone` command. Do the following:

#### 1 Log in as `root` and start the `crm` interactive shell:

```
root # crm configure
```

#### 2 Configure the primitive, for example:

```
crm(live)configure# primitive Apache lsb:apache
```

#### 3 Clone the primitive:

```
crm(live)configure# clone cl-apache Apache
```

## 7.3.10.2 Creating Stateful/Multi-State Clone Resources

To create a stateful clone resource, first create a primitive resource and then the multi-state resource. The multi-state resource must support at least promote and demote operations.

- 1 Log in as `root` and start the `crm` interactive shell:

```
root # crm configure
```

- 2 Configure the primitive. Change the intervals if needed:

```
crm(live)configure# primitive my-rsc ocf:myCorp:myAppl \  
  op monitor interval=60 \  
  op monitor interval=61 role=Master
```

- 3 Create the multi-state resource:

```
crm(live)configure# ms ms-rsc my-rsc
```

## 7.4 Managing Cluster Resources

Apart from the possibility to configure your cluster resources, the `crm` tool also allows you to manage existing resources. The following subsections gives you an overview.

### 7.4.1 Starting a New Cluster Resource

To start a new cluster resource you need the respective identifier. Proceed as follows:

- 1 Log in as `root` and start the `crm` interactive shell:

```
root #  
crm
```

- 2 Switch to the resource level:

```
crm(live)#  
resource
```

- 3 Start the resource with `start` and press the `→|` key to show all known resources:

```
crm(live)resource# start ID
```

## 7.4.2 Cleaning Up Resources

A resource will be automatically restarted if it fails, but each failure raises the resource's failcount. If a `migration-threshold` has been set for that resource, the node will no longer be allowed to run the resource as soon as the number of failures has reached the migration threshold.

- 1 Open a shell and log in as user `root`.

- 2 Get a list of all your resources:

```
root # crm resource list
...
Resource Group: dlm-clvm:1
    dlm:1 (ocf::pacemaker:controld) Started
    clvm:1 (ocf::lvm2:clvmd) Started
    cmirrord:1 (ocf::lvm2:cmirrord) Started
```

- 3 To clean up the resource `dlm`, for example:

```
root # crm resource cleanup dlm
```

## 7.4.3 Removing a Cluster Resource

Proceed as follows to remove a cluster resource:

- 1 Log in as `root` and start the `crm` interactive shell:

```
root # crm configure
```

- 2 Run the following command to get a list of your resources:

```
crm(live)# resource status
```

For example, the output can look like this (whereas `myIP` is the relevant identifier of your resource):

```
myIP (ocf::IPaddr:heartbeat) ...
```

- 3 Delete the resource with the relevant identifier (which implies a `commit` too):

```
crm(live)# configure delete YOUR_ID
```

#### 4 Commit the changes:

```
crm(live)# configure commit
```

## 7.4.4 Migrating a Cluster Resource

Although resources are configured to automatically fail over (or migrate) to other nodes of the cluster in the event of a hardware or software failure, you can also manually move a resource to another node in the cluster using either the Pacemaker GUI or the command line.

Use the `migrate` command for this task. For example, to migrate the resource `ipaddress1` to a cluster node named `bob`, use these commands:

```
root # crm resource
crm(live)resource# migrate ipaddress1 bob
```

## 7.4.5 Grouping/Tagging Resources

Tags are a way to refer to multiple resources at once, without creating any colocation or ordering relationship between them. This can be useful for grouping conceptually related resources. For example, if you have several resources related to a database, create a tag called `databases` and add all resources related to the database to this tag:

```
root # crm configure databases: db1 db2 db3
```

This allows you to start them all with a single command:

```
root # crm resource start databases
```

Similarly, you can stop them all too:

```
root # crm resource stop databases
```

---

### NOTE: Upgrading the CIB Syntax Version

Tags (for grouping resources) and some ACL features only work with the CIB syntax version `pacemaker-2.0` or higher. For details on how to check this and upgrade the CIB version, see the instructions in the *High Availability Guide* for SUSE Linux Enterprise High Availability Extension 11 SP4, section *Upgrading from SLE HA 11 SP3 to SLE HA 11 SP4*.

---

## 7.4.6 Using Maintenance Mode

Every now and then, you need to perform testing or maintenance tasks on individual cluster components or the whole cluster—be it changing the cluster configuration, updating software packages for individual nodes, or upgrading the cluster to a higher product version.

With regard to that, High Availability Extension provides maintenance options on several levels:

### Applying Maintenance Mode to your Cluster

In case you want to put the whole cluster in maintenance mode, use the following command:

```
root # crm configure property maintenance-mode=true
```

### Applying Maintenance Mode to Nodes

For example, to put the node `alice` into maintenance mode:

```
root # crm node maintenance alice
```

The `crm status` command will show the maintenance mode for `alice` and no more resources will be allocated to that node. To remove the maintenance flag from the node, use:

```
root # crm node ready alice
```

### Applying Maintenance Mode to Resources

If you need to set a specific resource into maintenance mode, use the `meta` command. For example, to put the resource `ipaddress` into maintenance mode, enter:

```
root # crm meta ipaddress set maintenance true
```

---

### **WARNING: Risk of Data Loss**

If you need to execute any testing or maintenance tasks while services are running under cluster control, make sure to follow this outline:

1. Before you start, set the individual resource, node or the whole cluster to maintenance mode. This helps to avoid unwanted side effects like resources not starting in an orderly fashion, the risk of unsynchronized CIBs across the cluster nodes or data loss.
2. Execute your maintenance task or tests.

3. After you have finished, remove the maintenance mode to start normal cluster operation.

---

For more details on what happens to the resources and the cluster while in maintenance mode, see Section 4.7, “Maintenance Mode” (page 88).

## 7.5 Setting Passwords Independent of cib.xml

In case your cluster configuration contains sensitive information, such as passwords, it should be stored in local files. That way, these parameters will never be logged or leaked in support reports.

Before using `secret`, better run the `show` command first to get an overview of all your resources:

```
root # crm configure show
primitive mydb ocf:heartbeat:mysql \
    params replication_user=admin ...
```

If you want to set a password for the above `mydb` resource, use the following commands:

```
root # crm resource secret mydb set passwd linux
INFO: syncing /var/lib/heartbeat/lrm/secrets/mydb/passwd to [your node list]
```

You can get the saved password back with:

```
root # crm resource secret mydb show passwd
linux
```

Note that the parameters need to be synchronized between nodes; the `crm resource secret` command will take care of that. We highly recommend to only use this command to manage secret parameters.

## 7.6 Retrieving History Information

Investigating the cluster history is a complex task. To simplify this task, the `crm` shell contains the `history` command with its subcommands. It is assumed SSH is configured correctly.



Each cluster moves states, migrates resources, or starts important processes. All these actions can be retrieved by subcommands of `history`. Alternatively, use Hawk as explained in Procedure 5.27, “Viewing Transitions with the History Explorer” (page 126).

By default, all `history` commands look at the events of the last hour. To change this time frame, use the `limit` subcommand. The syntax is:

```
root # crm history
crm(live)history# limit FROM_TIME [TO_TIME]
```

Some valid examples include:

```
limit4:00pm , limit16:00
```

Both commands mean the same, today at 4pm.

```
limit2012/01/12 6pm
January 12th 2012 at 6pm
```

```
limit"Sun 5 20:46"
```

In the current year of the current month at Sunday the 5th at 8:46pm

Find more examples and how to create time frames at <http://labix.org/python-dateutil>.

The `info` subcommand shows all the parameters which are covered by the `crm_report`:

```
crm(live)history# info
Source: live
Period: 2012-01-12 14:10:56 - end
Nodes: alice
Groups:
Resources:
```

To limit `crm_report` to certain parameters view the available options with the subcommand `help`.

To narrow down the level of detail, use the subcommand `detail` with a level:

```
crm(live)history# detail 2
```

The higher the number, the more detailed your report will be. Default is 0 (zero).

After you have set above parameters, use `log` to show the log messages.

To display the last transition, use the following command:

```
crm(live)history# transition -l
INFO: fetching new logs, please wait ...
```

This command fetches the logs and runs `dotty` (from the `graphviz` package) to show the transition graph. The shell opens the log file which you can browse with the `↓` and `↑` cursor keys.

If you do not want to open the transition graph, use the `nograph` option:

```
crm(live)history# transition -l nograph
```

## 7.7 For More Information

- The `crm` man page.
- Visit the upstream project documentation at <http://crmsh.github.io/documentation>.
- See Highly Available NFS Storage with DRBD and Pacemaker (↑Highly Available NFS Storage with DRBD and Pacemaker) for an exhaustive example.

# Adding or Modifying Resource Agents

All tasks that need to be managed by a cluster must be available as a resource. There are two major groups here to consider: resource agents and STONITH agents. For both categories, you can add your own agents, extending the abilities of the cluster to your own needs.

## 8.1 STONITH Agents

A cluster sometimes detects that one of the nodes is behaving strangely and needs to remove it. This is called *fencing* and is commonly done with a STONITH resource.

---

**WARNING: External SSH/STONITH Are Not Supported**

It is impossible to know how SSH might react to other system problems. For this reason, external SSH/STONITH agents (like `stonith:external/ssh`) are not supported for production environments. If you still want to use such agents for testing, install the `libglue-devel` package.

---

To get a list of all currently available STONITH devices (from the software side), use the command `crm ra list stonith`. If you do not find your favorite agent, install the `-devel` package.

As of yet there is no documentation about writing STONITH agents. If you want to write new STONITH agents, consult the examples available in the source of the `cluster-glue` package.

## 8.2 Writing OCF Resource Agents

All OCF resource agents (RAs) are available in `/usr/lib/ocf/resource.d/`, see Section 4.2.2, “Supported Resource Agent Classes” (page 57) for more information. Each resource agent must supported the following operations to control it:

```
start
    start or enable the resource

stop
    stop or disable the resource

status
    returns the status of the resource

monitor
    similar to status, but checks also for unexpected states

validate
    validate the resource's configuration

meta-data
    returns information about the resource agent in XML
```

The general procedure of how to create an OCF RA is like the following:

- 1 Load the file `/usr/lib/ocf/resource.d/pacemaker/Dummy` as a template.
- 2 Create a new subdirectory for each new resource agents to avoid naming contradictions. For example, if you have a resource group `kitchen` with the resource `coffee_machine`, add this resource to the directory `/usr/lib/ocf/resource.d/kitchen/`. To access this RA, execute the command `crm: configureprimitive coffee_1 ocf:coffee_machine:kitchen ...`
- 3 Implement the different shell functions and save your file under a different name.

More details about writing OCF resource agents can be found at [http://linux-ha.org/wiki/Resource\\_Agents](http://linux-ha.org/wiki/Resource_Agents). Find special information about several concepts at Chapter 1, *Product Overview* (page 3).

## 8.3 OCF Return Codes and Failure Recovery

According to the OCF specification, there are strict definitions of the exit codes an action must return. The cluster always checks the return code against the expected result. If the result does not match the expected value, then the operation is considered to have failed and a recovery action is initiated. There are three types of failure recovery:

**Table 8.1:** *Failure Recovery Types*

Recovery Type	Description	Action Taken by the Cluster
soft	A transient error occurred.	Restart the resource or move it to a new location.
hard	A non-transient error occurred. The error may be specific to the current node.	Move the resource elsewhere and prevent it from being retried on the current node.
fatal	A non-transient error occurred that will be common to all cluster nodes. This means a bad configuration was specified.	Stop the resource and prevent it from being started on any cluster node.

Assuming an action is considered to have failed, the following table outlines the different OCF return codes and the type of recovery the cluster will initiate when the respective error code is received.

**Table 8.2:** *OCF Return Codes*

<b>OCF Return Code</b>	<b>OCF Alias</b>	<b>Description</b>	<b>Recovery Type</b>
0	OCF_SUCCESS	Success. The command completed successfully. This is the expected result for all start, stop, promote and demote commands.	soft
1	OCF_ERR_- GENERIC	Generic “there was a problem” error code.	soft
2	OCF_ERR_ARGS	The resource’s configuration is not valid on this machine (for example, it refers to a location/tool not found on the node).	hard
3	OCF_- ERR_UN- IMPLEMENTED	The requested action is not implemented.	hard
4	OCF_ERR_PERM	The resource agent does not have sufficient privileges to complete the task.	hard
5	OCF_ERR_- INSTALLED	The tools required by the resource are not installed on this machine.	hard
6	OCF_ERR_- CONFIGURED	The resource’s configuration is invalid (for example, required parameters are missing).	fatal
7	OCF_NOT_- RUNNING	The resource is not running. The cluster will not attempt to stop a resource that returns this for any action.	N/A

<b>OCF Return Code</b>	<b>OCF Alias</b>	<b>Description</b>	<b>Recovery Type</b>
		This OCF return code may or may not require resource recovery—it depends on what is the expected resource status. If unexpected, then <code>soft</code> recovery.	
8	OCF_RUNNING_MASTER	The resource is running in Master mode.	soft
9	OCF_FAILED_MASTER	The resource is in Master mode but has failed. The resource will be demoted, stopped and then started (and possibly promoted) again.	soft
other	N/A	Custom error code.	soft





# Fencing and STONITH

Fencing is a very important concept in computer clusters for HA (High Availability). A cluster sometimes detects that one of the nodes is behaving strangely and needs to remove it. This is called *fencing* and is commonly done with a STONITH resource. Fencing may be defined as a method to bring an HA cluster to a known state.

Every resource in a cluster has a state attached. For example: “resource r1 is started on alice”. In an HA cluster, such a state implies that “resource r1 is stopped on all nodes except alice”, because the cluster must make sure that every resource may be started on only one node. Every node must report every change that happens to a resource. The cluster state is thus a collection of resource states and node states.

When the state of a node or resource cannot be established with certainty, fencing comes in. Even when the cluster is not aware of what is happening on a given node, fencing can ensure that the node does not run any important resources.

## 9.1 Classes of Fencing

There are two classes of fencing: resource level and node level fencing. The latter is the primary subject of this chapter.

### Resource Level Fencing

Using resource level fencing the cluster can ensure that a node cannot access one or more resources. One typical example is a SAN, where a fencing operation changes rules on a SAN switch to deny access from the node.

Resource level fencing can be achieved by using normal resources on which the resource you want to protect depends. Such a resource would simply refuse to start on this node and therefore resources which depend on it will not run on the same node.

### Node Level Fencing

Node level fencing ensures that a node does not run any resources at all. This is usually done in a simple if brutal way: reset or power off the node.

## 9.2 Node Level Fencing

In SUSE® Linux Enterprise High Availability Extension, the fencing implementation is STONITH (Shoot The Other Node in the Head). It provides node level fencing. The High Availability Extension includes the `stonith` command line tool, an extensible interface for remotely powering down a node in the cluster. For an overview of the available options, run `stonith --help` or refer to the man page of `stonith` for more information.

### 9.2.1 STONITH Devices

To use node level fencing, you first need to have a fencing device. To get a list of STONITH devices which are supported by the High Availability Extension, run one of the following commands on any of the nodes:

```
root # stonith -L
```

or

```
root # crm ra list stonith
```

STONITH devices may be classified into the following categories:

#### Power Distribution Units (PDU)

Power Distribution Units are an essential element in managing power capacity and functionality for critical network, server and data center equipment. They can provide remote load monitoring of connected equipment and individual outlet power control for remote power recycling.

#### Uninterruptible Power Supplies (UPS)

A stable power supply provides emergency power to connected equipment by supplying power from a separate source in the event of utility power failure.

### Blade Power Control Devices

If you are running a cluster on a set of blades, then the power control device in the blade enclosure is the only candidate for fencing. Of course, this device must be capable of managing single blade computers.

### Lights-out Devices

Lights-out devices (IBM RSA, HP iLO, Dell DRAC) are becoming increasingly popular and may even become standard in off-the-shelf computers. However, they are inferior to UPS devices, because they share a power supply with their host (a cluster node). If a node stays without power, the device supposed to control it would be useless. In that case, the CRM would continue its attempts to fence the node indefinitely while all other resource operations would wait for the fencing/STONITH operation to complete.

### Testing Devices

Testing devices are used exclusively for testing purposes. They are usually more gentle on the hardware. Before the cluster goes into production, they must be replaced with real fencing devices.

The choice of the STONITH device depends mainly on your budget and the kind of hardware you use.

## 9.2.2 STONITH Implementation

The STONITH implementation of SUSE® Linux Enterprise High Availability Extension consists of two components:

#### stonithd

stonithd is a daemon which can be accessed by local processes or over the network. It accepts the commands which correspond to fencing operations: reset, power-off, and power-on. It can also check the status of the fencing device.

The stonithd daemon runs on every node in the CRM HA cluster. The stonithd instance running on the DC node receives a fencing request from the CRM. It is up to this and other stonithd programs to carry out the desired fencing operation.

#### STONITH Plug-ins

For every supported fencing device there is a STONITH plug-in which is capable of controlling said device. A STONITH plug-in is the interface to the fencing device. The STONITH plug-ins contained in the `cluster-glue` package reside in `/usr/lib/stonith/plugins` (or in `/usr/lib64/`

`stonith/plugins` for 64-bit architectures) on each node. (If you installed the `fence-agents` package, too, the plug-ins contained there are installed in `/usr/sbin/fence_*`). All STONITH plug-ins look the same to `stonithd`, but are quite different on the other side reflecting the nature of the fencing device.

Some plug-ins support more than one device. A typical example is `ipmilan` (or `external/ipmi`) which implements the IPMI protocol and can control any device which supports this protocol.

## 9.3 STONITH Resources and Configuration

To set up fencing, you need to configure one or more STONITH resources—the `stonithd` daemon requires no configuration. All configuration is stored in the CIB. A STONITH resource is a resource of class `stonith` (see Section 4.2.2, “Supported Resource Agent Classes” (page 57)). STONITH resources are a representation of STONITH plug-ins in the CIB. Apart from the fencing operations, the STONITH resources can be started, stopped and monitored, like any other resource. Starting or stopping STONITH resources means loading and unloading the STONITH device driver on a node. Starting and stopping are thus only administrative operations and do not translate to any operation on the fencing device itself. However, monitoring does translate to logging it to the device (to verify that the device will work in case it is needed). When a STONITH resource fails over to another node it enables the current node to talk to the STONITH device by loading the respective driver.

STONITH resources can be configured like any other resource. For details how to do so with your preferred cluster management tool:

- Pacemaker GUI: Section 6.3.2, “Creating STONITH Resources” (page 145)
- Hawk: Section 5.3.3, “Creating STONITH Resources” (page 103)
- `crmsh`: Section 7.3.3, “Creating a STONITH Resource” (page 180)

The list of parameters (attributes) depends on the respective STONITH type. To view a list of parameters for a specific device, use the `stonith` command:

```
stonith -t stonith-device-type -n
```

For example, to view the parameters for the `ibmhmcc` device type, enter the following:

```
stonith -t ibmhmc -n
```

To get a short help text for the device, use the `-h` option:

```
stonith -t stonith-device-type -h
```

## 9.3.1 Example STONITH Resource Configurations

In the following, find some example configurations written in the syntax of the `crm` command line tool. To apply them, put the sample in a text file (for example, `sample.txt`) and run:

```
root # crm < sample.txt
```

For more information about configuring resources with the `crm` command line tool, refer to Chapter 7, *Configuring and Managing Cluster Resources (Command Line)* (page 169).

### **Example 9.1:** *Configuration of an IBM RSA Lights-out Device*

An IBM RSA lights-out device might be configured like this:

```
configure
primitive st-ibmrsa-1 stonith:external/ibmrsa-telnet \
params nodename=alice ipaddr=192.168.0.101 \
userid=USERID passwd=PASSWORD
primitive st-ibmrsa-2 stonith:external/ibmrsa-telnet \
params nodename=bob ipaddr=192.168.0.102 \
userid=USERID passwd=PASSWORD
location l-st-alice st-ibmrsa-1 -inf: alice
location l-st-bob st-ibmrsa-2 -inf: bob
commit
```

In this example, location constraints are used for the following reason: There is always a certain probability that the STONITH operation is going to fail. Therefore, a STONITH operation on the node which is the executioner as well is not reliable. If the node is reset, it cannot send the notification about the fencing operation outcome. The only way to do that is to assume that the operation is going to succeed and send the notification beforehand. But if the operation fails, problems could arise. Therefore, by convention, stonithd refuses to terminate its host.

### **Example 9.2:** *Configuration of a UPS Fencing Device*

The configuration of a UPS type fencing device is similar to the examples above. The details are not covered here. All UPS devices employ the same mechanics for

fencing. How the device is accessed varies. Old UPS devices only had a serial port, usually connected at 1200baud using a special serial cable. Many new ones still have a serial port, but often they also use a USB or Ethernet interface. The kind of connection you can use depends on what the plug-in supports.

For example, compare the `apcmaster` with the `apcsmart` device by using the `stonith -t stonith-device-type -n command`:

```
stonith -t apcmaster -h
```

returns the following information:

```
STONITH Device: apcmaster - APC MasterSwitch (via telnet)
NOTE: The APC MasterSwitch accepts only one (telnet)
connection/session a time. When one session is active,
subsequent attempts to connect to the MasterSwitch will fail.
For more information see http://www.apc.com/
List of valid parameter names for apcmaster STONITH device:
ipaddr
login
password
```

With

```
stonith -t apcsmart -h
```

you get the following output:

```
STONITH Device: apcsmart - APC Smart UPS
(via serial port - NOT USB!).
Works with higher-end APC UPSes, like
Back-UPS Pro, Smart-UPS, Matrix-UPS, etc.
(Smart-UPS may have to be >= Smart-UPS 700?).
See http://www.networkupstools.org/protocols/apcsmart.html
for protocol compatibility details.
For more information see http://www.apc.com/
List of valid parameter names for apcsmart STONITH device:
ttydev
hostlist
```

The first plug-in supports APC UPS with a network port and telnet protocol. The second plug-in uses the APC SMART protocol over the serial line, which is supported by many different APC UPS product lines.

### ***Example 9.3:*** *Configuration of a kdump Device*

Use the `stonith:fence_kdump` resource agent (provided by the package `fence-agents`) to monitor all nodes with the `kdump` function enabled. Find a configuration example for the resource below:

```
configure
```

```
primitive st-kdump stonith:fence_kdump \
params nodename="alice "\ ❶
params pcmk_host_check="dynamic-list" \
params pcmk_reboot_action="off" \
params pcmk_monitor_action="metadata" \
params pcmk_reboot_retries="1" \
params timeout="60"
commit
```

- ❶ Name of the node to be fenced. If you need to monitor more than one node, configure more STONITH resources. To prevent a specific node from using a fencing device, add location constraints.

The fencing action will be started after the timeout of the resource.

In `/etc/sysconfig/kdump` on each node, configure `KDUMP_POSTSCRIPT` to send a notification to all nodes when the `kdump` process is finished. For example:

```
/usr/lib64/fence_kdump_send -i INTERVAL -p PORT -c 1 alice bob charly [...]
```

The node that does a `kdump` will restart automatically after `kdump` has finished. Do not forget to open a port in the firewall for the `fence_kdump` resource. The default port is 7410.

## 9.4 Monitoring Fencing Devices

Like any other resource, the STONITH class agents also support the monitoring operation for checking status.

---

### NOTE: Monitoring STONITH Resources

Monitor STONITH resources regularly, yet sparingly. For most devices a monitoring interval of at least 1800 seconds (30 minutes) should suffice.

---

Fencing devices are an indispensable part of an HA cluster, but the less you need to use them, the better. Power management equipment is often affected by too much broadcast traffic. Some devices cannot handle more than ten or so connections per minute. Some get confused if two clients try to connect at the same time. Most cannot handle more than one session at a time.

Checking the status of fencing devices once every few hours should usually be enough. The probability that a fencing operation needs to be performed and the power switch fails is low.

For detailed information on how to configure monitor operations, refer to Procedure 6.3, “Adding or Modifying Meta and Instance Attributes” (page 144) for the GUI approach or to Section 7.3.8, “Configuring Resource Monitoring” (page 188) for the command line approach.

## 9.5 Special Fencing Devices

In addition to plug-ins which handle real STONITH devices, there are special purpose STONITH plug-ins.

---

### **WARNING: For Testing Only**

Some of the STONITH plug-ins mentioned below are for demonstration and testing purposes only. Do not use any of the following devices in real-life scenarios because this may lead to data corruption and unpredictable results:

- `external/ssh`
- `ssh`

---

#### `fence_kdump`

This plug-in checks if a Kernel dump is in progress on a node. If so, it returns `true`, and acts as if the node has been fenced. The node cannot run any resources during the dump anyway. This avoids fencing a node that is already down but doing a dump, which takes some time. The plug-in must be used in concert with another, real STONITH device.

For configuration details, see Example 9.3, “Configuration of a kdump Device” (page 208).

#### `external/sbd`

This is a self-fencing device. It reacts to a so-called “poison pill” which can be inserted into a shared disk. On shared-storage connection loss, it stops the node from operating. Learn how to use this STONITH agent to implement storage-based fencing in Chapter 17, *Storage Protection*, Section 17.1.3.5, “Configuring the Fencing Resource” (page 292). See also [http://www.linux-ha.org/wiki/SBD\\_Fencing](http://www.linux-ha.org/wiki/SBD_Fencing) for more details.



---

**IMPORTANT: `external/sbd` and DRBD**

The `external/sbd` fencing mechanism requires that the SBD partition is readable directly from each node. Thus, a DRBD\* device must not be used for an SBD partition.

However, you can use the fencing mechanism for a DRBD cluster, provided the SBD partition is located on a shared disk that is not mirrored or replicated.

---

**`external/ssh`**

Another software-based “fencing” mechanism. The nodes must be able to log in to each other as `root` without passwords. It takes a single parameter, `hostlist`, specifying the nodes that it will target. As it is not able to reset a truly failed node, it must not be used for real-life clusters—for testing and demonstration purposes only. Using it for shared storage would result in data corruption.

**`meatware`**

`meatware` requires help from the user to operate. Whenever invoked, `meatware` logs a CRIT severity message which shows up on the node's console. The operator then confirms that the node is down and issues a `meatclient(8)` command. This tells `meatware` to inform the cluster that the node should be considered dead. See `/usr/share/doc/packages/cluster-glue/README.meatware` for more information.

**`suicide`**

This is a software-only device, which can reboot a node it is running on, using the `reboot` command. This requires action by the node's operating system and can fail under certain circumstances. Therefore avoid using this device whenever possible. However, it is safe to use on one-node clusters.

`suicide` is the only exceptions to the “I do not shoot my host” rule.

## 9.6 Basic Recommendations

Check the following list of recommendations to avoid common mistakes:

- Do not configure several power switches in parallel.

- To test your STONITH devices and their configuration, pull the plug once from each node and verify that fencing the node does take place.
- Test your resources under load and verify the timeout values are appropriate. Setting timeout values too low can trigger (unnecessary) fencing operations. For details, refer to Section 4.2.9, “Timeout Values” (page 71).
- Use appropriate fencing devices for your setup. For details, also refer to Section 9.5, “Special Fencing Devices” (page 210).
- Configure one or more STONITH resources. By default, the global cluster option `stonith-enabled` is set to `true`. If no STONITH resources have been defined, the cluster will refuse to start any resources.
- Do not set the global cluster option `stonith-enabled` to `false` for the following reasons:
  - Clusters without STONITH enabled are not supported.
  - DLM/OCFS2 will block forever waiting for a fencing operation that will never happen.
- Do not set the global cluster option `startup-fencing` to `false`. By default, it is set to `true` for the following reason: If a node is in an unknown state during cluster start-up, the node will be fenced once to clarify its status.

## 9.7 For More Information

`/usr/share/doc/packages/cluster-glue`

In your installed system, this directory contains README files for many STONITH plug-ins and devices.

<http://www.linux-ha.org/wiki/STONITH>

Information about STONITH on the home page of the The High Availability Linux Project.

<http://www.clusterlabs.org/doc/>

- *Pacemaker Explained*: Explains the concepts used to configure Pacemaker. Contains comprehensive and very detailed information for reference.

[http://techthoughts.typepad.com/  
managing\\_computers/2007/10/split-brain-quo.html](http://techthoughts.typepad.com/managing_computers/2007/10/split-brain-quo.html)  
Article explaining the concepts of split brain, quorum and fencing in HA  
clusters.



# Access Control Lists

The cluster administration tools like `crm shell (crmsh)`, `Hawk` or the `Pacemaker GUI` can be used by `root` or any user in the group `haclient`. By default, these users have full read/write access. To limit access or assign more fine-grained access rights, you can use *Access control lists (ACLs)*.

Access control lists consist of an ordered set of access rules. Each rule allows read or write access or denies access to a part of the cluster configuration. Rules are typically combined to produce a specific role, then users may be assigned to a role that matches their tasks.

---

## NOTE: CIB Validation Version and ACL Differences

This ACL documentation only applies if your CIB is validated with the CIB syntax version `pacemaker-2.0` or higher. For details on how to check this and upgrade the CIB version, see *Upgrading the CIB Syntax Version* (page 347).

If you have upgraded from SUSE Linux Enterprise High Availability Extension 11 SP3 and kept your former CIB version, refer to the *Access Control List* chapter in the *High Availability Guide for SUSE Linux Enterprise High Availability Extension 11 SP3*. It is available from <http://www.suse.com/documentation/>.

---

# 10.1 Requirements and Prerequisites

Before you start using ACLs on your cluster, make sure the following conditions are fulfilled:

- Ensure you have the same users on all nodes in your cluster, either by using NIS, Active Directory, or by manually adding the same users to all nodes.
- All users for whom you want to modify access rights with ACLs must belong to the `haclient` group.
- All users need to run `crmsh` by its absolute path `/usr/sbin/crm`.
- If non-privileged users want to run `crmsh`, their `PATH` variable needs to be extended with `/usr/sbin`.

---

## IMPORTANT: Default Access Rights

- ACLs are an optional feature. By default, use of ACLs is disabled.
- If ACLs are not enabled, `root` and all users belonging to the `haclient` group have full read/write access to the cluster configuration.
- Even if ACLs are enabled and configured, both `root` and the default CRM owner `hacluster` *always* have full access to the cluster configuration.

---

To use ACLs you need some knowledge about XPath. XPath is a language for selecting nodes in an XML document. Refer to <http://en.wikipedia.org/wiki/XPath> or look into the specification at <http://www.w3.org/TR/xpath/>.

# 10.2 Enabling Use of ACLs In Your Cluster

Before you can start configuring ACLs, you need to *enable* use of ACLs. To do so, use the following command in the `crmsh`:

```
root # crm configure property enable-acl=true
```

Alternatively, use Hawk as described in Procedure 10.1, “Enabling Use of ACLs with Hawk” (page 217).

**Procedure 10.1:** *Enabling Use of ACLs with Hawk*

- 1 Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2 In the left navigation bar, select *Cluster Properties*.
- 3 In the *CRM Configuration* group, select the *enable-acl* attribute from the empty drop-down box and click the plus icon to add it.
- 4 To set `enable-acl=true`, activate the check box next to `enable-acl` and confirm your changes.

## 10.3 The Basics of ACLs

Access control lists consist of an ordered set of access rules. Each rule allows read or write access or denies access to a part of the cluster configuration. Rules are typically combined to produce a specific role, then users may be assigned to a role that matches their tasks. An ACL role is a set of rules which describe access rights to CIB. A rule consists of the following:

- an access right like `read`, `write`, or `deny`
- a specification where to apply the rule. This specification can be a type, an ID reference, or an XPath expression.

Usually, it is convenient to bundle ACLs into roles and assign a specific role to system users (ACL targets). There are two methods to create ACL rules:

- Section 10.3.1, “Setting ACL Rules via XPath Expressions” (page 218). You need to know the structure of the underlying XML to create ACL rules.
- Section 10.3.2, “Setting ACL Rules via Abbreviations” (page 220). Create a shorthand syntax and ACL rules to apply to the matched objects.

## 10.3.1 Setting ACL Rules via XPath Expressions

To manage ACL rules via XPath, you need to know the structure of the underlying XML. Retrieve the structure with the following command that shows your cluster configuration in XML (see Example 10.1):

```
root # crm configure show xml
```

### **Example 10.1:** *Excerpt of a Cluster Configuration in XML*

```
<num_updates="59"
  dc-uuid="175704363"
  crm_feature_set="3.0.9"
  validate-with="pacemaker-2.0"
  epoch="96"
  admin_epoch="0"
  cib-last-written="Fri Aug  8 13:47:28 2014"
  have-quorum="1">
<configuration>
  <crm_config>
    <cluster_property_set id="cib-bootstrap-options">
      <nvpair name="stonith-enabled" value="true" id="cib-bootstrap-
options-stonith-enabled"/>
      <nvpair name="no-quorum-policy" value="ignore" id="cib-bootstrap-
options-no-quorum-policy"/>
      [...]
    </cluster_property_set>
  </crm_config>
  <nodes>
    <node id="175704363" uname="alice"/>
    <node id="175704619" uname="bob"/>
  </nodes>
  <resources> [...] </resources>
  <constraints/>
  <rsc_defaults> [...] </rsc_defaults>
  <op_defaults> [...] </op_defaults>
</configuration>
</cib>
```

With the XPath language you can locate nodes in this XML document. For example, to select the root node (`cib`) use the XPath expression `/cib`. To locate the global cluster configurations, use the XPath expression `/cib/configuration/crm_config`.

As an example, Table 10.1, “Operator Role—Access Types and XPath Expressions” (page 219) shows the parameters (access type and XPath expression)



to create an “operator” role. Users with this role can only execute the tasks mentioned in the second column—they cannot reconfigure any resources (for example, change parameters or operations), nor change the configuration of colocation or ordering constraints.

**Table 10.1:** *Operator Role—Access Types and XPath Expressions*

Type	XPath/Explanation
Write	<pre>//crm_config// nvpair[@name='maintenance-mode']</pre> <p>Turn cluster maintenance mode on or off.</p>
Write	<pre>//op_defaults// nvpair[@name='record-pending']</pre> <p>Choose whether pending operations are recorded.</p>
Write	<pre>//nodes/node// nvpair[@name='standby']</pre> <p>Set node in online or standby mode.</p>
Write	<pre>//resources// nvpair[@name='target-role']</pre> <p>Start, stop, promote or demote any resource.</p>
Write	<pre>//resources// nvpair[@name='maintenance']</pre> <p>Select if a resource should be put to maintenance mode or not.</p>
Write	<pre>//constraints/rsc_location</pre> <p>Migrate/move resources from one node to another.</p>
Read	<pre>/cib</pre>

Type	XPath/Explanation
	View the status of the cluster.

## 10.3.2 Setting ACL Rules via Abbreviations

For users who do not want to deal with the XML structure there is an easier method.

For example, consider the following XPath:

```
//*[@id='rsc1']
```

which locates all the XML nodes with the ID `rsc1`.

The abbreviated syntax is written like this:

```
ref:"rsc1"
```

This also works for constraints. Here is the verbose XPath:

```
//constraints/rsc_location
```

The abbreviated syntax is written like this:

```
type:"rsc_location"
```

The abbreviated syntax can be used in `crmsh` and `Hawk`. The CIB daemon knows how to apply the ACL rules to the matching objects.

## 10.4 Configuring ACLs with the Pacemaker GUI

The following procedure shows how to configure a read-only access to the cluster configuration by defining a `monitor` role and assigning it to a user. Alternatively, you can use `Hawk` or `crmsh` to do so, as described in Section 10.5, “Configuring ACLs with Hawk” (page 222) and Procedure 10.4, “Adding a Monitor Role and Assigning a User with `crmsh`” (page 224).

**Procedure 10.2:** *Adding a Monitor Role and Assigning a User with the Pacemaker GUI*

- 1** Start the Pacemaker GUI and log in as described in Section 6.1.1, “Logging in to a Cluster” (page 138).
- 2** Click the *ACLs* entry in the *Configuration* tree.
- 3** Click *Add*. A dialog box appears. Choose between *ACL Target* and *ACL Role*.
- 4** To define your ACL role(s):
  - 4a** Choose *ACL Role*. A window opens in which you add your configuration options.
  - 4b** Add a unique identifier in the *ID* textfield, for example `monitor`.
  - 4c** Click *Add* and choose the rights (*Read*, *Write*, or *Deny*). In our example, select *Read*.
  - 4d** Enter the XPath expression `/cib` in the *Xpath* textfield. Proceed with *Ok*.

Sidenote: If you have resources or constraints, you can also use the abbreviated syntax as explained in Section 10.3.2, “Setting ACL Rules via Abbreviations” (page 220).
  - 4e** If you have other conditions, repeat the steps (Step 4c (page 221) and Step 4d (page 221)). In our example, this is not the case so your role is finished and you can close the window with *Ok*.
- 5** Assign your role to a user:
  - 5a** Click the *Add* button. A dialog box appears to choose between *ACL Target* and *ACL Role*.
  - 5b** Choose *ACL Target*. A window opens in which you add your configuration options.
  - 5c** Enter the username in the *ID* textfield. Make sure this user belongs to the `haclient` group.
  - 5d** Click and use the role name specified in Step 4b (page 221).

## 10.5 Configuring ACLs with Hawk

The following procedure shows how to configure a read-only access to the cluster configuration by defining a `monitor` role and assigning it to a user. Alternatively, you can use `crmsh` to do so, as described in Procedure 10.4, “Adding a Monitor Role and Assigning a User with `crmsh`” (page 224).

**Procedure 10.3:** *Adding a Monitor Role and Assigning a User with Hawk*

- 1** Start a Web browser and log in to the cluster as described in Section 5.1.1, “Starting Hawk and Logging In” (page 91).
- 2** In the left navigation bar, select *Access Control Lists*. The view shows the *Roles* and *Users* that are already defined.
- 3** To define your ACL role(s):
  - 3a** Select the *Roles* category and click the plus icon.
  - 3b** Enter a unique *Role ID*, for example, `monitor`.
  - 3c** For our example of defining a `monitor` role, select `read` from the *Right* drop-down box.
  - 3d** In the *Xpath* text box, enter the XPath expression `/cib` and click *Create Role*.

This creates a new role with name `monitor`, sets the `read` rights and applies it to all elements in the CIB by using the XPath expression `/cib`.
  - 3e** If necessary, you can add more access rights and XPath arguments by clicking the plus icon and specifying the respective parameters. Confirm your changes.
- 4** Assign the role to a user:
  - 4a** Select the *Users* category and click the plus icon.

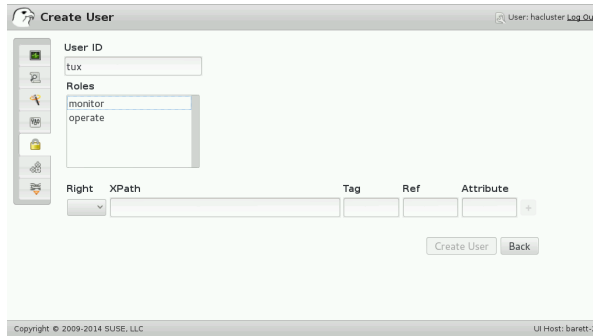
The *Create User* view shows the available roles. It also contains an additional line for configuring individual ACL rules for that user. The view lets you either assign one or multiple roles to the user *or* define one or more individual rules for the user. Selecting a role will make the line for

individual rules disappear and vice versa. Assigning a role plus individual rules is not possible.

- 4b** Enter a unique *User ID*, for example, `tux`. Make sure this user belongs to the `haclient` group.
- 4c** To assign a role to the user, select the respective entries from *Roles*. In our example, select the `monitor` role you have created.

To deselect one or multiple roles, click the respective entries once more. If no role is selected, the line for defining individual rules will appear again.

**Figure 10.1:** *Hawk—Assigning a Role or Rule to a User*



- 4d** If you want to define individual rules instead, select a *Right* and enter the respective Xpath parameters for your rule. Click the plus icon to define additional rules.
- 4e** Confirm your choice and assign the roles or rules to the user by clicking *Create User*.

To configure access rights for resources or constraints, you can also use the abbreviated syntax as explained in Section 10.3.2, “Setting ACL Rules via Abbreviations” (page 220).

## 10.6 Configuring ACLs with `crmsh`

The following procedure shows how to configure a read-only access to the cluster configuration by defining a `monitor` role and assigning it to a user.

**Procedure 10.4:** *Adding a Monitor Role and Assigning a User with crmsh*

**1** Log in as `root`.

**2** Start the interactive mode of `crmsh`:

```
root # crm configure
crm(live)configure#
```

**3** Define your ACL role(s):

**3a** Use the `role` command to define a new role:

```
crm(live)configure# role monitor read xpath: "/cib"
```

The previous command creates a new role with the name `monitor`, sets the `read` rights and applies it to all elements in the CIB by using the XPath expression `/cib`. If necessary, you can add more access rights and XPath arguments.

**3b** Add additional roles as needed.

**4** Assign your roles to one or multiple ACL targets, which are the corresponding system users. Make sure they belong to the `haclient` group.

```
crm(live)configure# acl_target tux monitor
```

**5** Check your changes:

```
crm(live)configure#
show
```

**6** Commit your changes:

```
crm(live)configure#
commit
```

To configure access rights for resources or constraints, you can also use the abbreviated syntax as explained in Section 10.3.2, “Setting ACL Rules via Abbreviations” (page 220).

# Network Device Bonding

For many systems, it is desirable to implement network connections that comply to more than the standard data security or availability requirements of a typical Ethernet device. In these cases, several Ethernet devices can be aggregated to a single bonding device.

The configuration of the bonding device is done by means of bonding module options. The behavior is determined through the mode of the bonding device. By default, this is `mode=active-backup`, which means that a different slave device will become active if the active slave fails.

When using OpenAIS, the bonding device is not managed by the cluster software. Therefore, the bonding device must be configured on each cluster node that might possibly need to access the bonding device.

## 11.1 Configuring Bonding Devices with YaST

To configure a bonding device, you need to have multiple Ethernet devices that can be aggregated to a single bonding device. Proceed as follows:

- 1 Start YaST as `root` and select *Network Devices > Network Settings*.
- 2 In the *Network Settings*, switch to the *Overview* tab, which shows the available devices.

**3** Check if the Ethernet devices to be aggregate to a bonding device have an IP address assigned. If yes, change it:

**3a** Select the device to change and click *Edit*.

**3b** In the *Address* tab of the *Network Card Setup* dialog that opens, select the option *No Link and IP Setup (Bonding Slaves)*.

The screenshot shows the 'Network Card Setup' dialog box with the 'Address' tab selected. The 'Device Type' is set to 'Ethernet' and the 'Configuration Name' is 'eth1'. Under the 'No Link and IP Setup (Bonding Slaves)' section, the 'Dynamic Address' option is selected with 'DHCP' as the method, and 'Use iBFT values' is checked. The 'IP Address' field is empty, 'Subnet Mask' is '255.255.255.0', and 'Hostname' is empty. Below these fields is a table for 'Additional Addresses' with columns for 'Alias Name', 'IP Address', and 'Netmask'. At the bottom of the dialog are 'Add', 'Edit', and 'Delete' buttons for the table, and 'Help', 'Cancel', 'Back', and 'Next' buttons at the very bottom.

**3c** Click *Next* to return to the *Overview* tab in the *Network Settings* dialog.

**4** To add a new bonding device:

**4a** Click *Add* and set the *Device Type* to *Bond*. Proceed with *Next*.

**4b** Select how to assign the IP address to the bonding device. Three methods are at your disposal:

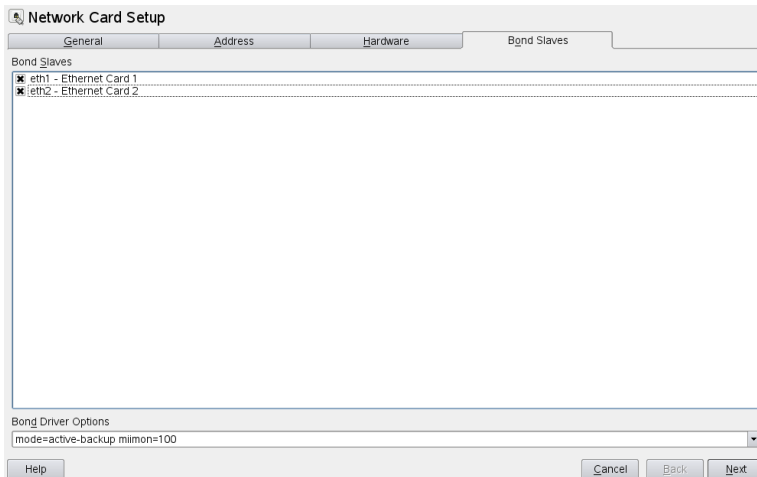
- No Link and IP Setup (Bonding Slaves)
- Dynamic Address (with DHCP or Zeroconf)
- Statically assigned IP Address



Use the method that is appropriate for your environment. If OpenAIS manages virtual IP addresses, select *Statically assigned IP Address* and assign an IP address to the interface.

**4c** Switch to the *Bond Slaves* tab.

**4d** It shows any Ethernet devices that have been configured as bonding slaves in Step 3b. To select the Ethernet devices that you want to include into the bond, activate the check box in front of the relevant *Bond Slave*.



**4e** Edit the *Bond Driver Options*. The following modes are available:

`balance-rr`

Provides load balancing and fault tolerance, at the cost of out-of-order packet transmission. This may cause delays, for example, for TCP reassembly.

`active-backup`

Provides fault tolerance.

`balance-xor`

Provides load balancing and fault tolerance.

`broadcast`

Provides fault tolerance.

`802.3ad`

Provides dynamic link aggregation if supported by the connected switch.

`balance-tlb`

Provides load balancing for outgoing traffic.

`balance-alb`

Provides load balancing for incoming and outgoing traffic, if the network devices used allow the modifying of the network device's hardware address while in use.

**4f** Make sure to add the parameter `miimon=100` to *Bond Driver Options*. Without this parameter, the link is not checked regularly, so the bonding driver might continue to lose packets on a faulty link.

**5** Click *Next* and leave YaST with *OK* to finish the configuration of the bonding device. YaST writes the configuration to `/etc/sysconfig/network/ifcfg-bondDEVICENUMBER`.

## 11.2 Hotplugging of Bonding Slaves

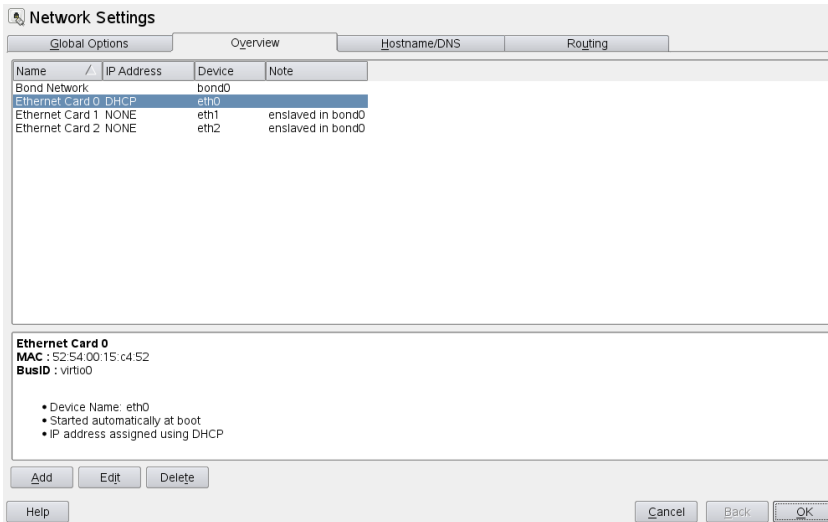
Sometimes it is necessary to replace a bonding slave interface with another one, for example, if the respective network device constantly fails. The solution is to set up hotplugging bonding slaves. It is also necessary to change the `udev` rules to match the device by bus ID instead of by MAC address. This enables you to replace defective hardware (a network card in the same slot but with a different MAC address), if the hardware allows for that.

### **Procedure 11.1:** *Configuring Hotplugging of Bonding Slaves with YaST*

If you prefer manual configuration instead, refer to the SUSE Linux Enterprise Server 11 Administration Guide, chapter *Basic Networking*, section *Hotplugging of Bonding Slaves*.

**1** Start YaST as `root` and select *Network Devices > Network Settings*.

- 2 In the *Network Settings*, switch to the *Overview* tab, which shows the already configured devices. If bonding slaves are already configured, the *Note* column shows it.



- 3 For each of the Ethernet devices that have been aggregated to a bonding device, execute the following steps:
- 3a Select the device to change and click *Edit*. The *Network Card Setup* dialog opens.
  - 3b Switch to the *General* tab and make sure that *Activate device* is set to On Hotplug.
  - 3c Switch to the *Hardware* tab.
  - 3d For the *Udev rules*, click *Change* and select the *BusID* option.
  - 3e Click *OK* and *Next* to return to the *Overview* tab in the *Network Settings* dialog. If you click the Ethernet device entry now, the bottom pane shows the device's details, including the bus ID.
- 4 Click *OK* to confirm your changes and leave the network settings.

At boot time, `/etc/init.d/network` does not wait for the hotplug slaves, but for the bond to become ready, which needs at least one available slave. When one of the slave interfaces is removed from the system (unbind from NIC driver, `rmmmod` of the NIC driver or true PCI hotplug removal), the Kernel removes it from the bond automatically. When a new card is added to the system (replacement of the hardware in the slot), `udev` renames it by applying the bus-based persistent name rule and calls `ifup` for it. The `ifup` call automatically joins it into the bond.

## 11.3 For More Information

All modes and many other options are explained in detail in the *Linux Ethernet Bonding Driver HOWTO*. The file can be found at `/usr/src/linux/Documentation/networking/bonding.txt` after you have installed the package `kernel-source`.

For High Availability setups, the following options described therein are especially important: `miimon` and `use_carrier`.

# Load Balancing with Linux Virtual Server

# 12

The goal of Linux Virtual Server (LVS) is to provide a basic framework that directs network connections to multiple servers that share their workload. Linux Virtual Server is a cluster of servers (one or more load balancers and several real servers for running services) which appears to be one large, fast server to an outside client. This apparent single server is called a *virtual server*. The Linux Virtual Server can be used to build highly scalable and highly available network services, such as Web, cache, mail, FTP, media and VoIP services.

The real servers and the load balancers may be interconnected by either high-speed LAN or by geographically dispersed WAN. The load balancers can dispatch requests to the different servers. They make parallel services of the cluster appear as a virtual service on a single IP address (the virtual IP address or VIP). Request dispatching can use IP load balancing technologies or application-level load balancing technologies. Scalability of the system is achieved by transparently adding or removing nodes in the cluster. High availability is provided by detecting node or daemon failures and reconfiguring the system appropriately.

## 12.1 Conceptual Overview

The following sections give an overview of the main LVS components and concepts.

## 12.1.1 Director

The main component of LVS is the `ip_vs` (or `IPVS`) Kernel code. It implements transport-layer load balancing inside the Linux Kernel (layer-4 switching). The node that runs a Linux Kernel including the `IPVS` code is called *director*. The `IPVS` code running on the director is the essential feature of LVS.

When clients connect to the director, the incoming requests are load-balanced across all cluster nodes: The director forwards packets to the real servers, using a modified set of routing rules that make the LVS work. For example, connections do not originate or terminate on the director, it does not send acknowledgments. The director acts as a specialized router that forwards packets from end users to real servers (the hosts that run the applications that process the requests).

By default, the Kernel does not need the `IPVS` module installed. The `IPVS` Kernel module is included in the `cluster-network-kmp-default` package.

## 12.1.2 User Space Controller and Daemons

The `ldirectord` daemon is a user-space daemon for managing Linux Virtual Server and monitoring the real servers in an LVS cluster of load balanced virtual servers. A configuration file, `/etc/ha.d/ldirectord.cf`, specifies the virtual services and their associated real servers and tells `ldirectord` how to configure the server as an LVS redirector. When the daemon is initialized, it creates the virtual services for the cluster.

By periodically requesting a known URL and checking the responses, the `ldirectord` daemon monitors the health of the real servers. If a real server fails, it will be removed from the list of available servers at the load balancer. When the service monitor detects that the dead server has recovered and is working again, it will add the server back to the list of available servers. In case that all real servers should be down, a fall-back server can be specified to which to redirect a Web service. Typically the fall-back server is `localhost`, presenting an emergency page about the Web service being temporarily unavailable.

The `ldirectord` uses the `ipvsadm` tool (package `ipvsadm`) to manipulate the virtual server table in the Linux Kernel.

## 12.1.3 Packet Forwarding

There are three different methods of how the director can send packets from the client to the real servers:

### Network Address Translation (NAT)

Incoming requests arrive at the virtual IP. They are forwarded to the real servers by changing the destination IP address and port to that of the chosen real server. The real server sends the response to the load balancer which in turn changes the destination IP address and forwards the response back to the client. Thus, the end user receives the replies from the expected source. As all traffic goes through the load balancer, it usually becomes a bottleneck for the cluster.

### IP Tunneling (IP-IP Encapsulation)

IP tunneling enables packets addressed to an IP address to be redirected to another address, possibly on a different network. The LVS sends requests to real servers through an IP tunnel (redirecting to a different IP address) and the real servers reply directly to the client using their own routing tables. Cluster members can be in different subnets.

### Direct Routing

Packets from end users are forwarded directly to the real server. The IP packet is not modified, so the real servers must be configured to accept traffic for the virtual server's IP address. The response from the real server is sent directly to the client. The real servers and load balancers need to be in the same physical network segment.

## 12.1.4 Scheduling Algorithms

Deciding which real server to use for a new connection requested by a client is implemented using different algorithms. They are available as modules and can be adapted to specific needs. For an overview of available modules, refer to the `ipvsadm(8)` man page. Upon receiving a connect request from a client, the director assigns a real server to the client based on a *schedule*. The scheduler is the part of the IPVS Kernel code which decides which real server will get the next new connection.

## 12.2 Configuring IP Load Balancing with YaST

You can configure Kernel-based IP load balancing with the YaST IP Load Balancing module. It is a front-end for `ldirectord`.

To access the IP Load Balancing dialog, start YaST as `root` and select *High Availability > IP Load Balancing*. Alternatively, start the YaST cluster module as `root` on a command line with `yast2 iplb`.

The YaST module writes its configuration to `/etc/ha.d/ldirectord.cf`. The tabs available in the YaST module correspond to the structure of the `/etc/ha.d/ldirectord.cf` configuration file, defining global options and defining the options for the virtual services.

For an example configuration and the resulting processes between load balancers and real servers, refer to Example 12.1, “Simple `ldirectord` Configuration” (page 237).

---

### NOTE: Global Parameters and Virtual Server Parameters

If a certain parameter is specified in both the virtual server section and in the global section, the value defined in the virtual server section overrides the value defined in the global section.

---

#### **Procedure 12.1:** *Configuring Global Parameters*

The following procedure describes how to configure the most important global parameters. For more details about the individual parameters (and the parameters not covered here), click *Help* or refer to the `ldirectord` man page.

- 1 With *Check Interval*, define the interval in which `ldirectord` will connect to each of the real servers to check if they are still online.
- 2 With *Check Timeout*, set the time in which the real server should have responded after the last check.
- 3 With *Failure Count* you can define how many times `ldirectord` will attempt to request the real servers until the check is considered failed.



- 4 With *Negotiate Timeout* define a timeout in seconds for negotiate checks.
- 5 In *Fallback*, enter the host name or IP address of the Web server onto which to redirect a Web service in case all real servers are down.
- 6 If you want the system to send alerts in case the connection status to any real server changes, enter a valid e-mail address in *Email Alert*.
- 7 With *Email Alert Frequency*, define after how many seconds the e-mail alert should be repeated if any of the real servers remains inaccessible.
- 8 In *Email Alert Status* specify the server states for which email alerts should be sent. If you want to define more than one state, use a comma-separated list.
- 9 With *Auto Reload* define, if `ldirectord` should continuously monitor the configuration file for modification. If set to *yes*, the configuration is automatically reloaded upon changes.
- 10 With the *Quiescent* switch, define if to remove failed real servers from the Kernel's LVS table or not. If set to *Yes*, failed servers are not removed. Instead their weight is set to 0 which means that no new connections will be accepted. Already established connections will persist until they time out.
- 11 If you want to use an alternative path for logging, specify a path for the logs in *Log File*. By default, `ldirectord` writes its logs to `/var/log/ldirectord.log`.

**Figure 12.1:** YaST IP Load Balancing—Global Parameters

**IPLB - Global Configuration**

Global Configuration Virtual Server Configuration

Check Interval: 5 Check Timeout: 3 Failure Count: Negotiate Timeout:

Fallback: Email Alert:

Callback: Email Alert Freq:

Execute: Email Alert Status:

Auto Reload: ☒ Quiescent: ☒ Fork: ☒ Supervised: ☒ Log File:

Help Cancel OK

## **Procedure 12.2:** *Configuring Virtual Services*

You can configure one or more virtual services by defining a couple of parameters for each. The following procedure describes how to configure the most important parameters for a virtual service. For more details about the individual parameters (and the parameters not covered here), click *Help* or refer to the `ldirectord` man page.

- 1** In the YaST IP Load Balancing module, switch to the *Virtual Server Configuration* tab.
- 2** *Add* a new virtual server or *Edit* an existing virtual server. A new dialog shows the available options.
- 3** In *Virtual Server* enter the shared virtual IP address (IPv4 or IPv6) and port under which the load balancers and the real servers are accessible as LVS. Instead of IP address and port number you can also specify a host name and a service. Alternatively, you can also use a firewall mark. A firewall mark is a way of aggregating an arbitrary collection of `VIP:port` services into one virtual service.
- 4** To specify the *Real Servers*, you need to enter the IP addresses (IPv4, IPv6, or host names) of the servers, the ports (or service names) and the forwarding method. The forwarding method must either be `gate`, `ipip` or `masq`, see Section 12.1.3, “Packet Forwarding” (page 233).

Click the *Add* button and enter the required arguments for each real server.

- 5** As *Check Type*, select the type of check that should be performed to test if the real servers are still alive. For example, to send a request and check if the response contains an expected string, select `Negotiate`.
- 6** If you have set the *Check Type* to `Negotiate`, you also need to define the type of service to monitor. Select it from the *Service* drop-down list.
- 7** In *Request*, enter the URI to the object that is requested on each real server during the check intervals.
- 8** If you want to check if the response from the real servers contains a certain string (“I’m alive” message), define a regular expression that needs to be matched. Enter the regular expression into *Receive*. If the response from a real server contains this expression, the real server is considered to be alive.

- 9 Depending on the type of *Service* you have selected in Step 6 (page 236), you also need to specify further parameters for authentication. Switch to the *Auth type* tab and enter the details like *Login*, *Password*, *Database*, or *Secret*. For more information, refer to the YaST help text or to the `ldirectord` man page.
- 10 Switch to the *Others* tab.
- 11 Select the *Scheduler* to be used for load balancing. For information on the available schedulers, refer to the `ipvsadm(8)` man page.
- 12 Select the *Protocol* to be used. If the virtual service is specified as an IP address and port, it must be either `tcp` or `udp`. If the virtual service is specified as a firewall mark, the protocol must be `fw`.
- 13 Define further parameters, if needed. Confirm your configuration with *OK*. YaST writes the configuration to `/etc/ha.d/ldirectord.cf`.

**Figure 12.2:** YaST IP Load Balancing—Virtual Services

IPLB - Virtual Servers Configuration

Virtual Server  
192.168.0.200:80

Real Servers  
192.168.0.110:80:gate  
192.168.0.210:80:gate

Add  
Delete  
Edit

Check type    Auth type    Others

Check Type: negotiate    Check Port: http    Service: http    Check Command:   
Http Method:   
Request: test.html    Receive:   
Virtual Host:    Fallback: 127.0.0.1:80

Help    Cancel    OK

### Example 12.1: Simple `ldirectord` Configuration

The values shown in Figure 12.1, “YaST IP Load Balancing—Global Parameters” (page 235) and Figure 12.2, “YaST IP Load Balancing—Virtual

Services” (page 237), would lead to the following configuration, defined in /etc/ha.d/ldirectord.cf:

```
autoreload = yes ❶
checkinterval = 5 ❷
checktimeout = 3 ❸
quiescent = yes ❹
    virtual = 192.168.0.200:80 ❺
    checktype = negotiate ❻
    fallback = 127.0.0.1:80 ❼
    protocol = tcp ❽
    real = 192.168.0.110:80 gate ❾
    real = 192.168.0.120:80 gate ❾
    receive = "still alive" ❿
    request = "test.html" ⓫
    scheduler = wlc ⓬
    service = http ⓭
```

- ❶ Defines that `ldirectord` should continuously check the configuration file for modification.
- ❷ Interval in which `ldirectord` will connect to each of the real servers to check if they are still online.
- ❸ Time in which the real server should have responded after the last check.
- ❹ Defines not to remove failed real servers from the Kernel's LVS table, but to set their weight to 0 instead.
- ❺ Virtual IP address (VIP) of the LVS. The LVS is available at port 80.
- ❻ Type of check that should be performed to test if the real servers are still alive.
- ❼ Server onto which to redirect a Web service all real servers for this service are down.
- ❽ Protocol to be used.
- ❾ Two real servers defined, both available at port 80. The packet forwarding method is `gate`, meaning that direct routing is used.
- ❿ Regular expression that needs to be matched in the response string from the real server.
- ⓫ URI to the object that is requested on each real server during the check intervals.
- ⓬ Selected scheduler to be used for load balancing.
- ⓭ Type of service to monitor.

This configuration would lead to the following process flow: The `ldirectord` will connect to each real server once every 5 seconds (❷) and request `192.168.0.110:80/test.html` or `192.168.0.120:80/test.html` as specified in ❾ and ⓫. If it does not receive the expected `still alive` string (❿) from a real server within 3 seconds (❸) of the last check, it will remove the real

server from the available pool. However, because of the `quiescent=yes` setting (❹), the real server will not be removed from the LVS table, but its weight will be set to 0 so that no new connections to this real server will be accepted. Already established connections will be persistent until they time out.

## 12.3 Further Setup

Apart from the configuration of `ldirectord` with YaST, you need to make sure the following conditions are fulfilled to complete the LVS setup:

- The real servers are set up correctly to provide the needed services.
- The load balancing server (or servers) must be able to route traffic to the real servers using IP forwarding. The network configuration of the real servers depends on which packet forwarding method you have chosen.
- To prevent the load balancing server (or servers) from becoming a single point of failure for the whole system, you need to set up one or several backups of the load balancer. In the cluster configuration, configure a primitive resource for `ldirectord`, so that `ldirectord` can fail over to other servers in case of hardware failure.
- As the backup of the load balancer also needs the `ldirectord` configuration file to fulfill its task, make sure the `/etc/ha.d/ldirectord.cf` is available on all servers that you want to use as backup for the load balancer. You can synchronize the configuration file with `Csync2` as described in Section 3.5.4, “Transferring the Configuration to All Nodes” (page 42).

## 12.4 For More Information

To learn more about Linux Virtual Server, refer to the project home page available at <http://www.linuxvirtualserver.org/>.

For more information about `ldirectord`, refer to its comprehensive man page.



## Geo Clusters (Multi-Site Clusters)

Apart from local clusters and metro area clusters, SUSE® Linux Enterprise High Availability Extension 11 SP4 also supports geographically dispersed clusters (Geo clusters, sometimes also called multi-site clusters). That means you can have multiple, geographically dispersed sites with a local cluster each. Failover between these clusters is coordinated by a higher level entity, the so-called `booth`. Support for Geo clusters is available as a separate extension to Geo Clustering for SUSE Linux Enterprise High Availability Extension. For details on how to use and set up Geo clusters, refer to the Quick Start Geo Clustering for SUSE Linux Enterprise High Availability Extension. It is available from <http://www.suse.com/documentation/> or in your installed system under `/usr/share/doc/manual/sle-ha-geo-manuals_en/`.





## **Part III. Storage and Data Replication**



## OCFS2

Oracle Cluster File System 2 (OCFS2) is a general-purpose journaling file system that has been fully integrated since the Linux 2.6 Kernel. OCFS2 allows you to store application binary files, data files, and databases on devices on shared storage. All nodes in a cluster have concurrent read and write access to the file system. A user-space control daemon, managed via a clone resource, provides the integration with the HA stack, in particular with OpenAIS/Corosync and the Distributed Lock Manager (DLM).

### 14.1 Features and Benefits

OCFS2 can be used for the following storage solutions for example:

- General applications and workloads.
- Xen image store in a cluster. Xen virtual machines and virtual servers can be stored on OCFS2 volumes that are mounted by cluster servers. This provides quick and easy portability of Xen virtual machines between servers.
- LAMP (Linux, Apache, MySQL, and PHP | Perl | Python) stacks.

As a high-performance, symmetric and parallel cluster file system, OCFS2 supports the following functions:

- An application's files are available to all nodes in the cluster. Users simply install it once on an OCFS2 volume in the cluster.

- All nodes can concurrently read and write directly to storage via the standard file system interface, enabling easy management of applications that run across the cluster.
- File access is coordinated through DLM. DLM control is good for most cases, but an application's design might limit scalability if it contends with the DLM to coordinate file access.
- Storage backup functionality is available on all back-end storage. An image of the shared application files can be easily created, which can help provide effective disaster recovery.

OCFS2 also provides the following capabilities:

- Metadata caching.
- Metadata journaling.
- Cross-node file data consistency.
- Support for multiple-block sizes up to 4 KB, cluster sizes up to 1 MB, for a maximum volume size of 4 PB (Petabyte).
- Support for up to 32 cluster nodes.
- Asynchronous and direct I/O support for database files for improved database performance.

## 14.2 OCFS2 Packages and Management Utilities

The OCFS2 Kernel module (`ocfs2`) is installed automatically in the High Availability Extension on SUSE® Linux Enterprise Server 11 SP4. To use OCFS2, make sure the following packages are installed on each node in the cluster: `ocfs2-tools` and the matching `ocfs2-kmp-*` packages for your Kernel.

The `ocfs2-tools` package provides the following utilities for management of OFS2 volumes. For syntax information, see their man pages.

**Table 14.1:** *OCFS2 Utilities*

OCFS2 Utility	Description
debugfs.ocfs2	Examines the state of the OCFS file system for the purpose of debugging.
fsck.ocfs2	Checks the file system for errors and optionally repairs errors.
mkfs.ocfs2	Creates an OCFS2 file system on a device, usually a partition on a shared physical or logical disk.
mounted.ocfs2	Detects and lists all OCFS2 volumes on a clustered system. Detects and lists all nodes on the system that have mounted an OCFS2 device or lists all OCFS2 devices.
tunefs.ocfs2	Changes OCFS2 file system parameters, including the volume label, number of node slots, journal size for all node slots, and volume size.

## 14.3 Configuring OCFS2 Services and a STONITH Resource

Before you can create OCFS2 volumes, you must configure the following resources as services in the cluster: DLM, O2CB and a STONITH resource. OCFS2 uses the cluster membership services from Pacemaker which run in user space. Therefore, DLM and O2CB need to be configured as clone resources that are present on each node in the cluster.

The following procedure uses the `crm` shell to configure the cluster resources. Alternatively, you can also use the Pacemaker GUI to configure the resources.

---

**NOTE: DLM Resource for Both cLVM and OCFS2**

Both cLVM and OCFS2 need a DLM resource that runs on all nodes in the cluster and therefore usually is configured as a clone. If you have a setup that includes both OCFS2 and cLVM, configuring *one* DLM resource for both OCFS2 and cLVM is enough.

---

***Procedure 14.1: Configuring DLM and O2CB Resources***

The configuration consists of a base group that includes several primitives and a base clone. Both base group and base clone can be used in various scenarios afterwards (for both OCFS2 and cLVM, for example). You only need to extend the base group with the respective primitives as needed. As the base group has internal colocation and ordering, this facilitates the overall setup as you do not have to specify several individual groups, clones and their dependencies.

Follow the steps below for one node in the cluster:

- 1** Start a shell and log in as `root` or equivalent.
- 2** Run `crm configure`.
- 3** Enter the following to create the primitive resources for DLM and O2CB:

```
crm(live)configure# primitive dlm ocf:pacemaker:controld \  
    op monitor interval="60" timeout="60" \  
primitive o2cb ocf:ocfs2:o2cb \  
    op monitor interval="60" timeout="60"
```

The `dlm` clone resource controls the distributed lock manager service and makes sure this service is started on all nodes in the cluster. Due to the base group's internal colocation and ordering, the `o2cb` service is only started on nodes where a copy of the `dlm` service is already running.

- 4** Enter the following to create a base group and a base clone:

```
crm(live)configure# group base-group dlm o2cb \  
clone base-clone base-group \  
    meta interleave="true"
```

- 5** Review your changes with `show`.
- 6** If everything is correct, submit your changes with `commit` and leave the `crm` live configuration with `exit`.

## **Procedure 14.2:** *Configuring a STONITH Resource*

---

### **NOTE: STONITH Device Needed**

You need to configure a fencing device. Without a STONITH mechanism (like `external/sbd`) in place the configuration will fail.

---

- 1 Start a shell and log in as `root` or equivalent.
- 2 Create an SBD partition as described in Section 17.1.3.1, “Creating the SBD Partition” (page 288).
- 3 Run `crm configure`.
- 4 Configure `external/sdb` as fencing device with `/dev/sdb2` being a dedicated partition on the shared storage for heartbeating and fencing:

```
crm(live)configure# primitive sbd_stonith stonith:external/sbd \  
    params pcmk_delay_max="15" \  
    op monitor interval="15" timeout="15"
```
- 5 Review your changes with `show`.
- 6 If everything is correct, submit your changes with `commit` and leave the `crm` live configuration with `exit`.

## **14.4 Creating OCFS2 Volumes**

After you have configured DLM and O2CB as cluster resources as described in Section 14.3, “Configuring OCFS2 Services and a STONITH Resource” (page 247), configure your system to use OCFS2 and create OCFS2 volumes.

---

### **NOTE: OCFS2 Volumes for Application and Data Files**

We recommend that you generally store application files and data files on different OCFS2 volumes. If your application volumes and data volumes have different requirements for mounting, it is mandatory to store them on different volumes.

---

Before you begin, prepare the block devices you plan to use for your OCFS2 volumes. Leave the devices as free space.

Then create and format the OCFS2 volume with the `mkfs.ocfs2` as described in Procedure 14.3, “Creating and Formatting an OCFS2 Volume” (page 252). The most important parameters for the command are listed in Table 14.2, “Important OCFS2 Parameters” (page 250). For more information and the command syntax, refer to the `mkfs.ocfs2` man page.

**Table 14.2:** *Important OCFS2 Parameters*

OCFS2 Parameter	Description and Recommendation
Volume Label (-L)	A descriptive name for the volume to make it uniquely identifiable when it is mounted on different nodes. Use the <code>tuneefs.ocfs2</code> utility to modify the label as needed.
Cluster Size (-C)	Cluster size is the smallest unit of space allocated to a file to hold the data. For the available options and recommendations, refer to the <code>mkfs.ocfs2</code> man page.
Number of Node Slots (-N)	<p>The maximum number of nodes that can concurrently mount a volume. For each of the nodes, OCFS2 creates separate system files, such as the journals, for each of the nodes. Nodes that access the volume can be a combination of little-endian architectures (such as x86, x86-64, and ia64) and big-endian architectures (such as ppc64 and s390x).</p> <p>Node-specific files are referred to as local files. A node slot number is appended to the local file. For example: <code>journal:0000</code> belongs</p>



OCFS2 Parameter	Description and Recommendation
	<p>to whatever node is assigned to slot number 0.</p> <p>Set each volume's maximum number of node slots when you create it, according to how many nodes that you expect to concurrently mount the volume. Use the <code>tune fs . ocfs2</code> utility to increase the number of node slots as needed. Note that the value cannot be decreased.</p>
Block Size (-b)	<p>The smallest unit of space addressable by the file system. Specify the block size when you create the volume. For the available options and recommendations, refer to the <code>mkfs . ocfs2</code> man page.</p>
Specific Features On/Off (--fs-features)	<p>A comma separated list of feature flags can be provided, and <code>mkfs . ocfs2</code> will try to create the file system with those features set according to the list. To turn a feature on, include it in the list. To turn a feature off, prepend <code>no</code> to the name.</p> <p>For an overview of all available flags, refer to the <code>mkfs . ocfs2</code> man page.</p>
Pre-Defined Features (--fs-feature-level)	<p>Allows you to choose from a set of pre-determined file system features. For the available options, refer to the <code>mkfs . ocfs2</code> man page.</p>

If you do not specify any specific features when creating and formatting the volume with `mkfs . ocfs2`, the following features are enabled by default: `backup-`

super, sparse, inline-data, unwritten, metaecc, indexed-dirs, and xattr.

**Procedure 14.3:** *Creating and Formatting an OCFS2 Volume*

Execute the following steps only on *one* of the cluster nodes.

- 1 Open a terminal window and log in as `root`.
- 2 Check if the cluster is online with the command `crm status`.
- 3 Create and format the volume using the `mkfs.ocfs2` utility. For information about the syntax for this command, refer to the `mkfs.ocfs2` man page.

For example, to create a new OCFS2 file system on `/dev/sdb1` that supports up to 32 cluster nodes, enter the following commands:

```
root # mkfs.ocfs2 -N 32 /dev/sdb1
```

## 14.5 Mounting OCFS2 Volumes

You can either mount an OCFS2 volume manually or with the cluster manager, as described in Procedure 14.5, “Mounting an OCFS2 Volume with the Cluster Manager” (page 252).

**Procedure 14.4:** *Manually Mounting an OCFS2 Volume*

- 1 Open a terminal window and log in as `root`.
- 2 Check if the cluster is online with the command `crm status`.
- 3 Mount the volume from the command line, using the `mount` command.

---

**WARNING: Manually Mounted OCFS2 Devices**

If you mount the OCFS2 file system manually for testing purposes, make sure to unmount it again before starting to use it by means of OpenAIS.

---

**Procedure 14.5:** *Mounting an OCFS2 Volume with the Cluster Manager*

To mount an OCFS2 volume with the High Availability software, configure an `ocf` file system resource in the cluster. The following procedure uses the `crm` shell to

configure the cluster resources. Alternatively, you can also use the Pacemaker GUI to configure the resources.

- 1 Start a shell and log in as `root` or equivalent.
- 2 Run `crm configure`.
- 3 Configure Pacemaker to mount the OCFS2 file system on every node in the cluster:

```
crm(live)configure# primitive ocfs2-1 ocf:heartbeat:Filesystem \  
    params device="/dev/sdb1" directory="/mnt/shared" fstype="ocfs2" \  
    options="acl" \  
    op monitor interval="20" timeout="40"
```

- 4 Add the file system primitive to the base group you have configured in Procedure 14.1, “Configuring DLM and O2CB Resources” (page 248):

**4a** Enter

```
crm(live)configure# edit base-group
```

- 4b** In the vi editor that opens, modify the group as follows and save your changes:

```
crm(live)configure# group base-group dlm o2cb ocfs2-1
```

Due to the base group's internal colocation and ordering, Pacemaker will only start the `ocfs2-1` resource on nodes that also have an `o2cb` resource already running.

- 5 Review your changes with `show`. To check if you have configured all needed resources, also refer to Appendix C, *Example Configuration for OCFS2 and cLVM* (page 333).
- 6 If everything is correct, submit your changes with `commit` and leave the `crm` live configuration with `exit`.

## 14.6 Using Quotas on OCFS2 File Systems

To use quotas on an OCFS2 file system, create and mount the file system with the appropriate quota features or mount options, respectively: `ursquota` (quota

for individual users) or `grpquota` (quota for groups). These features can also be enabled later on an unmounted file system using `tunefs.ocfs2`.

When a file system has the appropriate quota feature enabled, it tracks in its metadata how much space and files each user (or group) uses. Since OCFS2 treats quota information as file system-internal metadata, you do not need to run the `quotacheck(8)` program. All functionality is built into `fsck.ocfs2` and the file system driver itself.

To enable enforcement of limits imposed on each user or group, run `quotaon(8)` like you would do for any other file system.

For performance reasons each cluster node performs quota accounting locally and synchronizes this information with a common central storage once per 10 seconds. This interval is tunable with `tunefs.ocfs2`, options `usrquota-sync-interval` and `grpquota-sync-interval`. Therefore quota information may not be exact at all times and as a consequence users or groups can slightly exceed their quota limit when operating on several cluster nodes in parallel.

## 14.7 For More Information

For more information about OCFS2, see the following links:

<http://oss.oracle.com/projects/ocfs2/>  
OCFS2 project home page at Oracle.

<http://oss.oracle.com/projects/ocfs2/documentation>  
*OCFS2 User's Guide*, available from the project documentation home page.

# DRBD

The *distributed replicated block device* (DRBD\*) allows you to create a mirror of two block devices that are located at two different sites across an IP network. When used with OpenAIS, DRBD supports distributed high-availability Linux clusters. This chapter shows you how to install and set up DRBD.

## 15.1 Conceptual Overview

DRBD replicates data on the primary device to the secondary device in a way that ensures that both copies of the data remain identical. Think of it as a networked RAID 1. It mirrors data in real-time, so its replication occurs continuously. Applications do not need to know that in fact their data is stored on different disks.

---

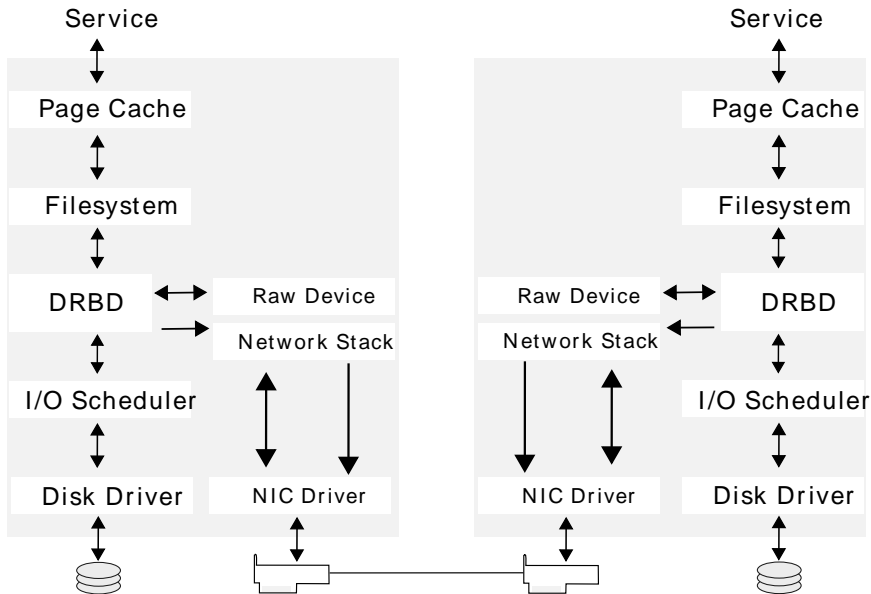
### IMPORTANT: Unencrypted Data

The data traffic between mirrors is not encrypted. For secure data exchange, you should deploy a Virtual Private Network (VPN) solution for the connection.

---

DRBD is a Linux Kernel module and sits between the I/O scheduler at the lower end and the file system at the upper end, see Figure 15.1, “Position of DRBD within Linux” (page 256). To communicate with DRBD, users use the high-level command `drbdadm`. For maximum flexibility DRBD comes with the low-level tool `drbdsetup`.

**Figure 15.1:** *Position of DRBD within Linux*



DRBD allows you to use any block device supported by Linux, usually:

- partition or complete hard disk
- software RAID
- Logical Volume Manager (LVM)
- Enterprise Volume Management System (EVMS)

By default, DRBD uses the TCP ports 7788 and higher for communication between DRBD nodes. Make sure that your firewall does not prevent communication on the used ports.

You must set up the DRBD devices before creating file systems on them. Everything pertaining to user data should be done solely via the `/dev/drbd_N` device and not on the raw device, as DRBD uses the last part of the raw device for metadata. Using the raw device will cause inconsistent data.

With udev integration, you will also get symbolic links in the form `/dev/drbd/by-res/RESOURCES` which are easier to use and provide safety against misremembering the devices' minor number.

For example, if the raw device is 1024 MB in size, the DRBD device has only 1023 MB available for data, with about 70 KB hidden and reserved for the metadata. Any attempt to access the remaining kilobytes via `/dev/drbdN` fails because it is not available for user data.

## 15.2 Installing DRBD Services

To install the needed packages for DRBD, install the High Availability Extension Add-On product on both SUSE Linux Enterprise Server machines in your networked cluster as described in Part I, “Installation and Setup” (page 1). Installing High Availability Extension also installs the DRBD program files.

If you do not need the complete cluster stack but just want to use DRBD, refer to Table 15.1, “DRBD RPM Packages” (page 257). It contains a list of all RPM packages for DRBD. Recently, the `drbd` package has been split into separate packages.

**Table 15.1:** *DRBD RPM Packages*

Filename	Explanation
<code>drbd</code>	Convenience package, split into other
<code>drbd-bash-completion</code>	Programmable bash completion support for <code>drbdadm</code>
<code>drbd-heartbeat</code>	Heartbeat resource agent for DRBD (only needed for Heartbeat)
<code>drbd-kmp-default</code>	Kernel module for DRBD (needed)
<code>drbd-kmp-xen</code>	Xen Kernel module for DRBD
<code>drbd-udev</code>	udev integration scripts for DRBD, managing symlinks to DRBD devices

Filename	Explanation
	in <code>/dev/drbd/by-res</code> and <code>/dev/drbd/by-disk</code>
<code>drbd-utils</code>	Management utilities for DRBD (needed)
<code>drbd-pacemaker</code>	Pacemaker resource agent for DRBD
<code>drbd-xen</code>	Xen block device management script for DRBD
<code>yast2-drbd</code>	YaST DRBD Configuration (recommended)

To simplify the work with `drbdadm`, use the Bash completion support in the RPM package `drbd-bash-completion`. If you want to enable it in your current shell session, insert the following command:

```
root # source /etc/bash_completion.d/drbdadm.sh
```

To use it permanently for `root`, create, or extend a file `/root/.bashrc` and insert the previous line.

## 15.3 Configuring the DRBD Service

### NOTE: Adjustments Needed

The following procedure uses the server names `alice` and `bob`, and the cluster resource name `r0`. It sets up `alice` as the primary node and `/dev/sda1` for storage. Make sure to modify the instructions to use your own nodes and file names.

Before you start configuring DRBD, make sure the block devices in your Linux nodes are ready and partitioned (if needed). The following procedure assumes you have two nodes, `alice` and `bob`, and that they should use the TCP port `7788`. Make sure this port is open in your firewall.



To set up DRBD manually, proceed as follows:

### **Procedure 15.1:** *Manually Configuring DRBD*

- 1** Put your cluster in maintenance mode, if the cluster is already using DRBD:

```
root # crm configure property maintenance-mode=true
```

If you skip this step when your cluster uses already DRBD, a syntax error in the live configuration will lead to a service shutdown.

- 2** Log in as user `root`.

- 3** Change DRBD's configuration files:

- 3a** Open the file `/etc/drbd.conf` and insert the following lines, if they do not exist yet:

```
include "drbd.d/global_common.conf";
include "drbd.d/*.res";
```

Beginning with DRBD 8.3 the configuration file is split into separate files, located under the directory `/etc/drbd.d/`.

- 3b** Open the file `/etc/drbd.d/global_common.conf`. It contains already some pre-defined values. Go to the `startup` section and insert these lines:

```
startup {
    # wfc-timeout degr-wfc-timeout outdated-wfc-timeout
    # wait-after-sb;
    wfc-timeout 100;
    degr-wfc-timeout 120;
}
```

These options are used to reduce the timeouts when booting, see <http://www.drbd.org/users-guide-emb/re-drbdconf.html> for more details.

- 3c** Create the file `/etc/drbd.d/r0.res`, change the lines according to your situation, and save it:

```
resource r0 { ❶
    device /dev/drbd0; ❷
    disk /dev/sdal; ❸
    meta-disk internal; ❹
```

```

on alice { ❸
    address 192.168.1.10:7788; ❹
}
on bob { ❸ (page 260)
    address 192.168.1.11:7788; ❹ (page 260)
}
syncer {
    rate 7M; ❺
}
}

```

- ❶ Name that allows some association to the service that needs them. For example, `nfs`, `http`, `mysql_0`, `postgres_wal`, etc.
- ❷ The device name for DRBD and its minor number.

In the example above, the minor number 0 is used for DRBD. The udev integration scripts will give you a symbolic link `/dev/drbd/by-res/nfs/0`. Alternatively, omit the device node name in the configuration and use the following line instead:

```
drbd0 minor 0 (/dev/ is optional) or /dev/drbd0
```

- ❸ The raw device that is replicated between nodes. Note, in this example the devices are the same on both nodes. If you need different devices, move the `disk` parameter into the `on host`.
- ❹ The meta-disk parameter usually contains the value `internal`, but it is possible to specify an explicit device to hold the meta data. See <http://www.drbd.org/users-guide-emb/ch-internals.html#s-metadata> for more information.
- ❺ The `on` section states which host this configuration statement applies to.
- ❻ The IP address and port number of the respective node. Each resource needs an individual port, usually starting with 7788.
- ❼ The synchronization rate. Set it to one third of the lower of the disk- and network bandwidth. It only limits the resynchronization, not the replication.

- 4 Check the syntax of your configuration file(s). If the following command returns an error, verify your files:

```
root # drbdadm dump all
```

- 5 If you have configured `Csync2` (which should be the default), the DRBD configuration files are already included in the list of files which need to be synchronized. To synchronize them, use:

```
root # csync2 -xv /etc/drbd.d/
```

If you do not have Csync2 (or do not want to use it), copy the DRBD configuration files manually to the other node:

```
root # scp /etc/drbd.conf bob:/etc/
scp /etc/drbd.d/* bob:/etc/drbd.d/
```

- 6** Initialize the meta data on *both* systems by entering the following on each node:

```
root # drbdadm create-md r0
root # rcdbrd start
```

If your disk already contains a file system that you do not need anymore, destroy the file system structure with the following command and repeat this step:

```
root # dd if=/dev/zero of=/dev/sdal count=16 bs=1M
```

- 7** Watch the DRBD status by entering the following on each node:

```
root # rcdbrd status
```

You should get something like this:

```
[... version string omitted ...]
m:res  cs          ro          ds          p
      mounted fstype
0:r0    Connected  Secondary/Secondary  Inconsistent/Inconsistent  C
```

- 8** Start the resync process on your intended primary node (alice in this case):

```
root # drbdadm -- --overwrite-data-of-peer primary r0
```

- 9** Check the status again with `rcdbrd status` and you get:

```
...
m:res  cs          ro          ds          p  mounted fstype
0:r0    Connected  Primary/Secondary  UpToDate/UpToDate  C
```

The status in the `ds` row (disk status) must be `UpToDate` on both nodes.

- 10** Create your file system on top of your DRBD device, for example:

```
root # mkfs.ext3 /dev/drbd/by-res/r0/0
```

- 11** Mount the file system and use it:

```
root # mount /dev/drbd /mnt/
```

## 12 Reset the cluster's maintenance mode flag:

```
root # crm configure property maintenance-mode=false
```

Alternatively, to use YaST to configure DRBD, proceed as follows:

### **Procedure 15.2:** *Using YaST to Configure DRBD*

- 1 Start YaST and select the configuration module *High Availability > DRBD*. If you already have a DRBD configuration, YaST warns you. YaST will change your configuration and will save your old DRBD configuration files as \*.YaSTsave.

Leave the booting flag in *Start-up Configuration > Booting* as it is (by default it is `off`); do not change that as Pacemaker manages this service.

- 2 The actual configuration of the resource is done in *Resource Configuration* (see Figure 15.2, “Resource Configuration” (page 262)).

**Figure 15.2:** *Resource Configuration*

Start-up Configuration  
Global Configuration  
Resource Configuration

### DRBD - Resource Basic Configuration

Resource Name  
r0

**Nodes Configurations**

Node 1	Node 2
Name jupiter	Name benus
Address:Port 192.168.1.10:7788	Address:Port 192.168.1.11:7788
Device /dev/drbd_r0 minor 0	Device /dev/drbd_r0 minor 0
Disk /dev/sda1	Disk /dev/sda1
Meta-disk internal	Meta-disk internal

Advanced Config Cancel OK  
Help Cancel Finish

Press *Add* to create a new resource. The following parameters need to be set twice:

Resource Name	The name of the resource (mandatory)
---------------	--------------------------------------

<i>Name</i>	The host name of the relevant node
<i>Address:Port</i>	The IP address and port number (default 7788) for the respective node
<i>Device</i>	The block device path that is used to access the replicated data. If the device contains a minor number, the associated block device is usually named <code>/dev/drbdX</code> , where <i>X</i> is the device minor number. If the device does not contain a minor number, make sure to add <code>minor 0</code> after the device name.
<i>Disk</i>	The device that is replicated between both nodes.
<i>Meta-disk</i>	<p>The <i>Meta-disk</i> is either set to the value <code>internal</code> or specifies an explicit device extended by an index to hold the meta data needed by DRBD.</p> <p>A real device may also be used for multiple drbd resources. For example, if your <i>Meta-Disk</i> is <code>/dev/sda6[0]</code> for the first resource, you may use <code>/dev/sda6[1]</code> for the second resource. However, there must be at least 128 MB space for each resource available on this disk. The fixed metadata size limits the maximum data size that you can replicate.</p>

All of these options are explained in the examples in the `/usr/share/doc/packages/drbd-utils/drbd.conf` file and in the man page of `drbd.conf(5)`.

- 3 If you have configured Csync2 (which should be the default), the DRBD configuration files are already included in the list of files which need to be synchronized. To synchronize them, use:

```
root # csync2 -xv /etc/drbd.d/
```

If you do not have Csync2 (or do not want to use it), copy the DRBD configuration files manually to the other node (here, another node with the name bob):

```
root # scp /etc/drbd.conf bob:/etc/
```

```
scp /etc/drbd.d/* bob:/etc/drbd.d/
```

- 4** Initialize and start the DRBD service on both systems by entering the following on each node:

```
root # drbdadm create-md r0
root # rcdbrd start
```

- 5** Configure `alice` as the primary node by entering the following on `alice`:

```
root # drbdsetup /dev/drbd0 primary --overwrite-data-of-peer
```

- 6** Check the DRBD service status by entering the following on each node:

```
rcdrbd status
```

Before proceeding, wait until the block devices on both nodes are fully synchronized. Repeat the `rcdrbd status` command to follow the synchronization progress.

- 7** After the block devices on both nodes are fully synchronized, format the DRBD device on the primary with your preferred file system. Any Linux file system can be used. It is recommended to use the `/dev/drbd/by-res/RESOURCE` name.

## 15.4 Testing the DRBD Service

If the install and configuration procedures worked as expected, you are ready to run a basic test of the DRBD functionality. This test also helps with understanding how the software works.

- 1** Test the DRBD service on `alice`.

- 1a** Open a terminal console, then log in as `root`.

- 1b** Create a mount point on `alice`, such as `/srv/r0`:

```
root # mkdir -p /srv/r0
```

- 1c** Mount the `drbd` device:

```
root # mount -o rw /dev/drbd0 /srv/r0
```

- 1d** Create a file from the primary node:

```
root # touch /srv/r0/from_alice
```

**1e** Unmount the disk on alice:

```
root # umount /srv/r0
```

**1f** Downgrade the DRBD service on alice by typing the following command on alice:

```
root # drbdadm secondary
```

**2** Test the DRBD service on bob.

**2a** Open a terminal console, then log in as `root` on bob.

**2b** On bob, promote the DRBD service to primary:

```
root # drbdadm primary
```

**2c** On bob, check to see if bob is primary:

```
root # rcdrbd status
```

**2d** On bob, create a mount point such as `/srv/r0mount`:

```
root # mkdir /srv/r0mount
```

**2e** On bob, mount the DRBD device:

```
root # mount -o rw /dev/drbd_r0 /srv/r0mount
```

**2f** Verify that the file you created on alice exists:

```
root # ls /srv/r0
```

The `/srv/r0mount/from_alice` file should be listed.

**3** If the service is working on both nodes, the DRBD setup is complete.

**4** Set up alice as the primary again.

**4a** Dismount the disk on bob by typing the following command on bob:

```
root # umount /srv/r0
```

**4b** Downgrade the DRBD service on bob by typing the following command on bob:

```
root # drbdadm secondary
```

**4c** On alice, promote the DRBD service to primary:

```
root # drbdadm primary
```

**4d** On alice, check to see if alice is primary:

```
rcdrbd status
```

- 5** To get the service to automatically start and fail over if the server has a problem, you can set up DRBD as a high availability service with OpenAIS. For information about installing and configuring OpenAIS for SUSE Linux Enterprise 11 see Part II, “Configuration and Administration” (page 51).

## 15.5 Tuning DRBD

There are several ways to tune DRBD:

1. Use an external disk for your metadata. This might help, at the cost of maintenance ease.
2. Create a udev rule to change the read-ahead of the DRBD device. Save the following line in the file `/etc/udev/rules.d/82-dm-ra.rules` and change the `read_ahead_kb` value to your workload:

```
ACTION=="add", KERNEL=="dm-*", ATTR{bdi/read_ahead_kb}="4100"
```

This line only works if you use LVM.

3. Tune your network connection, by changing the receive and send buffer settings via `sysctl`.
4. Change the `max-buffers`, `max-epoch-size` or both in the DRBD configuration.
5. Increase the `al-extents` value, depending on your IO patterns.
6. If you have a hardware RAID controller with a BBU (*Battery Backup Unit*), you might benefit from setting `no-disk-flushes`, `no-disk-barrier` and/or `no-md-flushes`.
7. Enable read-balancing depending on your workload. See <http://blogs.linbit.com/p/256/read-balancing/> for more details.



# 15.6 Troubleshooting DRBD

The DRBD setup involves many different components and problems may arise from different sources. The following sections cover several common scenarios and recommend various solutions.

## 15.6.1 Configuration

If the initial DRBD setup does not work as expected, there is probably something wrong with your configuration.

To get information about the configuration:

- 1 Open a terminal console, then log in as `root`.
- 2 Test the configuration file by running `drbdadm` with the `-d` option. Enter the following command:

```
root # drbdadm -d adjust r0
```

In a dry run of the `adjust` option, `drbdadm` compares the actual configuration of the DRBD resource with your DRBD configuration file, but it does not execute the calls. Review the output to make sure you know the source and cause of any errors.

- 3 If there are errors in the `/etc/drbd.d/*` and `drbd.conf` files, correct them before continuing.
- 4 If the partitions and settings are correct, run `drbdadm` again without the `-d` option.

```
root # drbdadm adjust r0
```

This applies the configuration file to the DRBD resource.

## 15.6.2 Host Names

For DRBD, host names are case-sensitive (`Node0` would be a different host than `node0`), and compared to the host name as stored in the Kernel (see the `uname -n` output).

If you have several network devices and want to use a dedicated network device, the host name will likely not resolve to the used IP address. In this case, use the parameter `disable-ip-verification`.

## 15.6.3 TCP Port 7788

If your system cannot connect to the peer, this might be a problem with your local firewall. By default, DRBD uses the TCP port 7788 to access the other node. Make sure that this port is accessible on both nodes.

## 15.6.4 DRBD Devices Broken after Reboot

In cases when DRBD does not know which of the real devices holds the latest data, it changes to a split brain condition. In this case, the respective DRBD subsystems come up as secondary and do not connect to each other. In this case, the following message can be found in the logging data:

```
Split-Brain detected, dropping connection!
```

To resolve this situation, enter the following on the node which has data to be discarded:

```
root # drbdadm secondary r0
root # drbdadm -- --discard-my-data connect r0
```

On the node which has the latest data enter the following:

```
root # drbdadm connect r0
```

That resolves the issue by overwriting one node's data with the peer's data, therefore getting a consistent view on both nodes.

## 15.7 For More Information

The following open source resources are available for DRBD:

- The project home page <http://www.drbd.org>.
- See Highly Available NFS Storage with DRBD and Pacemaker (↑Highly Available NFS Storage with DRBD and Pacemaker).

- [http://clusterlabs.org/wiki/DRBD\\_HowTo\\_1.0](http://clusterlabs.org/wiki/DRBD_HowTo_1.0) by the Linux Pacemaker Cluster Stack Project.
- The following man pages for DRBD are available in the distribution: `drbd(8)`, `drbddisk(8)`, `drbdsetup(8)`, `drbdsetup(8)`, `drbdadm(8)`, `drbd.conf(5)`.
- Find a commented example configuration for DRBD at `/usr/share/doc/packages/drbd/drbd.conf`
- Furthermore, for easier storage administration across your cluster, see the recent announcement about the *DRBD-Manager* at <http://blogs.linbit.com/p/666/drbd-manager>.



# Cluster Logical Volume Manager (cLVM)

# 16

When managing shared storage on a cluster, every node must be informed about changes that are done to the storage subsystem. The Linux Volume Manager 2 (LVM2), which is widely used to manage local storage, has been extended to support transparent management of volume groups across the whole cluster. Clustered volume groups can be managed using the same commands as local storage.

## 16.1 Conceptual Overview

Clustered LVM is coordinated with different tools:

Distributed Lock Manager (DLM)

Coordinates disk access for cLVM.

Logical Volume Manager2 (LVM2)

Enables flexible distribution of one file system over several disks. LVM provides a virtual pool of disk space.

Clustered Logical Volume Manager (cLVM)

Coordinates access to the LVM2 metadata so every node knows about changes. cLVM does not coordinate access to the shared data itself; to enable cLVM to do so, you must configure OCFS2 or other cluster-aware applications on top of the cLVM-managed storage.

## 16.2 Configuration of cLVM

Depending on your scenario it is possible to create a RAID 1 device with cLVM with the following layers:

- **LVM** This is a very flexible solution if you want to increase or decrease your file system size, add more physical storage, or create snapshots of your file systems. This method is described in Section 16.2.3, “Scenario: cLVM With iSCSI on SANs” (page 277).
- **DRBD** This solution only provides RAID 0 (striping) and RAID 1 (mirroring). The last method is described in Section 16.2.4, “Scenario: cLVM With DRBD” (page 280).
- **MD Devices (Linux Software RAID or mdadm)** Although this solution provides all RAID levels, it does not support clusters yet.

Although MD Devices (Linux Software RAID or mdadm) provides all RAID levels, it does not support clusters yet. Therefore it is not covered in the above list.

Make sure you have fulfilled the following prerequisites:

- A shared storage device is available, such as provided by a Fibre Channel, FCoE, SCSI, iSCSI SAN, or DRBD\*.
- In case of DRBD, both nodes must be primary (as described in the following procedure).
- Check if the locking type of LVM2 is cluster-aware. The keyword `locking_type` in `/etc/lvm/lvm.conf` must contain the value 3 (should be the default). Copy the configuration to all nodes, if necessary.

---

### NOTE: Create Cluster Resources First

First create your cluster resources as described in Section 16.2.2, “Creating the Cluster Resources” (page 275) and then your LVM volumes. Otherwise it is impossible to remove the volumes later.

---

## 16.2.1 Configuring Cmirrord

To track mirror log information in a cluster, the `cmirrord` daemon is used. Cluster mirrors are not possible without this daemon running.

We assume that `/dev/sda` and `/dev/sdb` are the shared storage devices. Replace these with your own device name(s), if necessary. Proceed as follows:

**1** Create a cluster with at least two nodes.

**2** Configure your cluster to run `dlm`, `clvmd`, and `STONITH`:

```
root # crm configure
crm(live)configure# primitive clvmd ocf:lvm2:clvmd \
    op stop interval="0" timeout="100" \
    op start interval="0" timeout="90" \
    op monitor interval="20" timeout="20"
crm(live)configure# primitive dlm ocf:pacemaker:controld \
    op start interval="0" timeout="90" \
    op stop interval="0" timeout="100" \
    op monitor interval="60" timeout="60"
crm(live)configure# primitive sbd_stonith stonith:external/sbd
crm(live)configure# group base-group dlm clvmd
crm(live)configure# clone base-clone base-group \
    meta interleave="true"
```

**3** Leave `crmsh` with `exit` and commit your changes.

**4** Create a clustered volume group (VG):

```
root # pvcreate /dev/sda /dev/sdb
root # vgcreate -cy vg /dev/sda /dev/sdb
```

**5** Create a mirrored-log logical volume (LV) in your cluster:

```
root # lvcreate -nlv -m1 -l10%VG vg --mirrorlog mirrored
```

**6** Use `lvs` to show the progress. If the percentage number has reached 100%, the mirrored disk is successfully synchronized.

**7** To test the clustered volume `/dev/vg/lv`, use the following steps:

**7a** Read or write to `/dev/vg/lv`.

**7b** Deactivate your LV with `lvchange -an`.

**7c** Activate your LV with `lvchange -ay`.

**7d** Use `lvconvert` to convert a mirrored log to a disk log.

- 8** Create a mirrored-log LV in another cluster VG. This is a different volume group from the previous one.

The current cLVM can only handle one physical volume (PV) per mirror side. If one mirror is actually made up of several PVs that need to be concatenated or striped, `lvcreate` does not understand this. For this reason, `lvcreate` and `cmirror` metadata needs to understand “grouping” of PVs into one side, effectively supporting RAID10.

To support RAID10 for `cmirror`, use the following procedure (assuming that `/dev/sda` and `/dev/sdb` are the shared storage devices):

- 1** Create a volume group (VG):

```
root # pvcreate /dev/sda /dev/sdb
root # vgcreate vg /dev/sda /dev/sdb
```

- 2** Open the file `/etc/lvm/lvm.conf` and go to the section `allocation`. Set the following line and save the file:

```
mirror_legs_require_separate_pvs = 1
```

- 3** Add your tags to your PVs:

```
root # pvchange --addtag @a /dev/sda
root # pvchange --addtag @b /dev/sdb
```

A tag is an unordered keyword or term assigned to the metadata of a storage object. Tagging allows you to classify collections of LVM storage objects in ways that you find useful by attaching an unordered list of tags to their metadata.

- 4** List your tags:

```
root # pvs -o pv_name,vg_name,pv_tags /dev/sd{a,b}
```

You should receive this output:

```
PV          VG      PV Tags
/dev/sda    vgtest  a
/dev/sdb    vgtest  b
```

If you need further information regarding LVM, refer to the SUSE Linux Enterprise Server 11 SP4 *Storage Administration Guide*, chapter *LVM Configuration*. It is available from <http://www.suse.com/documentation/>.



## 16.2.2 Creating the Cluster Resources

Preparing the cluster for use of cLVM includes the following basic steps:

- Creating a DLM Resource (page 275)
- Creating LVM and cLVM Resources (page 275)

### **Procedure 16.1:** *Creating a DLM Resource*

---

#### **NOTE: DLM Resource for Both cLVM and OCFS2**

Both cLVM and OCFS2 need a DLM resource that runs on all nodes in the cluster and therefore is usually configured as a clone. If you have a setup that includes both OCFS2 and cLVM, configuring *one* DLM resource for both OCFS2 and cLVM is enough.

---

- 1 Start a shell and log in as `root`.
- 2 Run `crm configure`.
- 3 Check the current configuration of the cluster resources with `show`.
- 4 If you have already configured a DLM resource (and a corresponding base group and base clone), continue with Procedure 16.2, “Creating LVM and cLVM Resources” (page 275).

Otherwise, configure a DLM resource and a corresponding base group and base clone as described in Procedure 14.1, “Configuring DLM and O2CB Resources” (page 248).

- 5 Leave the `crm` live configuration with `exit`.

### **Procedure 16.2:** *Creating LVM and cLVM Resources*

- 1 Start a shell and log in as `root`.
- 2 Run `crm configure`.
- 3 Configure a cLVM resource as follows:

```
crm(live)configure# primitive clvm ocf:lvm2:clvmd \
```

```
params daemon_timeout="30"
```

- 4 Configure an LVM resource for the volume group as follows:

```
crm(live)configure# primitive vg1 ocf:heartbeat:LVM \  
params volgrpname="cluster-vg" \  
op monitor interval="60" timeout="60"
```

- 5 If you want the volume group to be activated exclusively on *one* node, configure the LVM resource as described below and omit Step 6 (page 276):

```
crm(live)configure# primitive vg1 ocf:heartbeat:LVM \  
params volgrpname="cluster-vg" exclusive="yes" \  
op monitor interval="60" timeout="60"
```

In this case, cLVM will protect all logical volumes within the VG from being activated on multiple nodes, as an additional measure of protection for non-clustered applications.

- 6 To ensure that the cLVM and LVM resources are activated cluster-wide, add both primitives to the base group you have created in Procedure 14.1, “Configuring DLM and O2CB Resources” (page 248):

**6a** Enter

```
crm(live)configure# edit base-group
```

- 6b** In the vi editor that opens, modify the group as follows and save your changes:

```
crm(live)configure# group base-group dlm clvm vg1 ocfs2-1
```

---

### IMPORTANT: Setup Without OCFS2

If your setup does not include OCFS2, omit the `ocfs2-1` primitive from the base group. The `oc2cb` primitive can be configured and included in the group anyway, regardless of whether you use OCFS2 or not.

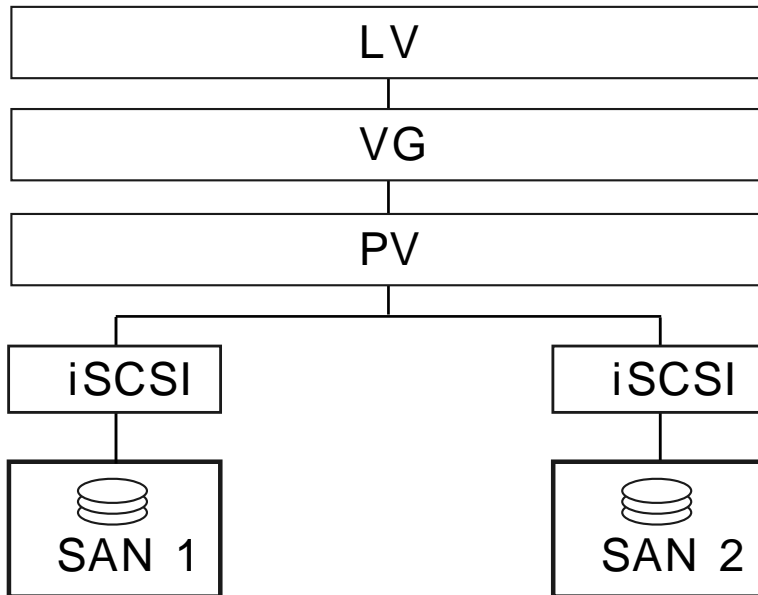
---

- 7 Review your changes with `show`. To check if you have configured all needed resources, also refer to Appendix C, *Example Configuration for OCFS2 and cLVM* (page 333).
- 8 If everything is correct, submit your changes with `commit` and leave the `crm` live configuration with `exit`.

## 16.2.3 Scenario: cLVM With iSCSI on SANs

The following scenario uses two SAN boxes which export their iSCSI targets to several clients. The general idea is displayed in Figure 16.1, “Setup of iSCSI with cLVM” (page 277).

**Figure 16.1:** *Setup of iSCSI with cLVM*



---

### **WARNING: Data Loss**

The following procedures will destroy any data on your disks!

---

Configure only one SAN box first. Each SAN box needs to export its own iSCSI target. Proceed as follows:

#### **Procedure 16.3:** *Configuring iSCSI Targets (SAN)*

- 1 Run YaST and click *Network Services > iSCSI Target* to start the iSCSI Server module.

- 2 If you want to start the iSCSI target whenever your computer is booted, choose *When Booting*, otherwise choose *Manually*.
- 3 If you have a firewall running, enable *Open Port in Firewall*.
- 4 Switch to the *Global* tab. If you need authentication enable incoming or outgoing authentication or both. In this example, we select *No Authentication*.
- 5 Add a new iSCSI target:
  - 5a Switch to the *Targets* tab.
  - 5b Click *Add*.
  - 5c Enter a target name. The name needs to be formatted like this:  
`iqn.DATE.DOMAIN`  
  
For more information about the format, refer to *Section 3.2.6.3.1. Type "iqn." (iSCSI Qualified Name)* at <http://www.ietf.org/rfc/rfc3720.txt>.
  - 5d If you want a more descriptive name, you can change it as long as your identifier is unique for your different targets.
  - 5e Click *Add*.
  - 5f Enter the device name in *Path* and use a *Scsiid*.
  - 5g Click *Next* twice.
- 6 Confirm the warning box with *Yes*.
- 7 Open the configuration file `/etc/iscsi/iscsi.conf` and change the parameter `node.startup` to `automatic`.

Now set up your iSCSI initiators as follows:

**Procedure 16.4:** *Configuring iSCSI Initiators*

- 1 Run YaST and click *Network Services > iSCSI Initiator*.

- 2 If you want to start the iSCSI initiator whenever your computer is booted, choose *When Booting*, otherwise set *Manually*.
- 3 Change to the *Discovery* tab and click the *Discovery* button.
- 4 Add your IP address and your port of your iSCSI target (see Procedure 16.3, “Configuring iSCSI Targets (SAN)” (page 277)). Normally, you can leave the port as it is and use the default value.
- 5 If you use authentication, insert the incoming and outgoing user name and password, otherwise activate *No Authentication*.
- 6 Select *Next*. The found connections are displayed in the list.
- 7 To test if the iSCSI initiator has been started successfully, select one of the displayed targets and click *Log In*.

**Procedure 16.5:** *Creating the LVM Volume Groups*

- 1 Open a `root` shell on one of the nodes you have run the iSCSI initiator from Procedure 16.4, “Configuring iSCSI Initiators” (page 278).
- 2 Prepare the physical volume for LVM with the command `pvcreate` on the disks `/dev/sdd` and `/dev/sde`:  

```
root # pvcreate /dev/sdd
root # pvcreate /dev/sde
```
- 3 Create the cluster-aware volume group on both disks:  

```
root # vgcreate --clustered y clustervg /dev/sdd /dev/sde
```
- 4 Create logical volumes as needed:  

```
root # lvcreate --name clusterlv --size 500M clustervg
```
- 5 Check the physical volume with `pvdisplay`:

```
--- Physical volume ---
PV Name               /dev/sdd
VG Name               clustervg
PV Size               509,88 MB / not usable 1,88 MB
Allocatable           yes
PE Size (KByte)       4096
Total PE              127
Free PE               127
Allocated PE           0
PV UUID               52okH4-nv3z-2AUL-GhAN-8DAZ-GMtU-Xrn9Kh
```

```

--- Physical volume ---
PV Name                /dev/sde
VG Name                clustervg
PV Size                509,84 MB / not usable 1,84 MB
Allocatable            yes
PE Size (KByte)        4096
Total PE               127
Free PE                127
Allocated PE           0
PV UUID                Ouj3Xm-AI58-1xB1-mWm2-xn51-agM2-0UuHFC

```

## 6 Check the volume group with vgdisplay:

```

--- Volume group ---
VG Name                clustervg
System ID
Format                 lvm2
Metadata Areas         2
Metadata Sequence No   1
VG Access               read/write
VG Status               resizable
Clustered              yes
Shared                 no
MAX LV                 0
Cur LV                 0
Open LV                 0
Max PV                 0
Cur PV                 2
Act PV                 2
VG Size                 1016,00 MB
PE Size                 4,00 MB
Total PE               254
Alloc PE / Size         0 / 0
Free PE / Size          254 / 1016,00 MB
VG UUID                UCyWw8-2jqV-enuT-KH4d-NXQI-JhH3-J24anD

```

After you have created the volumes and started your resources you should have a new device named `/dev/dm-*`. It is recommended to use a clustered file system on top of your LVM resource, for example OCFS. For more information, see Chapter 14, *OCFS2* (page 245).

## 16.2.4 Scenario: cLVM With DRBD

The following scenarios can be used if you have data centers located in different parts of your city, country, or continent.

### **Procedure 16.6:** *Creating a Cluster-Aware Volume Group With DRBD*

- 1 Create a primary/primary DRBD resource:

**1a** First, set up a DRBD device as primary/secondary as described in Procedure 15.1, “Manually Configuring DRBD” (page 259). Make sure the disk state is up-to-date on both nodes. Check this with `cat /proc/drbd` or with `rcdrbd status`.

**1b** Add the following options to your configuration file (usually something like `/etc/drbd.d/r0.res`):

```
resource r0 {
    startup {
        become-primary-on both;
    }

    net {
        allow-two-primaries;
    }
    ...
}
```

**1c** Copy the changed configuration file to the other node, for example:

```
root # scp /etc/drbd.d/r0.res venus:/etc/drbd.d/
```

**1d** Run the following commands on *both* nodes:

```
root # drbdadm disconnect r0
root # drbdadm connect r0
root # drbdadm primary r0
```

**1e** Check the status of your nodes:

```
root # cat /proc/drbd
...
0: cs:Connected ro:Primary/Primary ds:UpToDate/UpToDate C r----
```

**2** Include the `clvmd` resource as a clone in the pacemaker configuration, and make it depend on the DLM clone resource. See Procedure 16.1, “Creating a DLM Resource” (page 275) for detailed instructions. Before proceeding, confirm that these resources have started successfully on your cluster. You may use `crm_mon` or the Web interface to check the running services.

**3** Prepare the physical volume for LVM with the command `pvcreate`. For example, on the device `/dev/drbd_r0` the command would look like this:

```
root # pvcreate /dev/drbd_r0
```

**4** Create a cluster-aware volume group:

```
root # vgcreate --clustered y myclusterfs /dev/drbd_r0
```

- 5 Create logical volumes as needed. You may probably want to change the size of the logical volume. For example, create a 4 GB logical volume with the following command:

```
root # lvcreate --name testlv -L 4G myclusterfs
```

- 6 The logical volumes within the VG are now available as file system mounts or raw usage. Ensure that services using them have proper dependencies to collocate them with and order them after the VG has been activated.

After finishing these configuration steps, the LVM2 configuration can be done like on any stand-alone workstation.

## 16.3 Configuring Eligible LVM2 Devices Explicitly

When several devices seemingly share the same physical volume signature (as can be the case for multipath devices or DRBD), it is recommended to explicitly configure the devices which LVM2 scans for PVs.

For example, if the command `vgcreate` uses the physical device instead of using the mirrored block device, DRBD will be confused which may result in a split brain condition for DRBD.

To deactivate a single device for LVM2, do the following:

- 1 Edit the file `/etc/lvm/lvm.conf` and search for the line starting with `filter`.
- 2 The patterns there are handled as regular expressions. A leading “a” means to accept a device pattern to the scan, a leading “r” rejects the devices that follow the device pattern.
- 3 To remove a device named `/dev/sdb1`, add the following expression to the filter rule:

```
"r|^/dev/sdb1$|"
```

The complete filter line will look like the following:



```
filter = [ "r|^/dev/sdb1$|", "r|/dev/.*/by-path/.*/", "r|/dev/.*/by-id/.*/", "a/.*/" ]
```

A filter line, that accepts DRBD and MPIO devices but rejects all other devices would look like this:

```
filter = [ "a|/dev/drbd.*|", "a|/dev/.*/by-id/dm-uuid-mpath-.*/",  
  "r/.*/" ]
```

- 4 Write the configuration file and copy it to all cluster nodes.

## 16.4 For More Information

Thorough information is available from the pacemaker mailing list, available at <http://www.clusterlabs.org/wiki/Help:Contents>.

The official cLVM FAQ can be found at <http://sources.redhat.com/cluster/wiki/FAQ/CLVM>.



# Storage Protection

The High Availability cluster stack's highest priority is protecting the integrity of data. This is achieved by preventing uncoordinated concurrent access to data storage: For example, Ext3 file systems are only mounted once in the cluster, OCFS2 volumes will not be mounted unless coordination with other cluster nodes is available. In a well-functioning cluster Pacemaker will detect if resources are active beyond their concurrency limits and initiate recovery. Furthermore, its policy engine will never exceed these limitations.

However, network partitioning or software malfunction could potentially cause scenarios where several coordinators are elected. If this so-called split brain scenarios were allowed to unfold, data corruption might occur. Hence, several layers of protection have been added to the cluster stack to mitigate this.

The primary component contributing to this goal is IO fencing/STONITH since it ensures that all other access prior to storage activation is terminated. Other mechanisms are cLVM2 exclusive activation or OCFS2 file locking support to protect your system against administrative or application faults. Combined appropriately for your setup, these can reliably prevent split brain scenarios from causing harm.

This chapter describes an IO fencing mechanism that leverages the storage itself, followed by the description of an additional layer of protection to ensure exclusive storage access. These two mechanisms can be combined for higher levels of protection.

# 17.1 Storage-based Fencing

You can reliably avoid split brain scenarios by using one or more STONITH Block Devices (SBD), `watchdog` support and the `external/sbd` STONITH agent.

## 17.1.1 Overview

In an environment where all nodes have access to shared storage, a small partition of the device is formatted for use with SBD. The size of the partition depends on the block size of the used disk (1 MB for standard SCSI disks with 512 Byte block size; DASD disks with 4 kB block size need 4 MB). After the respective daemon is configured, it is brought online on each node before the rest of the cluster stack is started. It is terminated after all other cluster components have been shut down, thus ensuring that cluster resources are never activated without SBD supervision.

The daemon automatically allocates one of the message slots on the partition to itself, and constantly monitors it for messages addressed to itself. Upon receipt of a message, the daemon immediately complies with the request, such as initiating a power-off or reboot cycle for fencing.

The daemon constantly monitors connectivity to the storage device, and terminates itself in case the partition becomes unreachable. This guarantees that it is not disconnected from fencing messages. If the cluster data resides on the same logical unit in a different partition, this is not an additional point of failure: The work-load will terminate anyway if the storage connectivity has been lost.

Increased protection is offered through `watchdog` support. Modern systems support a `hardware watchdog` that needs to be “tickled” or “fed” by a software component. The software component (usually a daemon) regularly writes a service pulse to the watchdog—if the daemon stops feeding the watchdog, the hardware will enforce a system restart. This protects against failures of the SBD process itself, such as dying, or becoming stuck on an IO error.

If Pacemaker integration is activated, SBD will not self-fence if device majority is lost. For example, your cluster contains 3 nodes: A, B, and C. Because of a network split, A can only see itself while B and C can still communicate. In this case, there are two cluster partitions, one with quorum because of being the majority (B, C), and one without (A). If this happens while the majority of fencing devices are unreachable, node A would immediately commit suicide, but the nodes B and C would continue to run.

## 17.1.2 Number of SBD Devices

SBD supports the use of 1-3 devices:

### One Device

The most simple implementation. It is appropriate for clusters where all of your data is on the same shared storage.

### Two Devices

This configuration is primarily useful for environments that use host-based mirroring but where no third storage device is available. SBD will not terminate itself if it loses access to one mirror leg, allowing the cluster to continue.

However, since SBD does not have enough knowledge to detect an asymmetric split of the storage, it will not fence the other side while only one mirror leg is available. Thus, it cannot automatically tolerate a second failure while one of the storage arrays is down.

### Three Devices

The most reliable configuration. It is resilient against outages of one device—be it because of failures or maintenance. SBD will only terminate itself if more than one device is lost. Fencing messages can be successfully be transmitted if at least two devices are still accessible.

This configuration is suitable for more complex scenarios where storage is not restricted to a single array. Host-based mirroring solutions can have one SBD per mirror leg (not mirrored itself), and an additional tie-breaker on iSCSI.

## 17.1.3 Setting Up Storage-based Protection

The following steps are necessary to set up storage-based protection:

- 1 Creating the SBD Partition (page 288)
- 2 Setting Up the Software Watchdog (page 289)
- 3 Starting the SBD Daemon (page 291)
- 4 Testing SBD (page 292)

## 5 Configuring the Fencing Resource (page 292)

All of the following procedures must be executed as `root`. Before you start, make sure the following requirements are met:

---

### IMPORTANT: Requirements

- The environment must have shared storage reachable by all nodes.
  - The shared storage segment must not use host-based RAID, cLVM2, nor DRBD\*.
  - However, using storage-based RAID and multipathing is recommended for increased reliability.
- 

## 17.1.3.1 Creating the SBD Partition

It is recommended to create a 1MB partition at the start of the device. If your SBD device resides on a multipath group, you need to adjust the timeouts SBD uses, as MPIO's path down detection can cause some latency. After the `msgwait` timeout, the message is assumed to have been delivered to the node. For multipath, this should be the time required for MPIO to detect a path failure and switch to the next path. You may need to test this in your environment. The node will terminate itself if the SBD daemon running on it has not updated the watchdog timer fast enough. Test your chosen timeouts in your specific environment. In case you use a multipath storage with just one SBD device, pay special attention to the failover delays incurred.

---

### NOTE: Device Name for SBD Partition

In the following, this SBD partition is referred to by `/dev/SBD`. Replace it with your actual path name, for example: `/dev/sdc1`.

---

### IMPORTANT: Overwriting Existing Data

Make sure the device you want to use for SBD does not hold any data. The `sbd` command will overwrite the device without further requests for confirmation.

---

- 1 Initialize the SBD device with the following command:

```
root # sbd -d /dev/SBD create
```

This will write a header to the device, and create slots for up to 255 nodes sharing this device with default timings.

If you want to use more than one device for SBD, provide the devices by specifying the `-d` option multiple times, for example:

```
root # sbd -d /dev/SBD1 -d /dev/SBD2 -d /dev/SBD3 create
```

- 2** If your SBD device resides on a multipath group, adjust the timeouts SBD uses. This can be specified when the SBD device is initialized (all timeouts are given in seconds):

```
root # /usr/sbin/sbd -d /dev/SBD -4 180❶ -l 90❷ create
```

- ❶ The `-4` option is used to specify the `msgwait` timeout. In the example above, it is set to 180 seconds.
- ❷ The `-l` option is used to specify the `watchdog` timeout. In the example above, it is set to 90 seconds.

- 3** With the following command, check what has been written to the device:

```
root # sbd -d /dev/SBD dump
Header version      : 2
Number of slots     : 255
Sector size         : 512
Timeout (watchdog)  : 5
Timeout (allocate)  : 2
Timeout (loop)      : 1
Timeout (msgwait)   : 10
```

As you can see, the timeouts are also stored in the header, to ensure that all participating nodes agree on them.

## 17.1.3.2 Setting Up the Software Watchdog

Watchdog will protect the system against SBD failures, if no other software uses it.

---

### IMPORTANT: Accessing the Watchdog Timer

No other software must access the watchdog timer. Some hardware vendors ship systems management software that uses the watchdog for system resets (for example, HP ASR daemon). Disable such software, if watchdog is used by SBD.

---

In SUSE Linux Enterprise High Availability Extension, watchdog support in the Kernel is enabled by default: It ships with several different Kernel modules that provide hardware-specific watchdog drivers. The High Availability Extension uses the SBD daemon as software component that “feeds” the watchdog. If configured as described in Section 17.1.3.3, “Starting the SBD Daemon” (page 291), the SBD daemon will start automatically when the respective node is brought online with `rcopenais start`.

Usually, the appropriate watchdog driver for your hardware is automatically loaded during system boot. `softdog` is the most generic driver, but it is recommended to use a driver with actual hardware integration. For example:

- On HP hardware, this is the `hpwdt` driver.
- For systems with an Intel TCO, the `iTCO_wdt` driver can be used.

For a list of choices, refer to `/usr/src/KERNEL_VERSION/drivers/watchdog`. Alternatively, list the drivers that have been installed with your Kernel version with the following command:

```
root # rpm -ql kernel-VERSION | grep watchdog
```

As most watchdog driver names contain strings like `wd`, `wdt`, or `dog`, use the following command to check which driver is currently loaded:

```
root # lsmod | egrep "(wd|dog)"
```

To automatically load the watchdog driver, create the file `/etc/modules-load.d/watchdog.conf` containing a line with the driver name. For more information refer to the man page `modules-load.d`.

If you change the timeout for watchdog, the other two values (`msgwait` and `stonith-timeout`) must be changed as well. The watchdog timeout depends mostly on your storage latency. This value specifies that the majority of devices must successfully finish their read operation within this time frame. If not, the node will self-fence.

The following “formula” expresses roughly this relationship between these three values:

**Example 17.1:** *Cluster Timings with SBD as STONITH Device*

$$\text{Timeout (msgwait)} = (\text{Timeout (watchdog)} * 2)$$



```
stonith-timeout = Timeout (msgwait) + 20%
```

For example, if you set the timeout watchdog to 120, you need to set the `msgwait` to 240 and the `stonith-timeout` to 288. You can check the output with `cs_make_sbd_devices`:

```
root # cs_make_sbd_devices --dump
==Dumping header on disk /dev/sdb
Header version      : 2.1
UUID                : 619127f4-0e06-434c-84a0-ea82036e144c
Number of slots     : 255
Sector size         : 512
Timeout (watchdog)  : 20
Timeout (allocate)  : 2
Timeout (loop)      : 1
Timeout (msgwait)   : 40
==Header on disk /dev/sdb is dumped
```

If you set up a new cluster, the `sleha-init` command takes the above considerations into account.

For more details about timing variables related to SBD, see the technical information document *SBD Operation Guidelines for HAE Clusters*, available at <https://www.suse.com/support/kb/doc.php?id=7011346>.

### 17.1.3.3 Starting the SBD Daemon

The SBD daemon is a critical piece of the cluster stack. It needs to be running when the cluster stack is running, or even when part of it has crashed, so that the node can be fenced.

- 1 Run `sleha-init`. This script ensures that SBD is correctly configured and the configuration file `/etc/sysconfig/sbd` is added to the list of files that needs to be synchronized with Csync2.

If you want to configure SBD manually, perform the following step:

To make the OpenAIS init script start and stop SBD, edit the file `/etc/sysconfig/sbd` and search for the following line, replacing *SBD* with your SBD device:

```
SBD_DEVICE="/dev/SBD"
```

If you need to specify multiple devices in the first line, separate them by a semicolon (the order of the devices does not matter):

```
SBD_DEVICE="/dev/SBD1; /dev/SBD2; /dev/SBD3"
```

If the SBD device is not accessible, the daemon will fail to start and inhibit OpenAIS startup.

---

**NOTE: Starting Services at Boot Time**

If the SBD device becomes inaccessible from a node, this could cause the node to enter an infinite reboot cycle. This is technically correct behavior, but depending on your administrative policies, most likely a nuisance. In such cases, better do not automatically start up OpenAIS on boot.

---

- 2 Before proceeding, ensure that SBD has started on all nodes by executing `rcopenais restart`.

### 17.1.3.4 Testing SBD

- 1 The following command will dump the node slots and their current messages from the SBD device:

```
root # sbd -d /dev/SBD list
```

Now you should see all cluster nodes that have ever been started with SBD listed here, the message slot should show `clear`.

- 2 Try sending a test message to one of the nodes:

```
root # sbd -d /dev/SBD message alice test
```

- 3 The node will acknowledge the receipt of the message in the system log files:

```
Aug 29 14:10:00 alice sbd: [13412]: info: Received command test from bob
```

This confirms that SBD is indeed up and running on the node and that it is ready to receive messages.

### 17.1.3.5 Configuring the Fencing Resource

To complete the SBD setup, configure SBD as a STONITH/fencing mechanism in the CIB.

---

## TIP: STONITH Configuration for 2-Node Clusters

In two-node clusters (and other clusters where `no-quorum-policy` is set to `ignore`), mistimed fencing occurs quite frequently, because both nodes will try to fence each other in case of a split-brain situation. To avoid this double fencing, add the `pcmk_delay_max` parameter to the configuration of the STONITH resource. This gives servers with a working network card a better chance to survive.

---

**1** Log in to one of the nodes and start the interactive `crmsh` with `crm configure`.

**2** Enter the following:

```
crm(live)configure# property stonith-enabled="true"
crm(live)configure# property stonith-timeout="40s" ❶
crm(live)configure# primitive stonith_sbd stonith:external/sbd \
    pcmk_delay_max="30" ❷
```

The resource does not need to be cloned. As node slots are allocated automatically, no manual host list needs to be defined.

- ❶ Which value to set for `stonith-timeout` depends on the `msgwait` timeout. The `msgwait` timeout should be longer than the maximum allowed timeout for the underlying IO system. For example, this is 30 seconds for plain SCSI disks. Provided you set the `msgwait` timeout value to 30 seconds, setting `stonith-timeout` to 40 seconds is appropriate.
- ❷ The `pcmk_delay_max` parameter enables a random delay for STONITH actions on the fencing device. Its value specifies the maximum amount of time to wait before the start operation of the STONITH device is executed. As it takes time to detect the ring failure, become the DC and to start the STONITH resource, do not set this value too low (otherwise the prior DC will always start the fencing action first).

**3** Commit the changes:

```
crm(live)# configure commit
```

**4** Disable any other fencing devices you might have configured before, since the SBD mechanism is used for this function now.

After the resource has started, your cluster is successfully configured for shared-storage fencing and will use this method in case a node needs to be fenced.

### 17.1.3.6 For More Information

[http://www.linux-ha.org/wiki/SBD\\_Fencing](http://www.linux-ha.org/wiki/SBD_Fencing)

## 17.2 Ensuring Exclusive Storage Activation

This section introduces `sfex`, an additional low-level mechanism to lock access to shared storage exclusively to one node. Note that `sfex` does not replace STONITH. Since `sfex` requires shared storage, it is recommended that the `external/sbd` fencing mechanism described above is used on another partition of the storage.

By design, `sfex` cannot be used with workloads that require concurrency (such as OCFS2), but serves as a layer of protection for classic failover style workloads. This is similar to a SCSI-2 reservation in effect, but more general.

### 17.2.1 Overview

In a shared storage environment, a small partition of the storage is set aside for storing one or more locks.

Before acquiring protected resources, the node must first acquire the protecting lock. The ordering is enforced by Pacemaker, and the `sfex` component ensures that even if Pacemaker were subject to a split brain situation, the lock will never be granted more than once.

These locks must also be refreshed periodically, so that a node's death does not permanently block the lock and other nodes can proceed.

### 17.2.2 Setup

In the following, learn how to create a shared partition for use with `sfex` and how to configure a resource for the `sfex` lock in the CIB. A single `sfex` partition can hold any number of locks, it defaults to one, and needs 1 KB of storage space allocated per lock.

---

## IMPORTANT: Requirements

- The shared partition for sfex should be on the same logical unit as the data you want to protect.
  - The shared sfex partition must not use host-based RAID, nor DRBD.
  - Using a cLVM2 logical volume is possible.
- 

### ***Procedure 17.1: Creating an sfex Partition***

- 1 Create a shared partition for use with sfex. Note the name of this partition and use it as a substitute for `/dev/sfex` below.
- 2 Create the sfex meta data with the following command:

```
root # sfex_init -n 1 /dev/sfex
```

- 3 Verify that the meta data has been created correctly:

```
root # sfex_stat -i 1 /dev/sfex ; echo $?
```

This should return 2, since the lock is not currently held.

### ***Procedure 17.2: Configuring a Resource for the sfex Lock***

- 1 The sfex lock is represented via a resource in the CIB, configured as follows:

```
crm(live)configure# primitive sfex_1 ocf:heartbeat:sfex \  
# params device="/dev/sfex" index="1" collision_timeout="1" \  
lock_timeout="70" monitor_interval="10" \  
# op monitor interval="10s" timeout="30s" on_fail="fence"
```

- 2 To protect resources via an sfex lock, create mandatory ordering and placement constraints between the protectees and the sfex resource. If the resource to be protected has the id `filesystem1`:

```
crm(live)configure# order order-sfex-1 inf: sfex_1 filesystem1  
crm(live)configure# colocation colo-sfex-1 inf: filesystem1 sfex_1
```

- 3 If using group syntax, add the sfex resource as the first resource to the group:

```
crm(live)configure# group LAMP sfex_1 filesystem1 apache ipaddr
```

## 17.3 For More Information

See [http://www.linux-ha.org/wiki/SBD\\_Fencing](http://www.linux-ha.org/wiki/SBD_Fencing) and `man sbd`.

# Samba Clustering

A clustered Samba server provides a High Availability solution in your heterogeneous networks. This chapter explains some background information and how to set up a clustered Samba server.

## 18.1 Conceptual Overview

Trivial Database (TDB) has been used by Samba for many years. It allows multiple applications to write simultaneously. To make sure all write operations are successfully performed and do not collide with each other, TDB uses an internal locking mechanism.

Cluster Trivial Database (CTDB) is a small extension of the existing TDB. CTDB is described by the project as a “cluster implementation of the TDB database used by Samba and other projects to store temporary data”.

Each cluster node runs a local CTDB daemon. Samba communicates with its local CTDB daemon instead of writing directly to its TDB. The daemons exchange metadata over the network, but actual write and read operations are done on a local copy with fast storage. The concept of CTDB is displayed in Figure 18.1, “Structure of a CTDB Cluster” (page 298).

---

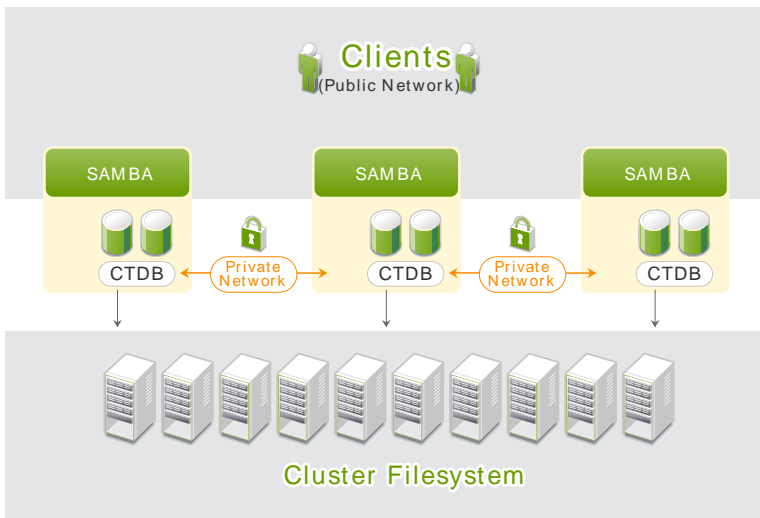
### NOTE: CTDB For Samba Only

The current implementation of the CTDB Resource Agent configures CTDB to only manage Samba. Everything else, including IP failover, should be configured with Pacemaker.

CTDB is only supported for completely homogeneous clusters. For example, all nodes in the cluster need to have the same architecture. You cannot mix i586 with x86\_64.

---

**Figure 18.1:** *Structure of a CTDB Cluster*



A clustered Samba server must share certain data:

- Mapping table that associates Unix user and group IDs to Windows users and groups.
- The user database must be synchronized between all nodes.
- Join information for a member server in a Windows domain must be available on all nodes.
- Metadata has to be available on all nodes, like active SMB sessions, share connections, and various locks.



The goal is that a clustered Samba server with N+1 nodes is faster than with only N nodes. One node is not slower than an unclustered Samba server.

## 18.2 Basic Configuration

---

### NOTE: Changed Configuration Files

The CTDB Resource Agent automatically changes `/etc/sysconfig/ctdb` and `/etc/samba/smb.conf`. Use `crm ra info CTDB` to list all parameters that can be specified for the CTDB resource.

---

To set up a clustered Samba server, proceed as follows:

#### **Procedure 18.1:** *Setting Up a Basic Clustered Samba Server*

##### **1** Prepare your cluster:

**1a** Configure your cluster (OpenAIS, Pacemaker, OCFS2) as described in this guide in Part II, “Configuration and Administration” (page 51).

**1b** Configure a shared file system, like OCFS2, and mount it, for example, on `/shared`.

**1c** If you want to turn on POSIX ACLs, enable it:

- For a new OCFS2 file system use:

```
root # mkfs.ocfs2 --fs-features=xattr ...
```

- For an existing OCFS2 file system use:

```
root # tuneefs.ocfs2 --fs-feature=xattrDEVICE
```

Make sure the `acl` option is specified in the file system resource. Use the `crm` shell as follows:

```
crm(live)configure# primary ocfs2-3 ocf:heartbeat:Filesystem
options="acl" ...
```

**1d** Make sure the services `ctdb`, `smb`, `nmb`, and `winbind` are disabled:

```
root # chkconfig ctdb off
chkconfig smb off
```

```
chkconfig nmb off
chkconfig winbind off
```

## 2 Create a directory for the CTDB lock on the shared file system:

```
root # mkdir -p /shared/samba/
```

## 3 In /etc/ctdb/nodes insert all nodes which contain all private IP addresses of each node in the cluster:

```
192.168.1.10
192.168.1.11
```

## 4 Copy the configuration file to all of your nodes by using csync2:

```
root # csync2 -xv
```

For more information, see Procedure 3.10, “Synchronizing the Configuration Files with Csync2” (page 43).

## 5 Add a CTDB resource to the cluster:

```
crm configure
crm(live)configure# primitive ctdb ocf:heartbeat:CTDB params \
    ctdb_manages_winbind="false" \
    ctdb_manages_samba="true" \
    ctdb_recovery_lock="/shared/samba/ctdb.lock" \
    op monitor interval="10" timeout="20" \
    op start interval="0" timeout="90" \
    op stop interval="0" timeout="100"
crm(live)configure# clone ctdb-clone ctdb \
    meta globally-unique="false" interleave="true"
crm(live)configure# colocation ctdb-with-fs inf: ctdb-clone fs-clone
crm(live)configure# order start-ctdb-after-fs inf: fs-clone ctdb-clone
crm(live)configure# commit
```

## 6 Add a clustered IP address:

```
crm(live)configure# primitive ip ocf:heartbeat:IPaddr2 params
    ip=192.168.2.222 \
    clusterip_hash="sourceip-sourceport" op monitor interval=60s
crm(live)configure# clone ip-clone ip meta globally-unique="true"
crm(live)configure# colocation ip-with-ctdb inf: ip-clone ctdb-clone
crm(live)configure# order start-ip-after-ctdb inf: ctdb-clone ip-clone
crm(live)configure# commit
```

## 7 Check the result:

```
root # crm status
Clone Set: dlm-clone
    Started: [ hex-14 hex-13 ]
```

```

Clone Set: o2cb-clone
  Started: [ hex-14 hex-13 ]
Clone Set: c-ocfs2-3
  Started: [ hex-14 hex-13 ]
Clone Set: ctddb-clone
  Started: [ hex-14 hex-13 ]
Clone Set: ip-clone (unique)
  ip:0      (ocf::heartbeat:IPAddr2):      Started hex-13
  ip:1      (ocf::heartbeat:IPAddr2):      Started hex-14

```

## 8 Test from a client machine. On a Linux client, add a user for Samba access:

```
root # smbpasswd -a USERNAME
```

## 9 Test if you can reach the new user's home directory:

```
root # smbclient -u USERNAME//192.168.2.222/USERNAME
```

# 18.3 Joining an Active Directory Domain

Active Directory (AD) is a directory service for Windows server systems.

The following instructions outline how to join a CTDB cluster to an Active Directory domain:

## 1 Consult your Windows Server documentation for instructions on how to set up an Active Directory domain. In this example, we use the following parameters:

AD and DNS server	win2k3.2k3test.example.com
AD domain	2k3test.example.com
Cluster AD member NetBIOS name	CTDB-SERVER

## 2 Procedure 18.2, “Configuring CTDB” (page 302)

## 3 Procedure 18.3, “Joining Active Directory” (page 302)

The next step is to configure the CTDB:

### **Procedure 18.2:** *Configuring CTDB*

- 1 Make sure you have configured your cluster as shown in Section 18.2, “Basic Configuration” (page 299).

- 2 Stop the CTDB resource on one node:

```
root # crm resource stop ctdb-clone
```

- 3 Open the `/etc/samba/smb.conf` configuration file, add your NetBIOS name, and close the file:

```
[global  
    netbios name = CTDB-SERVER
```

Other settings such as security, workgroup etc. are added by the YaST wizard.

- 4 Update on all nodes the file `/etc/samba.conf`:

```
root # csync2 -xv
```

- 5 Restart the CTDB resource:

```
root # crm resource start ctdb-clone
```

Finally, join your cluster to the Active Directory server:

### **Procedure 18.3:** *Joining Active Directory*

- 1 Make sure the following files are included in Csync2's configuration to become installed on all cluster hosts:

```
/etc/samba/smb.conf  
/etc/security/pam_winbind.conf  
/etc/krb5.conf  
/etc/nsswitch.conf  
/etc/security/pam_mount.conf.xml  
/etc/pam.d/common-session
```

You can also use YaST's *Configure Csync2* module for this task, see Section 3.5.4, “Transferring the Configuration to All Nodes” (page 42).

- 2 Create a CTDB resource as described in Procedure 18.1, “Setting Up a Basic Clustered Samba Server” (page 299).
- 3 Run YaST and open the *Windows Domain Membership* module from the *Network Services* entry.
- 4 Enter your domain or workgroup settings and finish with *Ok*.

## 18.4 Debugging and Testing Clustered Samba

To debug your clustered Samba server, the following tools which operate on different levels are available:

### `ctdb_diagnostics`

Run this tool to diagnose your clustered Samba server. Detailed debug messages should help you track down any problems you might have.

The `ctdb_diagnostics` command searches for the following files which must be available on all nodes:

```
/etc/krb5.conf
/etc/hosts
/etc/ctdb/nodes
/etc/sysconfig/ctdb
/etc/resolv.conf
/etc/nsswitch.conf
/etc/sysctl.conf
/etc/samba/smb.conf
/etc/fstab
/etc/multipath.conf
/etc/pam.d/system-auth
/etc/sysconfig/nfs
/etc/exports
/etc/vsftpd/vsftpd.conf
```

If the files `/etc/ctdb/public_addresses` and `/etc/ctdb/static-routes` exist, they will be checked as well.

### `ping_pong`

Check whether your file system is suitable for CTDB with `ping_pong`. It performs certain tests of your cluster file system like coherence and performance (see [http://wiki.samba.org/index.php/Ping\\_pong](http://wiki.samba.org/index.php/Ping_pong)) and gives some indication how your cluster may behave under high load.

### `send_arp` Tool and SendArp Resource Agent

The SendArp resource agent is located in `/usr/lib/heartbeat/send_arp` (or `/usr/lib64/heartbeat/send_arp`). The `send_arp` tool sends out a gratuitous ARP (Address Resolution Protocol) packet and can be used for updating other machines' ARP tables. It can help to identify communication problems after a failover process. If you cannot connect to a

node or ping it although it shows the clustered IP address for samba, use the `send_arp` command to test if the nodes only need an ARP table update.

For more information, refer to [http://wiki.wireshark.org/Gratuitous\\_ARP](http://wiki.wireshark.org/Gratuitous_ARP).

To test certain aspects of your cluster file system proceed as follows:

***Procedure 18.4: Test Coherence and Performance of Your Cluster File System***

- 1 Start the command `ping_pong` on one node and replace the placeholder *N* with the amount of nodes plus one. The file `ABSPATH/data.txt` is available in your shared storage and is therefore accessible on all nodes (`ABSPATH` indicates an absolute path):

```
ping_pong ABSPATH/data.txt N
```

Expect a very high locking rate as you are running only one node. If the program does not print a locking rate, replace your cluster file system.

- 2 Start a second copy of `ping_pong` on another node with the same parameters.

Expect to see a dramatic drop in the locking rate. If any of the following applies to your cluster file system, replace it:

- `ping_pong` does not print a locking rate per second,
- the locking rates in the two instances are not almost equal,
- the locking rate did not drop after you started the second instance.

- 3 Start a third copy of `ping_pong`. Add another node and note how the locking rates change.
- 4 Kill the `ping_pong` commands one after the other. You should observe an increase of the locking rate until you get back to the single node case. If you did not get the expected behavior, find more information in Chapter 14, *OCFS2* (page 245).

## 18.5 For More Information

- [http://linux-ha.org/wiki/CTDB\\_\(resource\\_agent\)](http://linux-ha.org/wiki/CTDB_(resource_agent))

- [http://wiki.samba.org/index.php/CTDB\\_Setup](http://wiki.samba.org/index.php/CTDB_Setup)
- <http://ctdb.samba.org>
- [http://wiki.samba.org/index.php/Samba\\_%26\\_Clustering](http://wiki.samba.org/index.php/Samba_%26_Clustering)





# Disaster Recovery with Rear

Relax-and-Recover (formerly “ReaR”, in this chapter abbreviated as Rear) is a disaster recovery framework for use by system administrators. It is a collection of Bash scripts that need to be adjusted to the specific production environment that is to be protected in case of disaster.

No disaster recovery solution will just work out-of-the-box. Therefore it is essential to take preparations *before* any disaster happens.

## 19.1 Conceptual Overview

The following sections describe the general disaster recovery concept and the basic steps you need to execute for successful recovery with Rear. They also provide some guidance on Rear requirements, Rear versions shipped with the current product, some limitations to be aware of, and scenarios and backup tools.

### 19.1.1 Creating a Disaster Recovery Plan

*Before* the worst scenario happens, take action: analyze your IT infrastructure for any substantial risks, evaluate your budget, and create a disaster recovery plan. If you do not already have a disaster recovery plan at hand, find some information on each step below:

- **Risk Analysis** Conduct a solid risk analysis of your infrastructure. List all the possible threats and evaluate how serious they are. Determine how likely these

threats are and prioritize them. It is recommended to use a simple categorization: probability and impact.

- **Budget Planning** The outcome of the analysis is an overview, which risks can be tolerated and which are critical for your business. Ask yourself how you can minimize risks and how much will it cost. Depending on how big your company is, spend two to fifteen percent of the overall IT budget on disaster recovery.
- **Disaster Recovery Plan Development** Make checklists, test procedures, establish and assign priorities, and inventory your IT infrastructure. Define how to deal with a problem when some services in your infrastructure fail.
- **Test** After defining an elaborate plan, test it. Test it at least once a year. Use the same testing hardware as your main IT infrastructure.

## 19.1.2 What Does Disaster Recovery Mean?

If a system in a production environment has been destroyed (for whatever reasons—be it broken hardware, a misconfiguration or software problems), you need to recreate the system. The recreation can be done either on the same hardware or on compatible replacement hardware. Recreating a system means more than just restoring files from a backup. It also includes preparing the system's storage (with regards to file systems, partitioning, or mount points), and reinstalling the boot loader.

## 19.1.3 How Does Disaster Recovery With Rear Work?

While the system is up and running, create a backup of the files and create a recovery system on a recovery medium. The recovery system contains a recovery installer.

In case the system has been destroyed, replace broken hardware (if needed), boot the recovery system from the recovery medium and launch the recovery installer. The recovery installer recreates the system: First, it prepares the storage (partitioning, file systems, mount points), then it restores the files from the backup. Finally, it reinstalls the boot loader.

## 19.1.4 Rear Requirements

To use Rear you need at least two identical systems: the machine that runs your production environment and an identical test machine. “Identical” in this context means that you can, for example, replace a network card with another one using the same Kernel driver.

---

### **WARNING: Identical Drivers Required**

If a hardware component does not use the same driver like the one in your production environment, it is not considered identical by Rear.

---

## 19.1.5 Rear Versions

SUSE Linux Enterprise High Availability Extension 11 SP4 ships two Rear versions in parallel:

- Rear version 1.10.0 (included in RPM package: `rear`). For the documentation of this Rear version, see the High Availability Guide for SUSE Linux Enterprise High Availability Extension 11 SP3. It is available from <http://www.suse.com/documentation/>.
- Rear version 1.16 (included in RPM package `rear116`). This version is documented in the current chapter.

---

### **IMPORTANT**

If you have a tested and fully functional disaster recovery procedure with Rear version 1.10.0 there is *no* need to upgrade. In that case, keep the Rear package and do not change your disaster recovery method!

However, if Rear version 1.10.0 does not fulfill your particular needs (see also Section 19.1.6 (page 310)), you can manually upgrade to Rear version 1.16. The packages `rear` and `rear116` intentionally conflict with each other to prevent that your installed version gets accidentally replaced with another version.

For each Rear version upgrade you must carefully re-validate whether your particular disaster recovery procedure still works.

---

## 19.1.6 Limitations with Btrfs

The following limitations apply if you use Btrfs.

Your System Includes Subvolumes, but No Snapshots Subvolumes

Rear version 1.16 is required. This version supports recreating “normal” Btrfs subvolume structure (no snapshot subvolumes).

Your System Includes Snapshot Subvolumes

---

### WARNING

Btrfs snapshot subvolumes *cannot* be backed up and restored as usual with file-based backup software.

---

While recent snapshots in subvolumes on Btrfs file systems need almost no disk space (because of Btrfs' copy-on-write functionality), those files would be backed up as complete files when using file-based backup software. They would end up twice in the backup with their original file size. Therefore, it would be impossible to restore the snapshots as they have been before on the original system.

## 19.1.7 Scenarios and Backup Tools

Rear can create a disaster recovery system (including a specific recovery installer) that can be booted from a local medium (like a hard disk, a flash disk, a DVD/CD-R) or via PXE. The backup data can be stored on a network file system, for example NFS, as described in Example 19.1 (page 312).

Rear does not replace a file backup, but complements it. By default, Rear supports the generic `tar` command, and several third-party backup tools (such as Tivoli Storage Manager, QNetix Galaxy, Symantec NetBackup, EMC NetWorker, or HP DataProtector). Refer to Example 19.2 (page 312) for an example configuration of using Rear with EMC NetWorker as backup tool.

## 19.1.8 Basic Steps

For a successful recovery with Rear in case of disaster, you need to execute the following basic steps:

Setting Up Rear and Your Backup Solution (page 311)

This includes tasks like editing the Rear configuration file, adjusting the Bash scripts, and configuring the backup solution that you want to use.

Creating the Recovery Installation System (page 313)

While the system to be protected is up and running use the `rear mkbbackup` command to create a file backup and to generate a recovery system that contains a system-specific Rear recovery installer.

Testing the Recovery Process (page 313)

Whenever you have created a disaster recovery medium with Rear, test the disaster recovery process thoroughly. It is essential to use a test machine that has *identical* hardware like the one that is part of your production environment. For details, refer to Section 19.1.4, “Rear Requirements” (page 309).

Recovering from Disaster (page 314)

After a disaster has occurred, replace any broken hardware (if necessary). Then boot the Rear recovery system and start the recovery installer with the `rear recover` command.

## 19.2 Setting Up Rear and Your Backup Solution

To set up Rear, you need to edit at least the Rear configuration file `/etc/rear/local.conf` and, if needed, the Bash scripts that are part of the Rear framework.

In particular, you need to define the following tasks that Rear should do:

- **How to back up files and how to create and store the disaster recovery system** This needs to be configured in `/etc/rear/local.conf`.
- **What to recreate exactly (partitioning, file systems, mount points etc.)** This can be defined in `/etc/rear/local.conf` (for example, what to exclude). To recreate non-standard systems, you may have to enhance the Bash scripts.
- **How the recovery process works** To change how Rear generates the recovery installer, or to adapt what the Rear recovery installer does, you have to edit the Bash scripts.

To configure Rear, add your options to the `/etc/rear/local.conf` configuration file. (The former configuration file `/etc/rear/sites.conf` has been removed from the package. However, if you have such a file from your last setup, Rear will still use it).

A default Rear configuration for reference is available at `/usr/share/rear/conf/default.conf`. Other example configurations (`*example.conf`) are available in the same directory. Find more information in the Rear man page.

After you have changed the Rear configuration file, run the following command and check its output:

```
rear dump
```

**Example 19.1:** *Using an NFS Server to Store the File Backup*

Rear can be used in different scenarios. The following example uses an NFS server as storage for the file backup.

- 1 Set up an NFS server with YaST as described in the SUSE Linux Enterprise Server 11 SP4 Administration Guide, chapter *Sharing File Systems with NFS*. It is available from <http://www.suse.com/documentation/>.
- 2 Define the configuration for your NFS server in the `/etc/exports` file. Make sure the directory on the NFS server (where you want the backup data to be available), has the right mount options. For example:

```
/srv/nfs *(fsid=0,crossmnt,rw,no_root_squash,sync,no_subtree_check)
```

If needed, replace `/srv/nfs` with the path to your backup data on the NFS server and adjust the mount options. You will probably need `no_root_squash` to access the backup data as the `rear mkbbackup` command runs as `root`.

- 3 Adjust the various BACKUP parameters in the configuration file `/etc/rear/local.conf` to make Rear store the file backup on the respective NFS server. Find an example in your installed system under `/usr/share/rear/conf/SLE11-ext3-example.conf`.

**Example 19.2:** *Using Third-Party Backup Tools Like EMC NetWorker*

Using third-party backup tools instead of `tar` requires appropriate settings in the Rear configuration file.

The following is an example configuration for EMC NetWorker. Add this configuration snippet to `/etc/rear/local.conf` and adjust it according to your setup:

```
BACKUP=NSR
OUTPUT=ISO
BACKUP_URL=nfs://host.example.com/path/to/your/rear/backup
OUTPUT_URL=nfs://host.example.com/path/to/your/rear/backup
NSRSERVER=backupserver.example.com
RETENTION_TIME="Month"
```

## 19.3 Creating the Recovery Installation System

After you have configured Rear as described in Section 19.2 (page 311), create the recovery installation system (including the Rear recovery installer) plus the file backup with the following command:

```
rear -d -D mkbackup
```

It executes the following steps:

1. Analyzing the target system and gathering information, in particular about the disk layout (partitioning, file systems, mount points) and about the boot loader.
2. Creating a bootable recovery system with the information gathered in the first step. The resulting Rear recovery installer is *specific* to the system that you want to protect from disaster. It can only be used to recreate this specific system.
3. Calling the configured backup tool to back up system and user files.

## 19.4 Testing the Recovery Process

After having created the recovery system, test the recovery process on a test machine with identical hardware. See also Section 19.1.4, “Rear Requirements” (page 309). Make sure the test machine is correctly set up and can serve as a replacement for your main machine.

---

**WARNING: Extensive Testing on Identical Hardware**

Thorough testing of the disaster recovery process on machines is required. Test the recovery procedure on a regular basis to ensure everything works as expected.

---

***Procedure 19.1: Performing a Disaster Recovery on a Test Machine***

- 1** Create a recovery medium by burning the recovery system that you have created in Section 19.3 (page 313) to a DVD or CD . Alternatively, you can use a network boot via PXE.
- 2** Boot the test machine from the recovery medium.
- 3** From the menu, select *Recover*.
- 4** Log in as `root` (no password needed).
- 5** Enter the following command to start the recovery installer:  

```
rear -d -D recover
```

For details about the steps that Rear takes during the process, see Recovery Process (page 314).
- 6** After the recovery process has finished, check if the system has been successfully recreated and can serve as a replacement for your original system in the production environment.

## 19.5 Recovering from Disaster

In case a disaster has occurred, exchange any broken hardware if necessary. Then proceed as described in Procedure 19.1 (page 314), using either the repaired machine (or a tested, identical machine that can serve as a replacement for your original system).

The `rear recover` command will execute the following steps:

***Recovery Process***

1. Restoring the disk layout (partitions, file systems, and mount points).



2. Restoring the system and user files from the backup.
3. Restoring the boot loader.

## 19.6 For More Information

- [http://en.opensuse.org/SDB:Disaster\\_Recovery](http://en.opensuse.org/SDB:Disaster_Recovery)
- rear man page
- /usr/share/doc/packages/rear/README



## **Part IV. Appendix**



# Troubleshooting

Strange problems may occur that are not easy to understand, especially when starting to experiment with High Availability. However, there are several utilities that allow you to take a closer look at the High Availability internal processes. This chapter recommends various solutions.

## 20.1 Installation and First Steps

Troubleshooting difficulties when installing the packages or bringing the cluster online.

Are the HA packages installed?

The packages needed for configuring and managing a cluster are included in the *High Availability* installation pattern, available with the High Availability Extension.

Check if High Availability Extension is installed as an add-on to SUSE Linux Enterprise Server 11 SP4 on each of the cluster nodes and if the *High Availability* pattern is installed on each of the machines as described in Section 3.3, “Installation as Add-on” (page 29).

Is the initial configuration the same for all cluster nodes?

To communicate with each other, all nodes belonging to the same cluster need to use the same `bindnetaddr`, `mcastaddr` and `mcastport` as described in Section 3.5, “Manual Cluster Setup (YaST)” (page 35).

Check if the communication channels and options configured in `/etc/corosync/corosync.conf` are the same for all cluster nodes.

In case you use encrypted communication, check if the `/etc/corosync/authkey` file is available on all cluster nodes.

All `corosync.conf` settings except for `nodeid` must be the same; `authkey` files on all nodes must be identical.

Does the Firewall allow communication via the `mcastport`?

If the `mcastport` used for communication between the cluster nodes is blocked by the firewall, the nodes cannot see each other. When configuring the initial setup with YaST or the bootstrap scripts as described in Section 3.5, “Manual Cluster Setup (YaST)” (page 35) and Section 3.4, “Automatic Cluster Setup (sleha-bootstrap)” (page 30), the firewall settings are usually automatically adjusted.

To make sure the `mcastport` is not blocked by the firewall, check the settings in `/etc/sysconfig/SuSEfirewall2` on each node. Alternatively, start the YaST firewall module on each cluster node. After clicking *Allowed Service > Advanced*, add the `mcastport` to the list of allowed *UDP Ports* and confirm your changes.

Is OpenAIS started on each cluster node?

Check the OpenAIS status on each cluster node with `/etc/init.d/openais status`. In case OpenAIS is not running, start it by executing `/etc/init.d/openais start`.

## 20.2 Logging

I enabled monitoring but there is no trace of monitoring operations in the log files?

The `lrmd` daemon does not log recurring monitor operations unless an error occurred. Logging all recurring operations would produce too much noise. Therefore recurring monitor operations are logged only once an hour.

I only get a `failed` message. Is it possible to get more information?

Add the `--verbose` parameter to your commands. If you do that multiple times, the debug output becomes quite verbose. See `/var/log/messages` for useful hints.

How can I get an overview of all my nodes and resources?

Use the `crm_mon` command. The following displays the resource operation history (option `-o`) and inactive resources (`-r`):

```
root # crm_mon -o -r
```

The display is refreshed when the status changes (to cancel this press `Ctrl + C`).

An example may look like:

### ***Example 20.1: Stopped Resources***

```
Last updated: Fri Aug 15 10:42:08 2014
Last change: Fri Aug 15 10:32:19 2014
Stack: corosync
Current DC: bob (175704619) - partition with quorum
Version: 1.1.12-ad083a8
2 Nodes configured
3 Resources configured

Online: [ alice bob ]

Full list of resources:

my_ipaddress      (ocf::heartbeat:Dummy): Started barett-2
my_filesystem     (ocf::heartbeat:Dummy): Stopped
my_webserver      (ocf::heartbeat:Dummy): Stopped

Operations:
* Node bob:
  my_ipaddress: migration-threshold=3
    + (14) start: rc=0 (ok)
    + (15) monitor: interval=10000ms rc=0 (ok)
* Node alice:
```

The *Pacemaker Explained* PDF, available at <http://www.clusterlabs.org/doc/>, covers three different recovery types in the *How are OCF Return Codes Interpreted?* section.

How can I create a report with an analysis of all my cluster nodes?

On the `crm` shell, use either `crm_report` or `hb_report` to create a report.

The tools are used to compile:

- Cluster-wide log files,
- Package states,
- DLM/OCFS2 states,
- System information,

- CIB history,
- Parsing of core dump reports, if a debuginfo package is installed.

Usually run `crm_report` with the following command:

```
root # crm_report -f 0:00 -n jupiter -n venus
```

The command extracts all information since 0am on the hosts `jupiter` and `venus` and creates a `*.tar.bz2` archive named `crm_report-DATE.tar.bz2` in the current directory, for example, `crm_report-Wed-03-Mar-2012`. If you are only interested in a specific time frame, add the end time with the `-t` option.

---

### **WARNING: Remove Sensitive Information**

The `crm_report` tool tries to remove any sensitive information from the CIB and the `peinput` files, however, it cannot do everything. If you have more sensitive information, supply additional patterns. The log files and the `crm_mon`, `ccm_tool`, and `crm_verify` output are *not* sanitized.

Before sharing your data in any way, check the archive and remove all information you do not want to expose.

---

Customize the command execution with further options. For example, if you have an OpenAIS cluster, you certainly want to add the option `-A`. In case you have another user who has permissions to the cluster, use the `-u` option and specify this user (in addition to `root` and `hacluster`). In case you have a non-standard SSH port, use the `-X` option to add the port (for example, with the port 3479, use `-X "-p 3479"`). Further options can be found in the man page of `crm_report`.

After `crm_report` has analyzed all the relevant log files and created the directory (or archive), check the log files for an uppercase `ERROR` string. The most important files in the top level directory of the report are:

`analysis.txt`

Compares files that should be identical on all nodes.

`crm_mon.txt`

Contains the output of the `crm_mon` command.



`corosync.txt`

Contains a copy of the Corosync configuration file.

`description.txt`

Contains all cluster package versions on your nodes. There is also the `sysinfo.txt` file which is node specific. It is linked to the top directory.

Node-specific files are stored in a subdirectory named by the node's name.

## 20.3 Resources

How can I clean up my resources?

Use the following commands:

```
root # crm resource list
crm resource cleanup rsid [node]
```

If you leave out the node, the resource is cleaned on all nodes. More information can be found in Section 7.4.2, “Cleaning Up Resources” (page 191).

How can I list my currently known resources?

Use the command `crm resource list` to display your current resources.

I configured a resource, but it always fails. Why?

To check an OCF script use `ocf-tester`, for instance:

```
ocf-tester -n ipl -o ip=YOUR_IP_ADDRESS \
/usr/lib/ocf/resource.d/heartbeat/IPaddr
```

Use `-o` multiple times for more parameters. The list of required and optional parameters can be obtained by running `crm ra info AGENT`, for example:

```
root # crm ra info ocf:heartbeat:IPaddr
```

Before running `ocf-tester`, make sure the resource is not managed by the cluster.

Why do resources not fail over and why are there no errors?

If your cluster is a two node cluster, terminating one node will leave the remaining node without quorum. Unless you set the `no-quorum-policy` property to `ignore`, nothing happens. For two-node clusters you need:

```
property no-quorum-policy="ignore"
```

Another possibility is that the terminated node is considered unclean. Then it is necessary to fence it. If the STONITH resource is not operational or does not

exist, the remaining node will be waiting for the fencing to happen. The fencing timeouts are typically high, so it may take quite a while to see any obvious sign of problems (if ever).

Yet another possible explanation is that a resource is simply not allowed to run on this node. That may be because of a failure which happened in the past and which was not “cleaned”. Or it may be because of an earlier administrative action, that is a location constraint with a negative score. Such a location constraint is for instance inserted by the `crm resource migrate` command.

Why can I never tell where my resource will run?

If there are no location constraints for a resource, its placement is subject to an (almost) random node choice. You are well advised to always express a preferred node for resources. That does not mean that you need to specify location preferences for *all* resources. One preference suffices for a set of related (colocated) resources. A node preference looks like this:

```
location rsc-prefers-alice rsc 100: alice
```

## 20.4 STONITH and Fencing

Why does my STONITH resource not start?

Start (or enable) operation includes checking the status of the device. If the device is not ready, the STONITH resource will fail to start.

At the same time the STONITH plugin will be asked to produce a host list. If this list is empty, there is no point in running a STONITH resource which cannot shoot anything. The name of the host on which STONITH is running is filtered from the list, since the node cannot shoot itself.

If you want to use single-host management devices such as lights-out devices, make sure that the STONITH resource is *not* allowed to run on the node which it is supposed to fence. Use an infinitely negative location node preference (constraint). The cluster will move the STONITH resource to another place where it can start, but not before informing you.

Why does fencing not happen, although I have the STONITH resource?

Each STONITH resource must provide a host list. This list may be inserted by hand in the STONITH resource configuration or retrieved from the device itself, for instance from outlet names. That depends on the nature of the STONITH

plugin. `stonithd` uses the list to find out which STONITH resource can fence the target node. Only if the node appears in the list can the STONITH resource shoot (fence) the node.

If `stonithd` does not find the node in any of the host lists provided by running STONITH resources, it will ask `stonithd` instances on other nodes. If the target node does not show up in the host lists of other `stonithd` instances, the fencing request ends in a timeout at the originating node.

Why does my STONITH resource fail occasionally?

Power management devices may give up if there is too much broadcast traffic. Space out the monitor operations. Given that fencing is necessary only once in a while (and hopefully never), checking the device status once a few hours is more than enough.

Also, some of these devices may refuse to talk to more than one party at the same time. This may be a problem if you keep a terminal or browser session open while the cluster tries to test the status.

## 20.5 Miscellaneous

How can I run commands on all cluster nodes?

Use the command `pssh` for this task. If necessary, install `pssh`. Create a file (for example `hosts.txt`) where you collect all your IP addresses or host names you want to visit. Make sure you can log in with `ssh` to each host listed in your `hosts.txt` file. If everything is correctly prepared, execute `pssh` and use the `hosts.txt` file (option `-h`) and the interactive mode (option `-i`) as shown in this example:

```
pssh -i -h hosts.txt "ls -l /corosync/*.conf"
[1] 08:28:32 [SUCCESS] root@venus.example.com
-rw-r--r-- 1 root root 1480 Nov 14 13:37 /etc/corosync/corosync.conf
[2] 08:28:32 [SUCCESS] root@192.168.2.102
-rw-r--r-- 1 root root 1480 Nov 14 13:37 /etc/corosync/corosync.conf
```

What is the state of my cluster?

To check the current state of your cluster, use one of the programs `crm_mon` or `crm status`. This displays the current DC and all the nodes and resources known by the current node.

Why can several nodes of my cluster not see each other?

There could be several reasons:

- Look first in the configuration file `/etc/corosync/corosync.conf`. Check if the multicast or unicast address is the same for every node in the cluster (look in the `interface` section with the key `mcastaddr`).
- Check your firewall settings.
- Check if your switch supports multicast or unicast addresses.
- Check if the connection between your nodes is broken. Most often, this is the result of a badly configured firewall. This also may be the reason for a *split brain* condition, where the cluster is partitioned.

Why can an OCFS2 device not be mounted?

Check `/var/log/messages` for the following line:

```
Jan 12 09:58:55 alice lrmd: [3487]: info: RA output: [...]
      ERROR: Could not load ocfs2_stackglue
Jan 12 16:04:22 alice modprobe: FATAL: Module ocfs2_stackglue not found.
```

In this case the Kernel module `ocfs2_stackglue.ko` is missing. Install the package `ocfs2-kmp-default`, `ocfs2-kmp-pae` or `ocfs2-kmp-xen`, depending on the installed Kernel.

ERROR: Tag Not Supported by the RNG Schema  
See Upgrading the CIB Syntax Version (page 347).

## 20.6 For More Information

For additional information about high availability on Linux, including configuring cluster resources and managing and customizing a High Availability cluster, see <http://clusterlabs.org/wiki/Documentation>.



# Naming Conventions

This guide uses the following naming conventions for cluster nodes and names, cluster resources, and constraints.

## Cluster Nodes

Cluster nodes use first names:

alice, bob, charly, doro, and eris

## Cluster Site Names

Clusters sites are named after cities:

amsterdam, berlin, canberra, dublin, fukuoka, gizeh, hanoi, and istanbul

## Cluster Resources

Primitives	No prefix
Groups	Prefix g-
Clones	Prefix cl-
Multi-state resources	Prefix ms-

## Constraints

Ordering constraints	Prefix o-
----------------------	-----------

Location constraints	Prefix <code>loc-</code>
Colocation constraints	Prefix <code>col-</code>

# B

## Example of Setting Up a Simple Testing Resource

This chapter provides a basic example for the configuration of a simple resource: an IP address. It demonstrates both approaches to do so, using either the Pacemaker GUI or the `crm` command line tool.

For the following example, we assume that you have set up your cluster as described in Chapter 3, *Installation and Basic Setup* (page 25) and that your cluster consists of at least two nodes. For an introduction and overview of how to configure cluster resources with the Pacemaker GUI and the `crm` shell, refer to the following chapters:

- *Configuring and Managing Cluster Resources (GUI)* (page 137)
- *Configuring and Managing Cluster Resources (Command Line)* (page 169)

## B.1 Configuring a Resource with the GUI

Creating a sample cluster resource and migrating it to another server can help you test to ensure your cluster is functioning properly. A simple resource to configure and migrate is an IP address.

### **Procedure B.1:** *Creating an IP Address Cluster Resource*

- 1 Start the Pacemaker GUI and log in to the cluster as described in Section 6.1.1, “Logging in to a Cluster” (page 138).

- 2** In the left pane, switch to the *Resources* view and in the right pane, select the group to modify and click *Edit*. The next window shows the basic group parameters and the meta attributes and primitives already defined for that resource.
- 3** Click the *Primitives* tab and click *Add*.
- 4** In the next dialog, set the following parameters to add an IP address as sub-resource of the group:
  - 4a** Enter a unique ID. For example, `myIP`.
  - 4b** From the *Class* list, select *ocf* as resource agent class.
  - 4c** As *Provider* of your OCF resource agent, select *heartbeat*.
  - 4d** From the *Type* list, select *IPaddr* as resource agent.
  - 4e** Click *Forward*.
  - 4f** In the *Instance Attribute* tab, select the *IP* entry and click *Edit* (or double-click the *IP* entry).
  - 4g** As *Value*, enter the desired IP address, for example, `10.10.0.1` and click *OK*.
  - 4h** Add a new instance attribute and specify `nic` as *Name* and `eth0` as *Value*, then click *OK*.

The name and value are dependent on your hardware configuration and what you chose for the media configuration during the installation of the High Availability Extension software.

- 5** Once all parameters are set according to your wishes, click *OK* to finish the configuration of that resource. The configuration dialog is closed and the main window shows the modified resource.

To start the resource with the Pacemaker GUI, select *Management* in the left pane. In the right pane, right-click the resource and select *Start* (or start it from the toolbar).

To migrate the IP address resource to another node (`saturn`) proceed as follows:



### ***Procedure B.2: Migrating Resources to Another Node***

- 1 Switch to the *Management* view in the left pane, then right-click the IP address resource in the right pane and select *Migrate Resource*.
- 2 In the new window, select `saturn` from the *To Node* drop-down list to move the selected resource to the node `saturn`.
- 3 If you want to migrate the resource only temporarily, activate *Duration* and enter the time frame for which the resource should migrate to the new node.
- 4 Click *OK* to confirm the migration.

## **B.2 Configuring a Resource Manually**

Resources are any type of service that a computer provides. Resources are known to High Availability when they may be controlled by RAs (Resource Agents), which are LSB scripts or OCF scripts. All resources can be configured with the `crm` command or as XML in the CIB (Cluster Information Base) in the `resources` section.

To add an IP address `10.10.0.1` as a resource to the current configuration, use the `crm` command:

### ***Procedure B.3: Creating an IP Address Cluster Resource***

- 1 Open a shell and become `root`.
- 2 Enter `crm configure` to open the internal shell.
- 3 Create an IP address resource:

```
crm(live)configure# resourceprimitive myIP ocf:heartbeat:IPaddr params  
ip=10.10.0.1
```

---

#### **NOTE**

When configuring a resource with High Availability, the same resource should not be initialized by `init`. High availability is responsible for all service start or stop actions.

---

If the configuration was successful, a new resource appears in `crm_mon` that is started on a random node of your cluster.

To migrate a resource to another node, do the following:

***Procedure B.4: Migrating Resources to Another Node***

- 1** Start a shell and become the user `root`.
- 2** Migrate your resource `myip` to node `saturn`:

```
crm resource migrate myIP saturn
```

# C

## Example Configuration for OCFS2 and cLVM

The following is an example configuration that can help you setting up your resources for use of either OCFS2, cLVM, or both. The configuration below does not represent a complete cluster configuration but is only an extract, including all resources needed for OCFS2 and cLVM, and ignoring any other resources that you might need. Attributes and attribute values may need adjustment to your specific setup.

### ***Example C.1: Cluster Configuration for OCFS2 and cLVM***

```
primitive clvm ocf:lvm2:clvmd \  
    params daemon_timeout="30" \  
primitive dlm ocf:pacemaker:controld \  
    op monitor interval="60" timeout="60" \  
primitive o2cb ocf:ocfs2:o2cb \  
    op monitor interval="60" timeout="60" \  
primitive ocfs2-1 ocf:heartbeat:Filesystem \  
    params device="/dev/sdb1" directory="/mnt/shared" fstype="ocfs2" \  
    options="acl" \  
    op monitor interval="20" timeout="40" \  
primitive sbd_stonith stonith:external/sbd \  
    params pcmk_delay_max="15" \  
    op monitor interval="15" timeout="15" \  
primitive vg1 ocf:heartbeat:LVM \  
    params volgrpname="cluster-vg" \  
    op monitor interval="60" timeout="60" \  
group base-group dlm o2cb clvm vg1 ocfs2-1 \  
clone base-clone base-group \  
    meta interleave="true"
```

The configuration with a base group (including several primitives) and a base clone simplifies the overall setup: The base group has internal colocation and ordering and can always remain the same, apart from two resources:

- `vg1`—the resource for the volume group. Only configure this resource if your setup includes cVLM. Otherwise omit it from the cluster configuration and from the base group.
- `ocfs2-1`—the resource for mounting the OCFS2 file system. Only configure this resource if your setup includes OCFS2. Otherwise omit it from the cluster configuration and from the base group.

All of the other resources mentioned in Example C.1, “Cluster Configuration for OCFS2 and cLVM” (page 333) can be configured and be running in the cluster regardless of your setup.

# Cluster Management Tools

High Availability Extension ships with a comprehensive set of tools to assist you in managing your cluster from the command line. This chapter introduces the tools needed for managing the cluster configuration in the CIB and the cluster resources. Other command line tools for managing resource agents or tools used for debugging (and troubleshooting) your setup are covered in Chapter 20, *Troubleshooting* (page 319).

---

**NOTE: Use crmsh**

This tool is for experts only. Usually the crm shell (crmsh) is the recommended way of managing your cluster.

---

The following list presents several tasks related to cluster management and briefly introduces the tools to use to accomplish these tasks:

## Monitoring the Cluster's Status

The `crm_mon` command allows you to monitor your cluster's status and configuration. Its output includes the number of nodes, uname, uuid, status, the resources configured in your cluster, and the current status of each. The output of `crm_mon` can be displayed at the console or printed into an HTML file. When provided with a cluster configuration file without the status section, `crm_mon` creates an overview of nodes and resources as specified in the file. See the `crm_mon` man page for a detailed introduction to this tool's usage and command syntax.

## Managing the CIB

The `cibadmin` command is the low-level administrative command for manipulating the CIB. It can be used to dump all or part of the CIB, update all or part of it, modify all or part of it, delete the entire CIB, or perform miscellaneous CIB administrative operations. See the `cibadmin` man page for a detailed introduction to this tool's usage and command syntax.

## Managing Configuration Changes

The `crm_diff` command assists you in creating and applying XML patches. This can be useful for visualizing the changes between two versions of the cluster configuration or saving changes so they can be applied at a later time using `cibadmin`. See the `crm_diff` man page for a detailed introduction to this tool's usage and command syntax.

## Manipulating CIB Attributes

The `crm_attribute` command lets you query and manipulate node attributes and cluster configuration options that are used in the CIB. See the `crm_attribute` man page for a detailed introduction to this tool's usage and command syntax.

## Validating the Cluster Configuration

The `crm_verify` command checks the configuration database (CIB) for consistency and other problems. It can check a file containing the configuration or connect to a running cluster. It reports two classes of problems. Errors must be fixed before the High Availability Extension can work properly while warning resolution is up to the administrator. `crm_verify` assists in creating new or modified configurations. You can take a local copy of a CIB in the running cluster, edit it, validate it using `crm_verify`, then put the new configuration into effect using `cibadmin`. See the `crm_verify` man page for a detailed introduction to this tool's usage and command syntax.

## Managing Resource Configurations

The `crm_resource` command performs various resource-related actions on the cluster. It lets you modify the definition of configured resources, start and stop resources, or delete and migrate resources between nodes. See the `crm_resource` man page for a detailed introduction to this tool's usage and command syntax.

## Managing Resource Fail Counts

The `crm_failcount` command queries the number of failures per resource on a given node. This tool can also be used to reset the failcount, allowing

the resource to again run on nodes where it had failed too often. See the `crm_failcount` man page for a detailed introduction to this tool's usage and command syntax.

### Managing a Node's Standby Status

The `crm_standby` command can manipulate a node's standby attribute. Any node in standby mode is no longer eligible to host resources and any resources that are there must be moved. Standby mode can be useful for performing maintenance tasks, such as Kernel updates. Remove the standby attribute from the node for it to become a fully active member of the cluster again. See the `crm_standby` man page for a detailed introduction to this tool's usage and command syntax.





# Upgrading Your Cluster and Updating Software Packages



This chapter covers two different scenarios: upgrading a cluster to another version of SUSE® Linux Enterprise High Availability Extension (either a major release or a service pack) as opposed to updating individual packages on cluster nodes.

## E.1 Terminology

In the following, find definitions of the most important terms used in this chapter:

Major Release , General Availability (GA) Version

The Major Release of SUSE Linux Enterprise (or any software product) is a new version which brings new features and tools, decommissions previously deprecated components and comes with backward incompatible changes.

Service Pack (SP)

Combines several patches into a form which is easy to install or deploy. Service packs are numbered and usually contain security fixes, updates, upgrades, or enhancements of programs.

Update

Installation of a newer *minor* version of a package.

Upgrade

Installation of a newer *major* version of a package or distribution, which brings *new features*.

## E.2 Upgrading your Cluster to the Latest Product Version

Which upgrade path is supported and how to perform the upgrade depends on the current product version your cluster is running on and on the target version you want to migrate to. For general information on this, see the *SUSE Linux Enterprise Server 11 Deployment Guide*, chapter *Updating SUSE Linux Enterprise*. It is available at <http://www.suse.com/documentation/>.

---

### **IMPORTANT: Required Preparations Before Upgrading**

- Ensure that your system backup is up to date and restorable.
- Test the upgrade procedure on a staging instance of your cluster setup first, before performing it in a production environment.

This gives you an estimation of the time frame required for the maintenance window. It also helps to detect and solve any unexpected problems that might arise.

---

### E.2.1 Upgrading from SLES 10 to SLE HA 11

---

#### **IMPORTANT: Taking Cluster Offline is Required**

For migrating from SUSE Linux Enterprise Server 10 to SUSE Linux Enterprise Server 11 (any service pack), all cluster nodes must be *offline* and the cluster must be migrated as a whole—mixed clusters running on SUSE Linux Enterprise Server 10/SUSE Linux Enterprise Server 11 are not supported.

---

For convenience, SUSE Linux Enterprise High Availability Extension includes a `hb2openais.sh` script with which to convert your data while moving from the Heartbeat to the OpenAIS cluster stack. The script parses the configuration stored

in `/etc/ha.d/ha.cf` and generates a new configuration file for the OpenAIS cluster stack. Furthermore, it adjusts the CIB to match the OpenAIS conventions, converts the OCFS2 file system and replaces EVMS with cLVM. Any EVMS2 containers are converted to cLVM2 volumes. For volume groups referenced in existing resources in the CIB, new LVM resources are created.

To successfully migrate your cluster from SUSE Linux Enterprise Server 10 SP4 to SUSE Linux Enterprise Server 11, you need to execute the following steps:

1. Preparing your SUSE Linux Enterprise Server 10 SP4 Cluster (page 341)
2. Upgrading to SUSE Linux Enterprise 11 (page 342)
3. Testing the Conversion (page 343)
4. Converting the Data (page 343)

After the conversion has been successfully completed, you can bring the upgraded cluster online again.

---

**NOTE: Reverting after Upgrade**

After the upgrade process to SUSE Linux Enterprise Server 11, reverting back to SUSE Linux Enterprise Server 10 is *not* supported.

---

## E.2.1.1 Preparation and Backup

Before upgrading your cluster to the next product version and converting the data accordingly, you need to prepare your current cluster.

### ***Procedure E.1:*** *Preparing your SUSE Linux Enterprise Server 10 SP4 Cluster*

- 1 Log in to the cluster.
- 2 Review the Heartbeat configuration file `/etc/ha.d/ha.cf` and check that all communication media support multicasting.
- 3 Make sure the following files are equal on all nodes: `/etc/ha.d/ha.cf` and `/var/lib/heartbeat/crm/cib.xml`.

- 4 Take all nodes offline by executing `rcheartbeat stop` on each node.
- 5 In addition to the general system backup recommended before updating to the latest version, back up the following files, as you need them for running the conversion script after the upgrade to SUSE Linux Enterprise Server 11:
  - `/var/lib/heartbeat/crm/cib.xml`
  - `/var/lib/heartbeat/hostcache`
  - `/etc/ha.d/ha.cf`
  - `/etc/logd.cf`
- 6 If you have EVMS2 resources, convert non-LVM EVMS2 volumes to compatibility volumes on SUSE Linux Enterprise Server 10. During the conversion process (see Section E.2.1.3, “Data Conversion” (page 342)), these are then turned into LVM2 volume groups. After conversion, make sure to mark each volume group as a member of the High Availability cluster with `vgchange -c y`.

## E.2.1.2 Upgrade/Installation

After preparing the cluster and backing up the files, do a fresh installation of SUSE Linux Enterprise 11 on your cluster nodes.

### ***Procedure E.2:*** *Upgrading to SUSE Linux Enterprise 11*

- 1 Freshly install SUSE Linux Enterprise Server 11 on all cluster nodes.
- 2 On all cluster nodes, install SUSE Linux Enterprise High Availability Extension 11 as add-on on top of SUSE Linux Enterprise Server. For detailed information, see Section 3.3, “Installation as Add-on” (page 29).

## E.2.1.3 Data Conversion

After having installed SUSE Linux Enterprise Server 11 and the High Availability Extension, you can start with the data conversion. The conversion script shipped with the High Availability Extension has been set up with care, but it cannot handle all set-ups in fully automatic mode. It alerts you of the changes it makes, but needs

interaction and decisions from your side. You need to know your cluster in detail—it is up to you to verify that the changes are meaningful. The conversion script is located in `/usr/lib/heartbeat` (or in `/usr/lib64/heartbeat`, if you are using a 64-bit system).

---

**NOTE: Executing Test Runs**

To make yourself familiar with the conversion process, we highly recommend that you test the conversion first (without making any changes). You can use the same test directory to do repeated test runs, but you only need to copy the files once.

---

**Procedure E.3:** *Testing the Conversion*

- 1 On one of the nodes, create a test directory and copy the backup files to the test directory:

```
$ mkdir /tmp/hb2openais-testdir
$ cp /etc/ha.d/ha.cf /tmp/hb2openais-testdir
$ cp /var/lib/heartbeat/hostcache /tmp/hb2openais-testdir
$ cp /etc/logd.cf /tmp/hb2openais-testdir
$ sudo cp /var/lib/heartbeat/crm/cib.xml /tmp/hb2openais-testdir
```

- 2 Start the test run with

```
$ /usr/lib/heartbeat/hb2openais.sh -T /tmp/hb2openais-testdir -U
```

or with the following command, if you are using a 64-bit system:

```
$ /usr/lib64/heartbeat/hb2openais.sh -T /tmp/hb2openais-testdir -U
```

- 3 Read and verify the resulting `openais.conf` and `cib-out.xml` files:

```
$ cd /tmp/hb2openais-testdir
$ less openais.conf
$ crm_verify -V -x cib-out.xml
```

For detailed information about the conversion stages, refer to `/usr/share/doc/packages/pacemaker/README.hb2openais` in your installed High Availability Extension.

**Procedure E.4:** *Converting the Data*

After doing a test run and checking the output, you can now start with the data conversion. You only need to run the conversion on *one* node. The main cluster

configuration (the CIB) is automatically replicated to the other nodes. All other files that need to be replicated are automatically copied by the conversion script.

- 1 Make sure that `sshd` is running on all nodes with access allowed for `root` in order for the conversion script to successfully copy the files to the other cluster nodes.
- 2 Make sure that all `ocfs2` file systems are unmounted.
- 3 The High Availability Extension ships with a default OpenAIS configuration file. If you want to prevent the default configuration from being overwritten during the following steps, make a copy of the `/etc/ais/openais.conf` configuration file.
- 4 Start the conversion script as `root`. If using `sudo`, specify the privileged user using the `-u` option:

```
$ /usr/lib/heartbeat/hb2openais.sh -u root
```

Based on the configuration stored in `/etc/ha.d/ha.cf`, the script will generate a new configuration file for the OpenAIS cluster stack, `/etc/ais/openais.conf`. It will also analyze the CIB configuration and let you know if your cluster configuration requires changes, due to the change from Heartbeat to OpenAIS. All file processing is done on the node where conversion runs and replicated to the other nodes.

- 5 Follow the instructions on the screen.

After the conversion has been finished successfully, start the new cluster stack as described in Section 3.5.7, “Bringing the Cluster Online” (page 48).

After the upgrade process, reverting back to SUSE Linux Enterprise Server 10 is not supported.

## E.2.1.4 For More Information

For more details about the conversion script and the stages of the conversion, refer to `/usr/share/doc/packages/pacemaker/README.hb2openais` in your installed High Availability Extension.

## E.2.2 Upgrading from SLE HA 11 to SLE HA 11 SP1

---

### NOTE: Rolling Upgrade Between Service Pack Versions

To successfully migrate an existing cluster from one service pack version to the next one, you can do a “rolling upgrade”, meaning upgrading one node after the other.

---

As the main cluster configuration file has changed from `/etc/ais/openais.conf` to `/etc/corosync/corosync.conf` with SUSE Linux Enterprise High Availability Extension 11 SP1, a script takes care of the necessary conversions. They are executed automatically when the `openais` package is updated.

### *Procedure E.5: Performing a Rolling Upgrade*

---

#### WARNING: Active Cluster Stack During Update

To update any software packages on a node that is part of a running cluster, *stop* the cluster stack *on that node* before starting the software update. In some cases (see Conditions for Stopping the Cluster Stack (page 348), you can alternatively put the cluster into maintenance mode (which is available since SUSE Linux Enterprise High Availability Extension 11 SP4).

If the cluster resource manager on a node is active during the software update, this can lead to unpredictable results like fencing of active nodes.

---

- 1 Log in as `root` on the node that you want to upgrade and stop OpenAIS:  

```
rcopenais stop
```
- 2 Check that your system backup is up-to-date and restorable.
- 3 Perform an upgrade from SUSE Linux Enterprise Server 11 to SUSE Linux Enterprise Server 11 SP1 and from SUSE Linux Enterprise High Availability Extension 11 to SUSE Linux Enterprise High Availability Extension 11 SP1. For information on how to upgrade your product, refer to the SUSE Linux Enterprise Server 11 SP1 Deployment Guide, chapter *Updating SUSE Linux Enterprise*.

- 4 Restart OpenAIS/Corosync on the upgraded node to make the node rejoin the cluster:

```
rcopenais start
```

- 5 Take the next node offline and repeat the procedure for that node.

## E.2.3 Upgrading from SLE HA 11 SP1 to SLE HA 11 SP2

Migrating an existing cluster from SUSE Linux Enterprise High Availability Extension 11 SP1 to 11 SP2 is done via a `rolling upgrade`, similar to the upgrade procedure from version 11 to 11 SP1.

Proceed as described in Procedure E.5, “Performing a Rolling Upgrade” (page 345) with the following two deviations:

- In Step 3 (page 345), upgrade from SUSE Linux Enterprise Server 11 SP1 to SUSE Linux Enterprise Server 11 SP2 and from SUSE Linux Enterprise High Availability Extension 11 SP1 to SUSE Linux Enterprise High Availability Extension 11 SP2.

As Xen Hypervisor is discontinued for 32-bit architectures, you might need to solve dependencies for the package `drbd-xen` manually. Note that cross-platform clusters are not supported.

- Because of the Kernel update shipped with SP2, reboot the node between Step 3 (page 345) and Step 4 (page 346).

---

### IMPORTANT: Time Limit for Rolling Upgrade

The new features shipped with SUSE Linux Enterprise High Availability Extension 11 SP2 will only be available after *all* cluster nodes have been upgraded to the latest product version. Mixed SP1/SP2 clusters are only supported for a short time frame during the rolling upgrade. Complete the rolling upgrade within one week.

---



## E.2.4 Upgrading from SLE HA 11 SP2 to SLE HA 11 SP3

Migrating an existing cluster from SUSE Linux Enterprise High Availability Extension 11 SP2 to 11 SP3 is done via a `rolling upgrade`, similar to the upgrade procedure from version 11 to 11 SP1.

Proceed as described in Procedure E.5, “Performing a Rolling Upgrade” (page 345) with the following deviation: In Step 3 (page 345), upgrade from SUSE Linux Enterprise Server 11 SP2 to SUSE Linux Enterprise Server 11 SP3 and from SUSE Linux Enterprise High Availability Extension 11 SP2 to SUSE Linux Enterprise High Availability Extension 11 SP3.

---

### **IMPORTANT: Time Limit for Rolling Upgrade**

The new features shipped with SUSE Linux Enterprise High Availability Extension 11 SP3 will only be available after *all* cluster nodes have been upgraded to the latest product version. Mixed SP2/SP3 clusters are only supported for a short time frame during the rolling upgrade. Complete the rolling upgrade within one week.

---

## E.2.5 Upgrading from SLE HA 11 SP3 to SLE HA 11 SP4

Migrating an existing cluster from SUSE Linux Enterprise High Availability Extension 11 SP3 to 11 SP4 is done via a `rolling upgrade`, similar to the upgrade procedure from version 11 to 11 SP1.

Proceed as described in Procedure E.5, “Performing a Rolling Upgrade” (page 345) with the following deviation: In Step 3 (page 345), upgrade from SUSE Linux Enterprise Server 11 SP3 to SUSE Linux Enterprise Server 11 SP4 and from SUSE Linux Enterprise High Availability Extension 11 SP3 to SUSE Linux Enterprise High Availability Extension 11 SP4.

---

### **NOTE: Upgrading the CIB Syntax Version**

Tags (for grouping resources) and some ACL features only work with the CIB syntax version `pacemaker-2.0` or higher. (To check your version,

run `cibadmin -Q |grep validate-with`). If you have upgraded from SUSE Linux Enterprise High Availability Extension 11 SP3, your CIB version will *not* be upgraded by default. To manually upgrade to the latest CIB version use one of the following commands:

```
root # cibadmin --upgrade --force
```

or

```
root # crm configure upgrade force
```

---

### **IMPORTANT: Time Limit for Rolling Upgrade**

The new features shipped with SUSE Linux Enterprise High Availability Extension 11 SP4 will only be available after *all* cluster nodes have been upgraded to the latest product version. Mixed SP3/SP4 clusters are only supported for a short time frame during the rolling upgrade. Complete the rolling upgrade within one week.

---

## **E.3 Updating Software Packages on Cluster Nodes**

---

### **WARNING: Active Cluster Stack During Update**

To update any software packages on a node that is part of a running cluster, *stop* the cluster stack *on that node* before starting the software update. In some cases (see Conditions for Stopping the Cluster Stack (page 348), you can alternatively put the cluster into maintenance mode (which is available since SUSE Linux Enterprise High Availability Extension 11 SP4).

If the cluster resource manager on a node is active during the software update, this can lead to unpredictable results like fencing of active nodes.

---

Before installing any package updates on a node, check the following:

#### ***Conditions for Stopping the Cluster Stack***

- Does the update affect any packages belonging to SUSE Linux Enterprise High Availability Extension or the Geo Clustering for SUSE Linux Enterprise High

Availability Extension add-on? If yes: Stop the cluster stack on the node before starting the software update.

```
root # rcopenais stop
```

- Does the package update require a reboot? If yes: Stop the cluster stack on the node before starting the software update:

```
root # rcopenais stop
```

If none of the situations above apply, you do not need to stop the cluster stack. In that case, put the cluster into maintenance mode before starting the software update:

```
root # crm configure property maintenance-mode=true
```

For more details on maintenance mode, see Section 4.7, “Maintenance Mode” (page 88).

After the update has been successfully installed, remove the cluster maintenance mode:

```
root # crm configure property maintenance-mode=true
```

or restart the cluster stack on the respective node with:

```
root # rcopenais start
```

## E.4 For More Information

For detailed information about any changes and new features of the product you are upgrading to, refer to its release notes. They are available from <https://www.suse.com/releasesnotes/>.



# What's New?

The most important software modifications from version to version are outlined in the following sections. This summary indicates, for example, whether basic settings have been completely reconfigured, configuration files have been moved to other places, or other significant changes happened.

For more details and the most recent information, refer to the release notes of the respective product version. They are available in the installed system at `/usr/share/doc/release-notes`, or online at <https://www.suse.com/releasesnotes/>.

## F.1 Version 10 SP3 to Version 11

With SUSE Linux Enterprise Server 11, the cluster stack has changed from Heartbeat to OpenAIS. OpenAIS implements an industry standard API, the Application Interface Specification (AIS), published by the Service Availability Forum. The cluster resource manager from SUSE Linux Enterprise Server 10 has been retained but has been significantly enhanced, ported to OpenAIS and is now known as Pacemaker.

For more details what changed in the High Availability components from SUSE® Linux Enterprise Server 10 SP3 to SUSE Linux Enterprise Server 11, refer to the following sections.

## F.1.1 New Features and Functions Added

### Migration Threshold and Failure Timeouts

The High Availability Extension now comes with the concept of a migration threshold and failure timeout. You can define a number of failures for resources, after which they will migrate to a new node. By default, the node will no longer be allowed to run the failed resource until the administrator manually resets the resource's failcount. However it is also possible to expire them by setting the resource's `failure-timeout` option.

### Resource and Operation Defaults

You can now set global defaults for resource options and operations.

### Support for Offline Configuration Changes

Often it is desirable to preview the effects of a series of changes before updating the configuration atomically. You can now create a “shadow” copy of the configuration that can be edited with the command line interface, before committing it and thus changing the active cluster configuration atomically.

### Reusing Rules, Options and Sets of Operations

Rules, `instance_attributes`, `meta_attributes` and sets of operations can be defined once and referenced in multiple places.

### Using XPath Expressions for Certain Operations in the CIB

The CIB now accepts XPath-based `create`, `modify`, `delete` operations. For more information, refer to the `cibadmin help` text.

### Multi-dimensional Colocation and Ordering Constraints

For creating a set of collocated resources, previously you could either define a resource group (which could not always accurately express the design) or you could define each relationship as an individual constraint—causing a constraint explosion as the number of resources and combinations grew. Now you can also use an alternate form of colocation constraints by defining `resource_sets`.

### Connection to the CIB From Non-cluster Machines

Provided Pacemaker is installed on a machine, it is possible to connect to the cluster even if the machine itself is not a part of it.

### Triggering Recurring Actions at Known Times

By default, recurring actions are scheduled relative to when the resource started, but this is not always desirable. To specify a date/time that the operation should

be relative to, set the operation's interval-origin. The cluster uses this point to calculate the correct start-delay such that the operation will occur at origin + (interval \* N).

## F.1.2 Changed Features and Functions

### Naming Conventions for Resource and Cluster Options

All resource and cluster options now use dashes (-) instead of underscores (\_). For example, the `master_max` meta option has been renamed to `master-max`.

### Renaming of `master_slave` Resource

The `master_slave` resource has been renamed to `master`. Master resources are a special type of clone that can operate in one of two modes.

### Container Tag for Attributes

The `attributes` container tag has been removed.

### Operation Field for Prerequisites

The `pre-req` operation field has been renamed `requires`.

### Interval for Operations

All operations must have an interval. For start/stop actions the interval must be set to 0 (zero).

### Attributes for Colocation and Ordering Constraints

The attributes of colocation and ordering constraints were renamed for clarity.

### Cluster Options for Migration Due to Failure

The `resource-failure-stickiness` cluster option has been replaced by the `migration-threshold` cluster option. See also “Migration Threshold and Failure Timeouts” (page 352).

### Arguments for Command Line Tools

The arguments for command-line tools have been made consistent. See also “Naming Conventions for Resource and Cluster Options” (page 353).

### Validating and Parsing XML

The cluster configuration is written in XML. Instead of a Document Type Definition (DTD), now a more powerful RELAX NG schema is used to define the pattern for the structure and content. `libxml2` is used as parser.

#### id Fields

id fields are now XML IDs which have the following limitations:

- IDs cannot contain colons.
- IDs cannot begin with a number.
- IDs must be globally unique (not just unique for that tag).

#### References to Other Objects

Some fields (such as those in constraints that refer to resources) are IDREFs.

This means that they must reference existing resources or objects in order for the configuration to be valid. Removing an object which is referenced elsewhere will therefore fail.

## F.1.3 Removed Features and Functions

#### Setting Resource Meta Options

It is no longer possible to set resource meta-options as top-level attributes. Use meta attributes instead. See also the `crm_resource` man page.

#### Setting Global Defaults

Resource and operation defaults are no longer read from `crm_config`.

## F.2 Version 11 to Version 11 SP1

#### Cluster Configuration File

The main cluster configuration file has changed from `/etc/ais/openais.conf` to `/etc/corosync/corosync.conf`. Both files are very similar. When upgrading from SUSE Linux Enterprise High Availability Extension 11 to SP1, a script takes care of the minor differences between those files.

#### Rolling Upgrade

In order to migrate existing clusters with minimal downtime, SUSE Linux Enterprise High Availability Extension allows you to perform a “rolling upgrade” from SUSE Linux Enterprise High Availability Extension 11 to 11 SP1. The cluster is still online while you upgrade one node after the other.



## Automatic Cluster Deployment

For easier cluster deployment, AutoYaST allows you to clone existing nodes. AutoYaST is a system for installing one or more SUSE Linux Enterprise systems automatically and without user intervention, using an AutoYaST profile that contains installation and configuration data. The profile tells AutoYaST what to install and how to configure the installed system to get a completely ready-to-use system in the end. This profile can be used for mass deployment in different ways.

## Transfer of Configuration Files

SUSE Linux Enterprise High Availability Extension ships with Csync2, a tool for replication of configuration files across all nodes in the cluster. It can handle any number of hosts and it is also possible to synchronize files among certain subgroups of hosts only. Use YaST to configure the host names and the files that should be synchronized with Csync2.

## Web-Interface for Cluster Management

The High Availability Extension now also includes the HA Web Konsole (Hawk), a Web-based user interface for management tasks. It allows you to monitor and administer your Linux cluster also from non-Linux machines. It is also an ideal solution in case your system does not provide or allow a graphical user interface.

## Templates for Resource Configuration

When using the command line interface to create and configure resources, you can now choose from various resource templates for quicker and easier configuration.

## Load-based Placement of Resources

By defining the capacity a certain node *provides*, the capacity a certain resource *requires* and by choosing one of several placement strategies in the cluster, resources can be placed according to their load impact to prevent decrease of cluster performance.

## Cluster-aware Active/Active RAID1

It is now possible to create disaster-resilient storage configurations from two independent SANs, using `cmirror`.

## Read-only GFS2 Support

For easier migration from GFS2 to OCFS2, you can mount your GFS2 file systems in read-only mode to copy the data to an OCFS2 file system. OCFS2 is fully supported by SUSE Linux Enterprise High Availability Extension.

### SCTP Support for OCFS2

If redundant rings are configured, OCFS2 and DLM can automatically use redundant communication paths via SCTP, independent of network device bonding.

### Storage Protection

For additional layers of security in protecting your storage from data corruption, you can use a combination of IO fencing (with the `external/sbd` fencing device) and the `sfex` resource agent to ensure exclusive storage access.

### Samba Clustering

The High Availability Extension now supports CTDB, the cluster implementation of the trivial database. This allows you configure a clustered Samba server—providing an High Availability solution also for heterogeneous environments.

### YaST Module for IP Load Balancing

The new module allows configuration of Kernel-based load balancing with a graphical user interface. It is a front-end for `ldirectord`, a user-space daemon for managing Linux Virtual Server and monitoring the real servers.

## F.3 Version 11 SP1 to Version 11 SP2

### *Geo Clusters (Multi-Site Clusters)* (page 241)

Apart from local clusters and metro area clusters, SUSE® Linux Enterprise High Availability Extension 11 SP4 also supports multi-site clusters. That means you can have multiple, geographically dispersed sites with a local cluster each. Failover between these clusters is coordinated by a higher level entity, the so-called `booth`. Support for multi-site clusters is available as a separate option to SUSE Linux Enterprise High Availability Extension.

### *Access Control Lists* (page 215)

For defining fine-grained access rights to any part of the cluster configuration ACLs are supported. If this feature is enabled in the CRM, the available functions in the cluster management tools depend on the role and access rights assigned to a user.

### Automatic Cluster Setup (sleha-bootstrap) (page 30)

For quick and easy cluster setup, use the bootstrap scripts `sleha-init` and `sleha-join` to get a one-node cluster up and running in a few minutes and to make other nodes join, respectively. Any options set during the bootstrap process can be modified later with the YaST cluster module.

### Corosync Unicast Mode

While multicast is still default, using unicast for the communication between nodes is now also supported. For more information, refer to Section 3.5.2, “Defining the Communication Channels” (page 36).

### HA Web Konsole (Hawk)

Hawk's functionality has been considerably extended. Now you can configure global cluster properties, basic and advanced types of resources, constraints and resource monitoring. For detailed analysis of the cluster status, Hawk generates a cluster report (`hb_report`). View the cluster history or explore potential failure scenarios with the simulator. For details, refer to Chapter 5, *Configuring and Managing Cluster Resources (Web Interface)* (page 91).

### Resource Templates (page 59)

To ease configuration of similar resources, all cluster management tools now let you define resource templates that can be referenced in primitives or certain types of constraints.

### Virtualization and Cloud Integration

For placing resources based on load impact, the High Availability Extension now offers automatic detection of both the capacity of a node and the capacities a resource requires. The minimal requirements of a virtual machine (for example, the memory assigned to a Xen or KVM guest or the number of CPU cores) can be detected by a resource agent. Utilization attributes (used to define the requirements or capacity) will automatically be added to the CIB. For more information, refer to Section 4.4.6, “Placing Resources Based on Their Load Impact” (page 80).

To protect a node's network connection from being overloaded by a large number of parallel Xen or KVM live migrations, a new global cluster property has been introduced: `migration-limit`. It allows you to limit the number of migration jobs that the TE may execute in parallel on a node. By default, it is set to `-1`, which means the number of parallel migrations is unlimited.

### conntrack Tools

To synchronize the connection status between cluster nodes, the High Availability Extension uses the `conntrack-tools`. They allow interaction with the in-kernel Connection Tracking System for enabling *stateful* packet inspection for iptables. For more information, refer to Section 3.5.5, “Synchronizing Connection Status Between Cluster Nodes” (page 45).

### Parallel SSH (`pssh`)

To execute commands on all cluster nodes without having to log in to each node, use `pssh`. For more information, refer to Section 20.5, “Miscellaneous” (page 325).

### `crm resource secret`

To set passwords for STONITH or other resources independent of `cib.xml`, use `crm resource secret`. For more information, refer to Section 7.5, “Setting Passwords Independent of `cib.xml`” (page 194).

### Samba Clustering

The CTDB functionality to join Active Directory Domains has been improved. For more information, refer to Section 18.3, “Joining an Active Directory Domain” (page 301).

### *Disaster Recovery with Rear* (page 307)

Rear (Relax and Recover) is an administrator tool-set for creating disaster recovery images. The disaster recovery information can either be stored via the network or locally on hard disks, USB devices, DVD/CD-R, tape or similar. The backup data is stored on a network file system (NFS).

### Quotas on OCFS2

To use quotas on OCFS2 file systems, create and mount the files system with the appropriate quota features or mount options, respectively: `ursquota` (quota for individual users) or `grpquota` (quota for groups).

## F.4 Version 11 SP2 to Version 11 SP3

### *Configuring and Managing Cluster Resources (Web Interface)* (page 91)

Hawk's functionality has again been extended. For example, you can monitor multiple clusters with Hawk's new *Cluster Dashboard*. Hawk's simulator

now also allows you to change the status of nodes, add or edit resources and constraints, or change the cluster configuration for a simulator run. Apart from this, many other details in the HA Web Konsole have been enhanced.

#### Pacemaker GUI

The X11-based Pacemaker GUI is now in maintenance mode and is not scheduled to receive new functionality during the SUSE Linux Enterprise High Availability Extension 11 lifecycle. For SUSE Linux Enterprise High Availability Extension 12, the HA Web Konsole (Hawk) will become the default graphical user interface for the High Availability Extension.

#### Removing Nodes From An Existing Cluster (page 34)

The `sleha-remove` bootstrap script now makes it easier to remove single nodes from a cluster.

#### Using Maintenance Mode (page 123)

Sometimes it is necessary to put single nodes into maintenance mode. If your cluster consists of more than 3 nodes, you can easily set one node to maintenance mode, while the other nodes continue their normal operation.

#### Configuring a Cluster Resource Group (page 188)

The `group` command in the `crm` shell has been extended to allow modification of groups: it is possible to add resources to a group, to delete resources from a group, and to change the order of group members.

#### Utilization Attributes for Groups

If multiple resources with utilization attributes are grouped or have colocation constraints, the High Availability Extension takes that into account. If possible, the resources will be placed on a node that can fulfill *all* capacity requirements. For details about utilization attributes, refer to Section 4.4.6, “Placing Resources Based on Their Load Impact” (page 80).

#### Default Probe Timeout

If no specific timeout is configured for the resource's monitoring operation, the CRM will now automatically calculate a timeout for probing. For details, refer to Section 4.2.9, “Timeout Values” (page 71). Up to now, the default timeout for probing had been inherited from the cluster-wide `default operation timeout` if no specific timeout was configured.

#### Monitoring System Health (page 87)

To avoid a node running out of disk space and thus being no longer able to adequately manage any resources that have been assigned

to it, the High Availability Extension provides a resource agent, `ocf:pacemaker:SysInfo`. Use it to monitor a node's health with respect to disk partitions.

Section 4.5.1, “Monitoring Services on Remote Hosts with Nagios Plugins” (page 84)

#### Cluster Diagram

Both the `crm` shell and Hawk now offer the possibility to display a graphical representation of the nodes and the resources configured in the CIB. For details, refer to Section 7.1.7, “Cluster Diagram” (page 177) and Section 5.1.2, “Main Screen: Cluster Status” (page 94).

#### Hotplugging of Bonding Slaves (page 228)

Sometimes it is necessary to replace a bonding slave interface with another one, for example, if the respective network device constantly fails. The solution is to set up hotplugging of bonding slaves, which is now supported by YaST.

#### Configuring Cmirror (page 273)

The High Availability Extension supports RAID 10 in `cmirror`: it is now possible to add multiple physical volumes per mirror leg. Also the `mirrored` option of the `lvcreate` command is supported, which means that temporarily broken mirrors are much faster to resynchronize.

#### Joining an Active Directory Domain (page 301)

YaST now supports the CTDB functionality to join Active Directory domains.

## F.5 Version 11 SP3 to Version 11 SP4

### General

- Introduced a consistent naming scheme for cluster names, node names, cluster resources, and constraints and applied it to the documentation. See Appendix A, *Naming Conventions* (page 327). (Fate#314938).
- Improved the consistency of `crm` shell examples.

- Removed chapter *HA OCF Agents* with the agent list. As a result, removed the part *Troubleshooting and Reference*, too, and moved the chapter *Troubleshooting* (page 319) to the appendix.
- Moved documentation for Geo Clustering for SUSE Linux Enterprise High Availability Extension into a separate document (Fate#316120). For details on how to use and set up geographically dispersed clusters, refer to the Quick Start Geo Clustering for SUSE Linux Enterprise High Availability Extension. It is available from <http://www.suse.com/documentation/> or in your installed system under `/usr/share/doc/manual/sle-ha-geo-manuals_en/`.
- Changed terminology for `master-slave` resources, which are now called `multi-state` resources in the upstream documentation.
- Updated the screenshots.
- Mentioned both `hb_report` and `crm_report` as command line tools for creating detailed cluster reports.
- Numerous bug fixes and additions to the manual based on technical feedback.

#### Chapter 1, *Product Overview* (page 3)

- Changed terminology from multi-site clusters to geographically dispersed (or Geo) clusters for consistency reasons.
- Added section about availability of SUSE Linux Enterprise High Availability Extension and Geo Clustering for SUSE Linux Enterprise High Availability Extension as add-on products: Section 1.1, “Availability as Add-On/Extension” (page 4).

#### Chapter 2, *System Requirements and Recommendations* (page 19)

- Restructured contents.
- Mentioned how to create a cluster report when using a non-standard SSH port (Fate#314906). See “SSH” (page 23).

#### Chapter 3, *Installation and Basic Setup* (page 25)

- Added note *No-Start-on-Boot Parameter* (Fate#317778).

## Chapter 4, *Configuration and Administration Basics* (page 53)

- Section 4.1.3, “Option `stonith-enabled`” (page 55) mentions policy change in DLM services when the global cluster option `stonith-enabled` is set to `false` (Fate#315195).
- Section 4.4.7, “Grouping Resources by Using Tags” (page 83) describes a new option to group conceptually related resources, without creating any colocation or ordering relationship between them (Fate#318109).
- Section 4.4.1.1, “Resource Sets” (page 74) explains the concept of resource sets as an alternative format for defining constraints.
- Restructured Section 4.7, “Maintenance Mode” (page 88) to also cover the option of setting a whole cluster to maintenance mode.
- Added attributes for `pacemaker_remote` service to Table 4.1, “Options for a Primitive Resource” (page 64) and added new section: Section 4.5, “Managing Services on Remote Hosts” (page 84).

## Chapter 5, *Configuring and Managing Cluster Resources (Web Interface)* (page 91)

- Updated the chapter to reflect the new features that have been described in Chapter 4, *Configuration and Administration Basics* (page 53).
- The description of all Hawk functions that are related to Geo clusters has been moved to a separate document. See the new *Geo Clustering for SUSE Linux Enterprise High Availability Extension Quick Start*, available from <http://www.suse.com/documentation/>.

## Chapter 7, *Configuring and Managing Cluster Resources (Command Line)* (page 169)

- Added Section 7.3.4.3, “Collocating Sets for Resources Without Dependency” (page 183) (Fate#314917).
- Added a section about tagging resources (Fate#318109): Section 7.4.5, “Grouping/Tagging Resources” (page 192).
- Updated the chapter to reflect the new features that have been described in Chapter 4, *Configuration and Administration Basics* (page 53).



## Chapter 9, *Fencing and STONITH* (page 203)

- Removed cloning of STONITH resources from examples as this is no longer needed.
- Removed some STONITH devices that are for testing purposes only.
- Removed `external/kdumpcheck` resource agent and added example configuration for the `fence_kdump` resource agent instead.

## Chapter 10, *Access Control Lists* (page 215)

- Added Section 10.5, “Configuring ACLs with Hawk” (page 222).
- Updated chapter according to the new ACL features that become available after upgrading the CIB validation version. For details, see *Upgrading the CIB Syntax Version* (page 347).

If you have upgraded from SUSE Linux Enterprise High Availability Extension 11 SP3 and kept your former CIB version, refer to the *Access Control List* chapter in the High Availability Guide for SUSE Linux Enterprise High Availability Extension 11 SP3. It is available from <http://www.suse.com/documentation/>.

## Chapter 17, *Storage Protection* (page 285)

- Section 17.1.3.2, “Setting Up the Software Watchdog” (page 289): Added note about watchdog and other software that accesses the watchdog timer ([https://bugzilla.suse.com/show\\_bug.cgi?id=891340](https://bugzilla.suse.com/show_bug.cgi?id=891340)).
- Section 17.1.3.2, “Setting Up the Software Watchdog” (page 289): Added how to load the watchdog driver at boot time ([https://bugzilla.suse.com/show\\_bug.cgi?id=892344](https://bugzilla.suse.com/show_bug.cgi?id=892344)).
- Section 17.1.3.5, “Configuring the Fencing Resource” (page 292): Added advice about length of `stonith-timeout` in relation to `msgwait timeout` ([https://bugzilla.suse.com/show\\_bug.cgi?id=891346](https://bugzilla.suse.com/show_bug.cgi?id=891346)).
- Adjusted Section 17.1.3.3, “Starting the SBD Daemon” (page 291) ([https://bugzilla.suse.com/show\\_bug.cgi?id=891499](https://bugzilla.suse.com/show_bug.cgi?id=891499)).

- Described how to avoid double fencing in clusters with `no-quorum-policy=ignore` by using the `pcmk_delay_max` parameter for the STONITH resource configuration (Fate#31713).

## Chapter 19, *Disaster Recovery with Rear* (page 307)

The chapter has been completely revised and updated to Rear version 1.16.

SUSE Linux Enterprise High Availability Extension 11 SP4 ships two Rear versions in parallel: version 1.10.0 (included in RPM package: `rear`) and version 1.16 (included in RPM package `rear116`). For the documentation of Rear version 1.10.0, see the High Availability Guide for SUSE Linux Enterprise High Availability Extension 11 SP3. It is available from <http://www.suse.com/documentation/>.

## Chapter 20, *Troubleshooting* (page 319)

- Mentioned how to create a cluster report when using a non-standard SSH port (Fate#314906). See “How can I create a report with an analysis of all my cluster nodes?” (page 321).

## *HA OCF Agents*

This chapter has been removed. The latest information about the OCF resource agents can be viewed in the installed system as described in Section 7.1.3, “Displaying Information about OCF Resource Agents” (page 172).

## Appendix A, *Naming Conventions* (page 327)

- New appendix explaining the naming scheme used in the documentation.

## Appendix E, *Upgrading Your Cluster and Updating Software Packages* (page 339)

- Added definition of terms: Section E.1, “Terminology” (page 339).
- Added Section E.2.5, “Upgrading from SLE HA 11 SP3 to SLE HA 11 SP4” (page 347) to Section E.2, “Upgrading your Cluster to the Latest Product Version” (page 340).
- Added Section E.3, “Updating Software Packages on Cluster Nodes” (page 348).

# Terminology

## active/active, active/passive

A concept of how services are running on nodes. An active-passive scenario means that one or more services are running on the active node and the passive node waits for the active node to fail. Active-active means that each node is active and passive at the same time. For example, it has *some* services running, but can take over other services from the other node. Compare with primary/secondary and dual-primary in DRBD speak.

## arbiter

Additional instance in a Geo cluster that helps to reach consensus about decisions such as failover of resources across sites. Arbitrators are single machines that run one or more booth instances in a special mode.

## AutoYaST

AutoYaST is a system for installing one or more SUSE Linux Enterprise systems automatically and without user intervention.

## bindnetaddr (bind network address)

The network address the Corosync executive should bind to.

## booth

The instance that manages the failover process between the sites of a Geo cluster. It aims to get multi-site resources active on one and only one site. This is achieved by using so-called tickets that are treated as failover domain between cluster sites, in case a site should be down.

## boothd (booth daemon)

Each of the participating clusters and arbiters in a Geo cluster runs a service, the `boothd`. It connects to the booth daemons running at the other sites and exchanges connectivity details.

## CCM (consensus cluster membership)

The CCM determines which nodes make up the cluster and shares this information across the cluster. Any new addition and any loss of nodes or quorum is delivered by the CCM. A CCM module runs on each node of the cluster.

### CIB (cluster information base)

A representation of the whole cluster configuration and status (cluster options, nodes, resources, constraints and the relationship to each other). It is written in XML and resides in memory. A master CIB is kept and maintained on the DC (designated coordinator) and replicated to the other nodes. Normal read and write operations on the CIB are serialized through the master CIB.

### cluster

A *high-performance* cluster is a group of computers (real or virtual) sharing the application load to achieve faster results. A *high-availability* cluster is designed primarily to secure the highest possible availability of services.

### cluster partition

Whenever communication fails between one or more nodes and the rest of the cluster, a cluster partition occurs. The nodes of a cluster are split into partitions but still active. They can only communicate with nodes in the same partition and are unaware of the separated nodes. As the loss of the nodes on the other partition cannot be confirmed, a split brain scenario develops (see also split brain (page 370)).

### concurrency violation

A resource that should be running on only one node in the cluster is running on several nodes.

### conntrack tools

Allow interaction with the in-kernel connection tracking system for enabling *stateful* packet inspection for iptables. Used by the High Availability Extension to synchronize the connection status between cluster nodes.

### CRM (cluster resource manager)

The main management entity responsible for coordinating all non-local interactions. The High Availability Extension uses Pacemaker as CRM. Each node of the cluster has its own CRM instance, but the one running on the DC is the one elected to relay decisions to the other non-local CRMs and process their input. A CRM interacts with several components: local resource managers, both on its own node and on the other nodes, non-local CRMs, administrative commands, the fencing functionality, the membership layer, and booth.

### crmd (cluster resource manager daemon)

The CRM is implemented as daemon, crmd. It has an instance on each cluster node. All cluster decision-making is centralized by electing one of the crmd

instances to act as a master. If the elected crmd process fails (or the node it ran on), a new one is established.

## crmsh

The command line utility crmsh manages your cluster, nodes, and resources.

See Chapter 7, *Configuring and Managing Cluster Resources (Command Line)* (page 169) for more information.

## Csync2

A synchronization tool that can be used to replicate configuration files across all nodes in the cluster, and even across Geo clusters.

## DC (designated coordinator)

One CRM in the cluster is elected as the Designated Coordinator (DC). The DC is the only entity in the cluster that can decide that a cluster-wide change needs to be performed, such as fencing a node or moving resources around. The DC is also the node where the master copy of the CIB is kept. All other nodes get their configuration and resource allocation information from the current DC. The DC is elected from all nodes in the cluster after a membership change.

## Disaster

Unexpected interruption of critical infrastructure induced by nature, humans, hardware failure, or software bugs.

## Disaster Recovery

Disaster recovery is the process by which a business function is restored to the normal, steady state after a disaster.

## Disaster Recover Plan

A strategy to recover from a disaster with minimum impact on IT infrastructure.

## DLM (distributed lock manager)

DLM coordinates disk access for clustered file systems and administers file locking to increase performance and availability.

## DRBD

DRBD® is a block device designed for building high availability clusters. The whole block device is mirrored via a dedicated network and is seen as a network RAID-1.

### existing cluster

The term “existing cluster” is used to refer to any cluster that consists of at least one node. Existing clusters have a basic Corosync configuration that defines the communication channels, but they do not necessarily have resource configuration yet.

### failover

Occurs when a resource or node fails on one machine and the affected resources are started on another node.

### failover domain

A named subset of cluster nodes that are eligible to run a cluster service if a node fails.

### fencing

Describes the concept of preventing access to a shared resource by isolated or failing cluster members. Should a cluster node fail, it will be shut down or reset to prevent it from causing trouble. This way, resources are locked out of a node whose status is uncertain.

### geo cluster (geographically dispersed cluster)

See Geo cluster (page 369).

### load balancing

The ability to make several servers participate in the same service and do the same work.

### local cluster

A single cluster in one location (for example, all nodes are located in one data center). Network latency can be neglected. Storage is typically accessed synchronously by all nodes.

### LRM (local resource manager)

Responsible for performing operations on resources. It uses the resource agent scripts to carry out these operations. The LRM is “dumb” in that it does not know of any policy. It needs the DC to tell it what to do.

### mcastaddr (multicast address)

IP address to be used for multicasting by the Corosync executive. The IP address can either be IPv4 or IPv6.

**mcastport (multicast port)**

The port to use for cluster communication.

**metro cluster**

A single cluster that can stretch over multiple buildings or data centers, with all sites connected by fibre channel. Network latency is usually low (<5 ms for distances of approximately 20 miles). Storage is frequently replicated (mirroring or synchronous replication).

**multicast**

A technology used for a one-to-many communication within a network that can be used for cluster communication. Corosync supports both multicast and unicast.

**Geo cluster**

Consists of multiple, geographically dispersed sites with a local cluster each. The sites communicate via IP. Failover across the sites is coordinated by a higher-level entity, the booth. Geo clusters need to cope with limited network bandwidth and high latency. Storage is replicated asynchronously.

**node**

Any computer (real or virtual) that is a member of a cluster and invisible to the user.

**PE (policy engine)**

The policy engine computes the actions that need to be taken to implement policy changes in the CIB. The PE also produces a transition graph containing a list of (resource) actions and dependencies to achieve the next cluster state. The PE always runs on the DC.

**quorum**

In a cluster, a cluster partition is defined to have quorum (is “quorate”) if it has the majority of nodes (or votes). Quorum distinguishes exactly one partition. It is part of the algorithm to prevent several disconnected partitions or nodes from proceeding and causing data and service corruption (split brain). Quorum is a prerequisite for fencing, which then ensures that quorum is indeed unique.

**RA (resource agent)**

A script acting as a proxy to manage a resource (for example, to start, stop or monitor a resource). The High Availability Extension supports three different kinds of resource agents: OCF (Open Cluster Framework) resource agents, LSB (Linux Standards Base) resource agents (Standard LSB init scripts),

and Heartbeat resource agents. For more information, refer to Section 4.2.2, “Supported Resource Agent Classes” (page 57).

### Rear (Relax and Recover)

An administrator tool set for creating disaster recovery images.

### resource

Any type of service or application that is known to Pacemaker. Examples include an IP address, a file system, or a database.

The term “resource” is also used for DRBD, where it names a set of block devices that are using a common connection for replication.

### RRP (redundant ring protocol)

Allows the use of multiple redundant local area networks for resilience against partial or total network faults. This way, cluster communication can still be kept up as long as a single network is operational. Corosync supports the Totem Redundant Ring Protocol.

### SBD (STONITH block device)

In an environment where all nodes have access to shared storage, a small partition is used for disk-based fencing.

### SFEX (shared disk file exclusiveness)

SFEX provides storage protection over SAN.

### split brain

A scenario in which the cluster nodes are divided into two or more groups that do not know of each other (either through a software or hardware failure). STONITH prevents a split brain situation from badly affecting the entire cluster. Also known as a “partitioned cluster” scenario.

The term split brain is also used in DRBD but means that the two nodes contain different data.

### SPOF (single point of failure)

Any component of a cluster that, should it fail, triggers the failure of the entire cluster.

### STONITH

The acronym for “Shoot the other node in the head”. It refers to the fencing mechanism that shuts down a misbehaving node to prevent it from causing trouble in a cluster.



### switchover

Planned, on-demand moving of services to other nodes in a cluster. See failover (page 368).

### ticket

A component used in Geo clusters. A ticket grants the right to run certain resources on a specific cluster site. A ticket can only be owned by one site at a time. Resources can be bound to a certain ticket by dependencies. Only if the defined ticket is available at a site, the respective resources are started. Vice versa, if the ticket is removed, the resources depending on that ticket are automatically stopped.

### unicast

A technology for sending messages to a single network destination. Corosync supports both multicast and unicast. In Corosync, unicast is implemented as UDP-unicast (UDPU).





# GNU Licenses

This appendix contains the GNU Free Documentation License version 1.2.

## GNU Free Documentation License

Copyright (C) 2000, 2001, 2002 Free Software Foundation, Inc. 51 Franklin St, Fifth Floor, Boston, MA 02110-1301 USA. Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

## 0. PREAMBLE

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondly, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

## 1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "Invariant Sections" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "Cover Texts" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "Opaque".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "Title Page" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

A section "Entitled XYZ" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "Acknowledgements", "Dedications", "Endorsements", or "History".) To "Preserve the Title" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

## 2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

## 3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has

access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

## 4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

- A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.
- B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.
- C. State on the Title page the name of the publisher of the Modified Version, as the publisher.
- D. Preserve all the copyright notices of the Document.
- E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
- F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.
- G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.
- H. Include an unaltered copy of this License.
- I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.
- J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions if there was one. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.
- K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.
- L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.
- M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.
- N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.
- O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties--for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

## 5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements".

## 6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

## 7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

## 8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

## 9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

## 10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See <http://www.gnu.org/copyleft/>.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

## ADDENDUM: How to use this License for your documents

```
Copyright (c) YEAR YOUR NAME.  
Permission is granted to copy, distribute and/or modify this document  
under the terms of the GNU Free Documentation License, Version 1.2  
or any later version published by the Free Software Foundation;  
with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts.  
A copy of the license is included in the section entitled "GNU  
Free Documentation License".
```

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with...Texts." line with this:

```
with the Invariant Sections being LIST THEIR TITLES, with the  
Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.
```

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.

