Pacemaker Administration

Managing Pacemaker Clusters

Andrew Beekhof
Ken Gaillot
Pacemaker Administration
Managing Pacemaker Clusters
Edition 1

Author Andrew Beekhof andrew@beekhof.net
Author Ken Gaillot kgaillot@redhat.com

Copyright © 2009-2019 Andrew Beekhof.

The text of and illustrations in this document are licensed under version 4.0 or later of the Creative Commons Attribution-ShareAlike International Public License ("CC-BY-SA").

In accordance with CC-BY-SA, if you distribute this document or an adaptation of it, you must provide the URL for the original version.

In addition to the requirements of this license, the following activities are looked upon favorably:
1. If you are distributing Open Publication works on hardcopy or CD-ROM, you provide email notification to the authors of your intent to redistribute at least thirty days before your manuscript or media freeze, to give the authors time to provide updated documents. This notification should describe modifications, if any, made to the document.

2. All substantive modifications (including deletions) be either clearly marked up in the document or else described in an attachment to the document.

3. Finally, while it is not mandatory under this license, it is considered good form to offer a free copy of any hardcopy or CD-ROM expression of the author(s) work.

This document has instructions and tips for system administrators who need to manage high-availability clusters using Pacemaker.

1 An explanation of CC-BY-SA is available at https://creativecommons.org/licenses/by-sa/4.0/
Table of Contents

Preface .................................................................................................................. xi
  1. Document Conventions ...................................................................................... xi
     1.1. Typographic Conventions ................................................................................ xi
     1.2. Pull-quote Conventions .................................................................................. xii
     1.3. Notes and Warnings ..................................................................................... xiii
  2. We Need Feedback! ............................................................................................ xiii

1. Read-Me-First .................................................................................................... 1
   1.1. The Scope of this Document .......................................................................... 1
   1.2. What Is Pacemaker? ..................................................................................... 1
   1.3. Cluster Architecture .................................................................................... 2
   1.4. Pacemaker Architecture ............................................................................. 3
   1.5. Node Redundancy Designs .......................................................................... 4

2. Installing Cluster Software ............................................................................... 7
   2.1. Installing the Software ................................................................................ 7
   2.2. Enabling Pacemaker ................................................................................... 7
       2.2.1. Enabling Pacemaker For Corosync version 2 and greater .................. 7

3. The Cluster Layer ............................................................................................. 9
   3.1. Pacemaker and the Cluster Layer ................................................................ 9
   3.2. Managing Nodes in a Corosync-Based Cluster ........................................... 9
       3.2.1. Adding a New Corosync Node ................................................................. 9
       3.2.2. Removing a Corosync Node ................................................................. 9
       3.2.3. Replacing a Corosync Node ................................................................. 10

4. Configuring Pacemaker .................................................................................... 11
   4.1. How Should the Configuration be Updated? .............................................. 11
       4.1.1. Editing the CIB Using XML ................................................................. 11
       4.1.2. Quickly Deleting Part of the Configuration .......................................... 12
       4.1.3. Updating the Configuration Without Using XML ............................. 12
   4.2. Making Configuration Changes in a Sandbox .......................................... 13
   4.3. Testing Your Configuration Changes ....................................................... 14
       4.3.1. Small Cluster Transition ..................................................................... 15
       4.3.2. Complex Cluster Transition ............................................................... 16
   4.4. Do I Need to Update the Configuration on All Cluster Nodes? ............... 16
   4.5. Working with CIB Properties .................................................................. 17
   4.6. Querying and Setting Cluster Options .................................................... 17
       4.6.1. When Options are Listed More Than Once ......................................... 18
   4.7. Connecting from a Remote Machine .......................................................... 18

5. Monitoring a Pacemaker Cluster ...................................................................... 21
   5.1. Using crm_mon ......................................................................................... 21

6. Troubleshooting Cluster Problems .................................................................. 23
   6.1. Logging ..................................................................................................... 23
   6.2. Transitions .............................................................................................. 23
   6.3. Further Information About Troubleshooting ........................................... 24

7. Upgrading a Pacemaker Cluster ...................................................................... 25
   7.1. Pacemaker Versioning .............................................................................. 25
   7.2. Upgrading Cluster Software ..................................................................... 26
       7.2.1. Complete Cluster Shutdown ............................................................. 26
       7.2.2. Rolling (node by node) ................................................................. 27
       7.2.3. Detach and Reattach ...................................................................... 28
List of Figures

1.1. Example Cluster Stack .......................................................................................................... 3
1.2. Internal Components ............................................................................................................. 3
1.3. Active/Passive Redundancy ................................................................................................ 5
1.4. Shared Failover .................................................................................................................... 5
1.5. N to N Redundancy .............................................................................................................. 6
List of Tables

4.1. Environment Variables Used to Connect to Remote Instances of the CIB ............................... 18
4.2. Extra top-level CIB properties for remote access .................................................................. 18
7.1. Upgrade Methods ............................................................................................................... 26
7.2. Version Compatibility Table ............................................................................................... 28
8.1. Required Actions for OCF Agents ......................................................................................... 33
8.2. Optional Actions for OCF Resource Agents ........................................................................ 34
8.3. Types of recovery performed by the cluster ........................................................................ 34
8.4. OCF Return Codes and their Recovery Types ..................................................................... 35
List of Examples

2.1. Corosync configuration file for two nodes myhost1 and myhost2 .............................................. 7
2.2. Corosync configuration file for three nodes myhost1, myhost2 and myhost3 .......................... 8
4.1. Safely using an editor to modify the cluster configuration .......................................................... 11
4.2. Safely using an editor to modify only the resources section ...................................................... 12
4.3. Searching for STONITH-related configuration items .............................................................. 12
4.4. Creating and displaying the active sandbox ............................................................................. 13
4.5. Use sandbox to make multiple changes all at once, discard them, and verify real
    configuration is untouched ........................................................................................................... 13
4.6. Attributes set for a cib object .................................................................................................... 17
4.7. Deleting an option that is listed twice .......................................................................................... 18
5.1. Sample output from crm_mon -1 ............................................................................................... 21
5.2. Sample output from crm_mon -n -1 ........................................................................................... 21
Preface

Table of Contents

1. Document Conventions ................................................................. xi
   1.1. Typographic Conventions ......................................................... xi
   1.2. Pull-quote Conventions ........................................................... xii
   1.3. Notes and Warnings ............................................................... xiii
2. We Need Feedback! ...................................................................... xiii

1. Document Conventions

This manual uses several conventions to highlight certain words and phrases and draw attention to
specific pieces of information.

In PDF and paper editions, this manual uses typefaces drawn from the Liberation Fonts\(^1\) set. The
Liberation Fonts set is also used in HTML editions if the set is installed on your system. If not,
alternative but equivalent typefaces are displayed. Note: Red Hat Enterprise Linux 5 and later include
the Liberation Fonts set by default.

1.1. Typographic Conventions

Four typographic conventions are used to call attention to specific words and phrases. These
conventions, and the circumstances they apply to, are as follows.

**Mono-spaced Bold**

Used to highlight system input, including shell commands, file names and paths. Also used to highlight
keys and key combinations. For example:

To see the contents of the file `my_next_bestselling_novel` in your current
working directory, enter the `cat my_next_bestselling_novel` command at the
shell prompt and press `Enter` to execute the command.

The above includes a file name, a shell command and a key, all presented in mono-spaced bold and
all distinguishable thanks to context.

Key combinations can be distinguished from an individual key by the plus sign that connects each part
of a key combination. For example:

Press `Enter` to execute the command.

Press `Ctrl+Alt+F2` to switch to a virtual terminal.

The first example highlights a particular key to press. The second example highlights a key
combination: a set of three keys pressed simultaneously.

If source code is discussed, class names, methods, functions, variable names and returned values
mentioned within a paragraph will be presented as above, in **mono-spaced bold**. For example:

\(^1\) [https://fedorahosted.org/liberation-fonts/](https://fedorahosted.org/liberation-fonts/)
File-related classes include **filesystem** for file systems, **file** for files, and **dir** for directories. Each class has its own associated set of permissions.

**Proportional Bold**

This denotes words or phrases encountered on a system, including application names; dialog box text; labeled buttons; check-box and radio button labels; menu titles and sub-menu titles. For example:

Choose **System → Preferences → Mouse** from the main menu bar to launch **Mouse Preferences**. In the **Buttons** tab, select the **Left-handed mouse** check box and click **Close** to switch the primary mouse button from the left to the right (making the mouse suitable for use in the left hand).

To insert a special character into a **gedit** file, choose **Applications → Accessories → Character Map** from the main menu bar. Next, choose **Search → Find…** from the **Character Map** menu bar, type the name of the character in the **Search** field and click **Next**. The character you sought will be highlighted in the **Character Table**. Double-click this highlighted character to place it in the **Text to copy** field and then click the **Copy** button. Now switch back to your document and choose **Edit → Paste** from the **gedit** menu bar.

The above text includes application names; system-wide menu names and items; application-specific menu names; and buttons and text found within a GUI interface, all presented in proportional bold and all distinguishable by context.

**Mono-spaced Bold Italic or Proportional Bold Italic**

Whether mono-spaced bold or proportional bold, the addition of italics indicates replaceable or variable text. Italics denotes text you do not input literally or displayed text that changes depending on circumstance. For example:

To connect to a remote machine using ssh, type `ssh username@domain.name` at a shell prompt. If the remote machine is `example.com` and your username on that machine is john, type `ssh john@example.com`.

The `mount -o remount file-system` command remounts the named file system. For example, to remount the `/home` file system, the command is `mount -o remount /home`.

To see the version of a currently installed package, use the `rpm -q package` command. It will return a result as follows: `package-version-release`.

Note the words in bold italics above — username, domain.name, file-system, package, version and release. Each word is a placeholder, either for text you enter when issuing a command or for text displayed by the system.

Aside from standard usage for presenting the title of a work, italics denotes the first use of a new and important term. For example:

Publican is a DocBook publishing system.

### 1.2. Pull-quote Conventions

Terminal output and source code listings are set off visually from the surrounding text.

Output sent to a terminal is set in **mono-spaced roman** and presented thus:
Source-code listings are also set in **mono-spaced** roman but add syntax highlighting as follows:

```java
package org.jboss.book.jca.ex1;

import javax.naming.InitialContext;

public class ExClient {
    public static void main(String args[])
        throws Exception {
        InitialContext iniCtx = new InitialContext();
        Object         ref    = iniCtx.lookup("EchoBean");
        EchoHome       home   = (EchoHome) ref;
        Echo           echo   = home.create();

        System.out.println("Created Echo");
        System.out.println("Echo.echo('Hello') = " + echo.echo("Hello"));
    }
}
```

### 1.3. Notes and Warnings

Finally, we use three visual styles to draw attention to information that might otherwise be overlooked.

**Note**

Notes are tips, shortcuts or alternative approaches to the task at hand. Ignoring a note should have no negative consequences, but you might miss out on a trick that makes your life easier.

**Important**

Important boxes detail things that are easily missed: configuration changes that only apply to the current session, or services that need restarting before an update will apply. Ignoring a box labeled ‘Important’ will not cause data loss but may cause irritation and frustration.

**Warning**

Warnings should not be ignored. Ignoring warnings will most likely cause data loss.

### 2. We Need Feedback!
If you find a typographical error in this manual, or if you have thought of a way to make this manual better, we would love to hear from you! Please submit a report in Bugzilla\(^2\) against the product *Pacemaker Administration*.

When submitting a bug report, be sure to mention the manual's identifier: *Pacemaker_Administration*

If you have a suggestion for improving the documentation, try to be as specific as possible when describing it. If you have found an error, please include the section number and some of the surrounding text so we can find it easily.

\(^{2}\) [http://bugs.clusterlabs.org](http://bugs.clusterlabs.org)
Read-Me-First

Table of Contents

1.1. The Scope of this Document ................................................................................................. 1
1.2. What Is Pacemaker? ............................................................................................................. 1
1.3. Cluster Architecture ............................................................................................................... 2
1.4. Pacemaker Architecture ........................................................................................................ 3
1.5. Node Redundancy Designs ................................................................................................... 4

1.1. The Scope of this Document
The purpose of this document is to help system administrators learn how to manage a Pacemaker cluster.

System administrators may be interested in other parts of the Pacemaker documentation set, such as Clusters from Scratch, a step-by-step guide to setting up an example cluster, and Pacemaker Explained, an exhaustive reference for cluster configuration.

Multiple higher-level tools (both command-line and GUI) are available to simplify cluster management. However, this document focuses on the lower-level command-line tools that come with Pacemaker itself. The concepts are applicable to the higher-level tools, though the syntax would differ.

1.2. What Is Pacemaker?
Pacemaker is a high-availability cluster resource manager — software that runs on a set of hosts (a cluster of nodes) in order to preserve integrity and minimize downtime of desired services (resources). It is maintained by the ClusterLabs community.

Pacemaker's key features include:

• Detection of and recovery from node- and service-level failures
• Ability to ensure data integrity by fencing faulty nodes
• Support for one or more nodes per cluster
• Support for multiple resource interface standards (anything that can be scripted can be clustered)
• Support (but no requirement) for shared storage
• Support for practically any redundancy configuration (active/passive, N+1, etc.)
• Automatically replicated configuration that can be updated from any node
• Ability to specify cluster-wide relationships between services, such as ordering, colocation and anti-colocation

1 https://www.clusterlabs.org/pacemaker/doc/
2 Cluster is sometimes used in other contexts to refer to hosts grouped together for other purposes, such as high-performance computing (HPC), but Pacemaker is not intended for those purposes.
3 https://www.ClusterLabs.org/
• Support for advanced service types, such as *clones* (services that need to be active on multiple nodes), *stateful resources* (clones that can run in one of two modes), and containerized services

• Unified, scriptable cluster management tools

### Fencing

*Fencing*, also known as *STONITH* (an acronym for Shoot The Other Node In The Head), is the ability to ensure that it is not possible for a node to be running a service. This is accomplished via *fence devices* such as intelligent power switches that cut power to the target, or intelligent network switches that cut the target’s access to the local network.

Pacemaker represents fence devices as a special class of resource.

A cluster cannot safely recover from certain failure conditions, such as an unresponsive node, without fencing.

### 1.3. Cluster Architecture

At a high level, a cluster can viewed as having these parts (which together are often referred to as the *cluster stack*):

• **Resources**: These are the reason for the cluster’s being — the services that need to be kept highly available.

• **Resource agents**: These are scripts or operating system components that start, stop, and monitor resources, given a set of resource parameters. These provide a uniform interface between Pacemaker and the managed services.

• **Fence agents**: These are scripts that execute node fencing actions, given a target and fence device parameters.

• **Cluster membership layer**: This component provides reliable messaging, membership, and quorum information about the cluster. Currently, Pacemaker supports *Corosync* as this layer.

• **Cluster resource manager**: Pacemaker provides the brain that processes and reacts to events that occur in the cluster. These events may include nodes joining or leaving the cluster; resource events caused by failures, maintenance, or scheduled activities; and other administrative actions. To achieve the desired availability, Pacemaker may start and stop resources and fence nodes.

• **Cluster tools**: These provide an interface for users to interact with the cluster. Various command-line and graphical (GUI) interfaces are available.

Most managed services are not, themselves, cluster-aware. However, many popular open-source cluster filesystems make use of a common *Distributed Lock Manager* (DLM), which makes direct use of *Corosync* for its messaging and membership capabilities and Pacemaker for the ability to fence nodes.

---

4 http://www.corosync.org/
1.4. Pacemaker Architecture

Pacemaker itself is composed of multiple daemons that work together:

- pacemakerd
- pacemaker-atrtd
- pacemaker-based
- pacemaker-controld
- pacemaker-execd
- pacemaker-fenced
- pacemaker-schedulerd
Chapter 1. Read-Me-First

The Pacemaker master process (pacemakerd) spawns all the other daemons, and respawns them if they unexpectedly exit.

The Cluster Information Base (CIB) is an XML representation of the cluster’s configuration and the state of all nodes and resources. The CIB manager (pacemaker-based) keeps the CIB synchronized across the cluster, and handles requests to modify it.

The attribute manager (pacemaker-attd) maintains a database of attributes for all nodes, keeps it synchronized across the cluster, and handles requests to modify them. These attributes are usually recorded in the CIB.

Given a snapshot of the CIB as input, the scheduler (pacemaker-schedulerd) determines what actions are necessary to achieve the desired state of the cluster.

The local executor (pacemaker-execd) handles requests to execute resource agents on the local cluster node, and returns the result.

The fencer (pacemaker-fenced) handles requests to fence nodes. Given a target node, the fencer decides which cluster node(s) should execute which fencing device(s), and calls the necessary fencing agents (either directly, or via requests to the fencer peers on other nodes), and returns the result.

The controller (pacemaker-controld) is Pacemaker’s coordinator, maintaining a consistent view of the cluster membership and orchestrating all the other components.

Pacemaker centralizes cluster decision-making by electing one of the controller instances as the Designated Controller (DC). Should the elected DC process (or the node it is on) fail, a new one is quickly established. The DC responds to cluster events by taking a current snapshot of the CIB, feeding it to the scheduler, then asking the executors (either directly on the local node, or via requests to controller peers on other nodes) and the fencer to execute any necessary actions.

Old daemon names

The Pacemaker daemons were renamed in version 2.0. You may still find references to the old names, especially in documentation targeted to version 1.1.

<table>
<thead>
<tr>
<th>Old name</th>
<th>New name</th>
</tr>
</thead>
<tbody>
<tr>
<td>attrd</td>
<td>pacemaker-attd</td>
</tr>
<tr>
<td>cib</td>
<td>pacemaker-based</td>
</tr>
<tr>
<td>crmd</td>
<td>pacemaker-controld</td>
</tr>
<tr>
<td>lrmd</td>
<td>pacemaker-execd</td>
</tr>
<tr>
<td>stonithd</td>
<td>pacemaker-fenced</td>
</tr>
<tr>
<td>pacemaker_remoted</td>
<td>pacemaker-remoted</td>
</tr>
</tbody>
</table>

1.5. Node Redundancy Designs

Pacemaker supports practically any node redundancy configuration including Active/Active, Active/Passive, N+1, N+M, N-to-1 and N-to-N.

---


Active/passive clusters with two (or more) nodes using Pacemaker and DRBD are a cost-effective high-availability solution for many situations. One of the nodes provides the desired services, and if it fails, the other node takes over.

Pacemaker also supports multiple nodes in a shared-failover design, reducing hardware costs by allowing several active/passive clusters to be combined and share a common backup node.

When shared storage is available, every node can potentially be used for failover. Pacemaker can even run multiple copies of services to spread out the workload.

---


Figure 1.5. N to N Redundancy
Chapter 2.

Installing Cluster Software

Table of Contents

2.1. Installing the Software ........................................................................................................... 7
2.2. Enabling Pacemaker ............................................................................................................. 7
   2.2.1. Enabling Pacemaker For Corosync version 2 and greater ............................................. 7

2.1. Installing the Software

Most major Linux distributions have pacemaker packages in their standard package repositories, or the software can be built from source code. See the Install wiki page for details.

2.2. Enabling Pacemaker

2.2.1. Enabling Pacemaker For Corosync version 2 and greater

High-level cluster management tools are available that can configure corosync for you. This document focuses on the lower-level details if you want to configure corosync yourself.

Corosync configuration is normally located in `/etc/corosync/corosync.conf`.

Example 2.1. Corosync configuration file for two nodes myhost1 and myhost2

```plaintext
totem {
   version: 2
   secauth: off
   cluster_name: mycluster
   transport: udpu
 }

nodelist {
   node {
      ring0_addr: myhost1
      nodeid: 1
   }
   node {
      ring0_addr: myhost2
      nodeid: 2
   }
 }

quorum {
   provider: corosync_votequorum
   two_node: 1
 }

logging {
   to_syslog: yes
 }
```

1 http://clusterlabs.org/wiki/Install
Example 2.2. Corosync configuration file for three nodes \texttt{myhost1}, \texttt{myhost2} and \texttt{myhost3}


totem {
  version: 2
  secauth: off
  cluster_name: mycluster
  transport: udpu
}

nodelist {
  node {
    ring0_addr: myhost1
    nodeid: 1
  }
  node {
    ring0_addr: myhost2
    nodeid: 2
  }
  node {
    ring0_addr: myhost3
    nodeid: 3
  };
}

quorum {
  provider: corosync_votequorum
}

logging {
  to_syslog: yes
}

In the above examples, the \texttt{totem} section defines what protocol version and options (including encryption) to use, \(^2\) and gives the cluster a unique name (\texttt{mycluster} in these examples).

The \texttt{node} section lists the nodes in this cluster.

The \texttt{quorum} section defines how the cluster uses quorum. The important thing is that two-node clusters must be handled specially, so \texttt{two_node: 1} must be defined for two-node clusters (and only for two-node clusters).

The \texttt{logging} section should be self-explanatory.

\(^2\) Please consult the Corosync website (http://www.corosync.org/) and documentation for details on enabling encryption and peer authentication for the cluster.
The Cluster Layer

Table of Contents

3.1. Pacemaker and the Cluster Layer ......................................................................................... 9
3.2. Managing Nodes in a Corosync-Based Cluster ....................................................................... 9
   3.2.1. Adding a New Corosync Node .................................................................................... 9
   3.2.2. Removing a Corosync Node ..................................................................................... 10
3.2.3. Replacing a Corosync Node .......................................................................................... 10

3.1. Pacemaker and the Cluster Layer

Pacemaker utilizes an underlying cluster layer for two purposes:

• obtaining quorum
• messaging between nodes

Currently, only Corosync 2 and later is supported for this layer.

3.2. Managing Nodes in a Corosync-Based Cluster

3.2.1. Adding a New Corosync Node

To add a new node:

1. Install Corosync and Pacemaker on the new host.

2. Copy /etc/corosync/corosync.conf and /etc/corosync/authkey (if it exists) from an existing node. You may need to modify the mcastaddr option to match the new node's IP address.

3. Start the cluster software on the new host. If a log message containing "Invalid digest" appears from Corosync, the keys are not consistent between the machines.

3.2.2. Removing a Corosync Node

Because the messaging and membership layers are the authoritative source for cluster nodes, deleting them from the CIB is not a complete solution. First, one must arrange for corosync to forget about the node (pcmk-1 in the example below).

1. Stop the cluster on the host to be removed. How to do this will vary with your operating system and installed versions of cluster software, for example, pcs cluster stop if you are using pcs for cluster management.

2. From one of the remaining active cluster nodes, tell Pacemaker to forget about the removed host, which will also delete the node from the CIB:
3.2.3. Replacing a Corosync Node

To replace an existing cluster node:

1. Make sure the old node is completely stopped.
2. Give the new machine the same hostname and IP address as the old one.
3. Follow the procedure above for adding a node.
Chapter 4.

Configuring Pacemaker

Table of Contents

4.1. How Should the Configuration be Updated? ................................................................. 11
   4.1.1. Editing the CIB Using XML ....................................................................................... 11
   4.1.2. Quickly Deleting Part of the Configuration ................................................................. 12
   4.1.3. Updating the Configuration Without Using XML .......................................................... 12
4.2. Making Configuration Changes in a Sandbox ................................................................. 13
4.3. Testing Your Configuration Changes .................................................................................... 14
   4.3.1. Small Cluster Transition ............................................................................................ 15
   4.3.2. Complex Cluster Transition ....................................................................................... 16
4.4. Do I Need to Update the Configuration on All Cluster Nodes? ............................................... 16
4.5. Working with CIB Properties ................................................................................................ 17
4.6. Querying and Setting Cluster Options .................................................................................. 17
   4.6.1. When Options are Listed More Than Once ................................................................ 18
4.7. Connecting from a Remote Machine .................................................................................... 18

4.1. How Should the Configuration be Updated?

There are three basic rules for updating the cluster configuration:

• Rule 1 - Never edit the cib.xml file manually. Ever. I’m not making this up.

• Rule 2 - Read Rule 1 again.

• Rule 3 - The cluster will notice if you ignored rules 1 & 2 and refuse to use the configuration.

Now that it is clear how not to update the configuration, we can begin to explain how you should.

4.1.1. Editing the CIB Using XML

The most powerful tool for modifying the configuration is the cibadmin command. With cibadmin, you can query, add, remove, update or replace any part of the configuration. All changes take effect immediately, so there is no need to perform a reload-like operation.

The simplest way of using cibadmin is to use it to save the current configuration to a temporary file, edit that file with your favorite text or XML editor, and then upload the revised configuration.  

Example 4.1. Safely using an editor to modify the cluster configuration

```bash
# cibadmin --query > tmp.xml
# vi tmp.xml
# cibadmin --replace --xml-file tmp.xml
```

Some of the better XML editors can make use of a Relax NG schema to help make sure any changes you make are valid. The schema describing the configuration can be found in pacemaker.rng.

---

1 This process might appear to risk overwriting changes that happen after the initial cibadmin call, but pacemaker will reject any update that is "too old". If the CIB is updated in some other fashion after the initial cibadmin, the second cibadmin will be rejected because the version number will be too low.
which may be deployed in a location such as /usr/share/pacemaker or /usr/lib/heartbeat depending on your operating system and how you installed the software.

If you want to modify just one section of the configuration, you can query and replace just that section to avoid modifying any others.

Example 4.2. Safely using an editor to modify only the resources section

# cibadmin --query --scope resources > tmp.xml
# vi tmp.xml
# cibadmin --replace --scope resources --xml-file tmp.xml

4.1.2. Quickly Deleting Part of the Configuration

Identify the object you wish to delete by XML tag and id. For example, you might search the CIB for all STONITH-related configuration:

Example 4.3. Searching for STONITH-related configuration items

# cibadmin -Q | grep stonith
<nvpair id="cib-bootstrap-options-stonith-action" name="stonith-action" value="reboot"/>
<nvpair id="cib-bootstrap-options-stonith-enabled" name="stonith-enabled" value="1"/>
<primitive id="child_DoFencing" class="stonith" type="external/vmware">
<lrm_resource id="child_DoFencing:0" type="external/vmware" class="stonith">
<lrm_resource id="child_DoFencing:0" type="external/vmware" class="stonith">
<lrm_resource id="child_DoFencing:1" type="external/vmware" class="stonith">
<lrm_resource id="child_DoFencing:0" type="external/vmware" class="stonith">
<lrm_resource id="child_DoFencing:2" type="external/vmware" class="stonith">
<lrm_resource id="child_DoFencing:0" type="external/vmware" class="stonith">
<lrm_resource id="child_DoFencing:3" type="external/vmware" class="stonith">

If you wanted to delete the primitive tag with id child_DoFencing, you would run:

# cibadmin --delete --xml-text '<primitive id="child_DoFencing"/>'

4.1.3. Updating the Configuration Without Using XML

Most tasks can be performed with one of the other command-line tools provided with pacemaker, avoiding the need to read or edit XML.

To enable STONITH for example, one could run:

# crm_attribute --name stonith-enabled --update 1

Or, to check whether somenode is allowed to run resources, there is:

# crm_standby --query --node somenode

Or, to find the current location of my-test-rsc, one can use:

# crm_resource --locate --resource my-test-rsc

Examples of using these tools for specific cases will be given throughout this document where appropriate.
4.2. Making Configuration Changes in a Sandbox

Often it is desirable to preview the effects of a series of changes before updating the configuration all at once. For this purpose, we have created `crm_shadow` which creates a "shadow" copy of the configuration and arranges for all the command line tools to use it.

To begin, simply invoke `crm_shadow --create` with the name of a configuration to create, and follow the simple on-screen instructions.

**Warning**

Read this section and the on-screen instructions carefully; failure to do so could result in destroying the cluster's active configuration!

Example 4.4. Creating and displaying the active sandbox

```
# crm_shadow --create test
Setting up shadow instance
Type Ctrl-D to exit the crm_shadow shell
shadow[test]:
shadow[test] # crm_shadow --which
```

From this point on, all cluster commands will automatically use the shadow copy instead of talking to the cluster's active configuration. Once you have finished experimenting, you can either make the changes active via the `--commit` option, or discard them using the `--delete` option. Again, be sure to follow the on-screen instructions carefully!

For a full list of `crm_shadow` options and commands, invoke it with the `--help` option.

Example 4.5. Use sandbox to make multiple changes all at once, discard them, and verify real configuration is untouched

```
shadow[test] # crm_failcount -r rsc_c001n01 -G scope=status name=fail-count-rsc_c001n01 value=0
shadow[test] # crm_standby --node c001n02 -v on
shadow[test] # crm_standby --node c001n02 -G scope=nodes name=standby value=on
shadow[test] # cibadmin --erase --force
shadow[test] # cibadmin --query
```

```
<cib crm_feature_set="3.0.14" validate-with="pacemaker-3.0" epoch="112" num_updates="2" admin_epoch="16" cib-last-written="Mon Jan  8 23:26:47 2018" update-origin="rhel7-1"
update-client="crm_node" update-user="root" have-quorum="1" dc-uuid="1">
<configuration>
<crm_config/>
<nodes/>
<resources/>
<constraints/>
</configuration>
<status/>
```

Shadow copies are identified with a name, making it possible to have more than one.

---

2 Shadow copies are identified with a name, making it possible to have more than one.
4.3. Testing Your Configuration Changes

We saw previously how to make a series of changes to a "shadow" copy of the configuration. Before loading the changes back into the cluster (e.g. `crm_shadow --commit mytest --force`), it is often advisable to simulate the effect of the changes with `crm_simulate`. For example:

```
# crm_simulate --live-check -VVVVV --save-graph tmp.graph --save-dotfile tmp.dot
```

This tool uses the same library as the live cluster to show what it would have done given the supplied input. Its output, in addition to a significant amount of logging, is stored in two files `tmp.graph` and `tmp.dot`. Both files are representations of the same thing: the cluster's response to your changes.

The graph file stores the complete transition from the existing cluster state to your desired new state, containing a list of all the actions, their parameters and their pre-requisites. Because the transition graph is not terribly easy to read, the tool also generates a Graphviz\(^3\) dot-file representing the same information.

For information on the options supported by `crm_simulate`, use its `--help` option.

**Interpreting the Graphviz output**

- Arrows indicate ordering dependencies
- Dashed arrows indicate dependencies that are not present in the transition graph
- Actions with a dashed border of any color do not form part of the transition graph
- Actions with a green border form part of the transition graph
- Actions with a red border are ones the cluster would like to execute but cannot run
- Actions with a blue border are ones the cluster does not feel need to be executed
- Actions with orange text are pseudo/pretend actions that the cluster uses to simplify the graph
- Actions with black text are sent to the LRM
- Resource actions have text of the form `rsc_action_interval node`
- Any action depending on an action with a red border will not be able to execute.

---

\(^3\) Graph visualization software. See [http://www.graphviz.org/](http://www.graphviz.org/) for details.
• Loops are really bad. Please report them to the development team.

4.3.1. Small Cluster Transition

In the above example, it appears that a new node, pcmk-2, has come online and that the cluster is checking to make sure rsc1, rsc2 and rsc3 are not already running there (Indicated by the rscN_monitor_0 entries). Once it did that, and assuming the resources were not active there, it would have liked to stop rsc1 and rsc2 on pcmk-1 and move them to pcmk-2. However, there appears to be some problem and the cluster cannot or is not permitted to perform the stop actions which implies it also cannot perform the start actions. For some reason the cluster does not want to start rsc3 anywhere.
4.3.2. Complex Cluster Transition

4.4. Do I Need to Update the Configuration on All Cluster Nodes?
No. Any changes are immediately synchronized to the other active members of the cluster.
To reduce bandwidth, the cluster only broadcasts the incremental updates that result from your changes and uses MD5 checksums to ensure that each copy is completely consistent.

4.5. Working with CIB Properties

Although these fields can be written to by the user, in most cases the cluster will overwrite any values specified by the user with the “correct” ones.

To change the ones that can be specified by the user, for example admin_epoch, one should use:

```
# cibadmin --modify --xml-text '<cib admin_epoch="42"/>'
```

A complete set of CIB properties will look something like this:

```xml
Example 4.6. Attributes set for a cib object

<cib crm_feature_set="3.0.7" validate-with="pacemaker-1.2"
     admin_epoch="42" epoch="116" num_updates="1"
     cib-last-written="Mon Jan 12 15:46:39 2015" update-origin="rhel7-1"
     update-client="crm_attribute" have-quorum="1" dc-uuid="1"/>
```

4.6. Querying and Setting Cluster Options

Cluster options can be queried and modified using the `crm_attribute` tool. To get the current value of cluster-delay, you can run:

```
# crm_attribute --query --name cluster-delay
```

which is more simply written as

```
# crm_attribute -G -n cluster-delay
```

If a value is found, you'll see a result like this:

```
# crm_attribute -G -n cluster-delay
scope=crm_config name=cluster-delay value=60s
```

If no value is found, the tool will display an error:

```
# crm_attribute -G -n clusta-deway
scope=crm_config name=clusta-deway value=(null)
Error performing operation: No such device or address
```

To use a different value (for example, 30 seconds), simply run:

```
# crm_attribute --name cluster-delay --update 30s
```

To go back to the cluster's default value, you can delete the value, for example:

```
# crm_attribute --name cluster-delay --delete
Deleted crm_config option: id=cib-bootstrap-options-cluster-delay name=cluster-delay
```
4.6.1. When Options are Listed More Than Once

If you ever see something like the following, it means that the option you’re modifying is present more than once.

Example 4.7. Deleting an option that is listed twice

```
# crm_attribute --name batch-limit --delete

Multiple attributes match name=batch-limit in crm_config:
Value: 50          (set=cib-bootstrap-options, id=cib-bootstrap-options-batch-limit)
Value: 100         (set=custom, id=custom-batch-limit)
Please choose from one of the matches above and supply the 'id' with --id
```

In such cases, follow the on-screen instructions to perform the requested action. To determine which value is currently being used by the cluster, refer to the *Rules* section of *Pacemaker Explained*.

**4.7. Connecting from a Remote Machine**

Provided Pacemaker is installed on a machine, it is possible to connect to the cluster even if the machine itself is not in the same cluster. To do this, one simply sets up a number of environment variables and runs the same commands as when working on a cluster node.

**Table 4.1. Environment Variables Used to Connect to Remote Instances of the CIB**

<table>
<thead>
<tr>
<th>Environment Variable</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIB_user</td>
<td>$USER</td>
<td>The user to connect as. Needs to be part of the haclient group on the target host.</td>
</tr>
<tr>
<td>CIB_passwd</td>
<td></td>
<td>The user’s password. Read from the command line if unset.</td>
</tr>
<tr>
<td>CIB_server</td>
<td>localhost</td>
<td>The host to contact</td>
</tr>
<tr>
<td>CIB_port</td>
<td></td>
<td>The port on which to contact the server; required.</td>
</tr>
<tr>
<td>CIB_encrypted</td>
<td>TRUE</td>
<td>Whether to encrypt network traffic</td>
</tr>
</tbody>
</table>

So, if c001n01 is an active cluster node and is listening on port 1234 for connections, and someuser is a member of the haclient group, then the following would prompt for someuser’s password and return the cluster’s current configuration:

```
# export CIB_port=1234; export CIB_server=c001n01; export CIB_user=someuser;
# cibadmin -Q
```

For security reasons, the cluster does not listen for remote connections by default. If you wish to allow remote access, you need to set the remote-tls-port (encrypted) or remote-clear-port (unencrypted) CIB properties (i.e., those kept in the cib tag, like num_updates and epoch).

**Table 4.2. Extra top-level CIB properties for remote access**

<table>
<thead>
<tr>
<th>Field</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>remote-tls-port</td>
<td>none</td>
<td>Listen for encrypted remote connections on this port.</td>
</tr>
</tbody>
</table>
Connecting from a Remote Machine

<table>
<thead>
<tr>
<th>Field</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>remote-clear-port</td>
<td>none</td>
<td>Listen for plaintext remote connections on this port.</td>
</tr>
</tbody>
</table>

**Important**

The Pacemaker version on the administration host must be the same or greater than the version(s) on the cluster nodes. Otherwise, it may not have the schema files necessary to validate the CIB.
Chapter 5.

Monitoring a Pacemaker Cluster

Table of Contents

5.1. Using crm_mon ......................................................... 21

5.1. Using crm_mon

The crm_mon utility displays the current state of an active cluster. It can show the cluster status organized by node or by resource, and can be used in either single-shot or dynamically updating mode. It can also display operations performed and information about failures.

Using this tool, you can examine the state of the cluster for irregularities, and see how it responds when you cause or simulate failures.

See the manual page or the output of crm_mon --help for a full description of its many options.

Example 5.1. Sample output from crm_mon -1

Stack: corosync
Current DC: node2 (version 2.0.0-1) - partition with quorum
Last updated: Mon Jan 29 12:18:42 2018
Last change: Mon Jan 29 12:18:40 2018 by root via crm_attribute on node3

5 nodes configured
2 resources configured

Online: [ node1 node2 node3 node4 node5 ]

Active resources:
Fencing (stonith:fence_xvm):    Started node1
IP      (ocf:heartbeat:IPaddr2):        Started node2

Example 5.2. Sample output from crm_mon -n -1

Stack: corosync
Current DC: node2 (version 2.0.0-1) - partition with quorum
Last updated: Mon Jan 29 12:21:48 2018
Last change: Mon Jan 29 12:18:40 2018 by root via crm_attribute on node3

5 nodes configured
2 resources configured

Node node1: online
    Fencing (stonith:fence_xvm):    Started
Node node2: online
    IP      (ocf:heartbeat:IPaddr2):        Started
Node node3: online
Node node4: online
Node node5: online

As mentioned in an earlier section, the DC is the node is where decisions are made. The cluster elects a node to be DC as needed. The only significance of the choice of DC to an administrator is the fact that its logs will have the most information about why decisions were made.
Troubleshooting Cluster Problems

Table of Contents

6.1. Logging .............................................................................................................................. 23
6.2. Transitions .......................................................................................................................... 23
6.3. Further Information About Troubleshooting ........................................................................... 24

6.1. Logging

Pacemaker by default logs messages of notice severity and higher to the system log, and messages of info severity and higher to the detail log, which by default is /var/log/pacemaker/pacemaker.log.

Logging options can be controlled via environment variables at Pacemaker start-up. Where these are set varies by operating system (often /etc/sysconfig/pacemaker or /etc/default/pacemaker).

Because cluster problems are often highly complex, involving multiple machines, cluster daemons, and managed services, Pacemaker logs rather verbosely to provide as much context as possible. It is an ongoing priority to make these logs more user-friendly, but by necessity there is a lot of obscure, low-level information that can make them difficult to follow.

The default log rotation configuration shipped with Pacemaker (typically installed in /etc/logrotate.d/pacemaker) rotates the log when it reaches 100MB in size, or weekly, whichever comes first.

If you configure debug or (Heaven forbid) trace-level logging, the logs can grow enormous quite quickly. Because rotated logs are by default named with the year, month, and day only, this can cause name collisions if your logs exceed 100MB in a single day. You can add dateformat -%Y%m%d-%H to the rotation configuration to avoid this.

6.2. Transitions

A key concept in understanding how a Pacemaker cluster functions is a transition. A transition is a set of actions that need to be taken to bring the cluster from its current state to the desired state (as expressed by the configuration).

Whenever a relevant event happens (a node joining or leaving the cluster, a resource failing, etc.), the controller will ask the scheduler to recalculate the status of the cluster, which generates a new transition. The controller then performs the actions in the transition in the proper order.

Each transition can be identified in the logs by a line like:

```
```

The file listed as the "inputs" is a snapshot of the cluster configuration and state at that moment (the CIB). This file can help determine why particular actions were scheduled. The crm_simulate command, described in Section 4.3, "Testing Your Configuration Changes", can be used to replay the file.
6.3. Further Information About Troubleshooting

Andrew Beekhof wrote a series of articles about troubleshooting in his blog, *The Cluster Guy*\(^1\):

- *Debugging Pacemaker*\(^2\)
- *Debugging the Policy Engine*\(^3\)
- *Pacemaker Logging*\(^4\)

The articles were written for an earlier version of Pacemaker, so many of the specific names and log messages to look for have changed, but the concepts are still valid.

---

\(^1\) [http://blog.clusterlabs.org/](http://blog.clusterlabs.org/)

\(^2\) [http://blog.clusterlabs.org/blog/2013/debugging-pacemaker](http://blog.clusterlabs.org/blog/2013/debugging-pacemaker)

\(^3\) [http://blog.clusterlabs.org/blog/2013/debugging-pengine](http://blog.clusterlabs.org/blog/2013/debugging-pengine)

\(^4\) [http://blog.clusterlabs.org/blog/2013/pacemaker-logging](http://blog.clusterlabs.org/blog/2013/pacemaker-logging)
Chapter 7.

Upgrading a Pacemaker Cluster

Table of Contents

7.1. Pacemaker Versioning ........................................................................................................ 25
7.2. Upgrading Cluster Software ................................................................................................. 26
  7.2.1. Complete Cluster Shutdown ...................................................................................... 26
  7.2.2. Rolling (node by node) ............................................................................................. 27
  7.2.3. Detach and Reattach ............................................................................................... 28
7.3. Upgrading the Configuration ............................................................................................... 28
7.4. What Changed in 2.0 .......................................................................................................... 31
7.5. What Changed in 1.0 .......................................................................................................... 31
  7.5.1. New .......................................................................................................................... 31
  7.5.2. Changed .................................................................................................................. 31
  7.5.3. Removed ................................................................................................................... 32

7.1. Pacemaker Versioning

Pacemaker has an overall release version, plus separate version numbers for certain internal components.

• **Pacemaker release version:** This version consists of three numbers (x.y.z).

  The major version number (the x in x.y.z) increases when at least some rolling upgrades are not possible from the previous major version. For example, a rolling upgrade from 1.0.8 to 1.1.15 should always be supported, but a rolling upgrade from 1.0.8 to 2.0.0 may not be possible.

  The minor version (the y in x.y.z) increases when there are significant changes in cluster default behavior, tool behavior, and/or the API interface (for software that utilizes Pacemaker libraries). The main benefit is to alert you to pay closer attention to the release notes, to see if you might be affected.

  The release counter (the z in x.y.z) is increased with all public releases of Pacemaker, which typically include both bug fixes and new features.

• **CRM feature set:** This version number applies to the communication between full cluster nodes, and is used to avoid problems in mixed-version clusters.

  The major version number increases when nodes with different versions would not work (rolling upgrades are not allowed). The minor version number increases when mixed-version clusters are allowed only during rolling upgrades. The minor-minor version number is ignored, but allows resource agents to detect cluster support for various features.¹

  Pacemaker ensures that the longest-running node is the cluster’s DC. This ensures new features are not enabled until all nodes are upgraded to support them.

• **LRMD protocol version:** This version applies to communication between a Pacemaker Remote node and the cluster. It increases when an older cluster node would have problems hosting the connection to a newer Pacemaker Remote node. To avoid these problems, Pacemaker Remote

¹ Before CRM feature set 3.1.0 (Pacemaker 2.0.0), the minor-minor version number was treated the same as the minor version.
nodes will accept connections only from cluster nodes with the same or newer LRMD protocol version.

Unlike with CRM feature set differences between full cluster nodes, mixed LRMD protocol versions between Pacemaker Remote nodes and full cluster nodes are fine, as long as the Pacemaker Remote nodes have the older version. This can be useful, for example, to host a legacy application in an older operating system version used as a Pacemaker Remote node.

- **XML schema version**: Pacemaker’s configuration syntax — what’s allowed in the Configuration Information Base (CIB) — has its own version. This allows the configuration syntax to evolve over time while still allowing clusters with older configurations to work without change.

### 7.2. Upgrading Cluster Software

There are three approaches to upgrading a cluster, each with advantages and disadvantages.

<table>
<thead>
<tr>
<th>Method</th>
<th>Available between all versions</th>
<th>Can be used with Pacemaker Remote nodes</th>
<th>Service outage during upgrade</th>
<th>Service recovery during upgrade</th>
<th>Exercises failover logic</th>
<th>Allows change of messaging layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete cluster shutdown</td>
<td>yes</td>
<td>yes</td>
<td>always</td>
<td>N/A</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Rolling (node by node)</td>
<td>no</td>
<td>yes</td>
<td>always</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Detach and reattach</td>
<td>yes</td>
<td>no</td>
<td>only due to failure</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>

1. Currently, Corosync version 2 and greater is the only supported cluster stack, but other stacks have been supported by past versions, and may be supported by future versions.

2. Any active resources will be moved off the node being upgraded, so there will be at least a brief outage unless all resources can be migrated "live".

#### 7.2.1. Complete Cluster Shutdown

In this scenario, one shuts down all cluster nodes and resources, then upgrades all the nodes before restarting the cluster.

1. On each node:
   a. Shutdown the cluster software (pacemaker and the messaging layer).
   b. Upgrade the Pacemaker software. This may also include upgrading the messaging layer and/or the underlying operating system.
   c. Check the configuration with the `crm_verify` tool.

2. On each node:
   a. Start the cluster software. Currently, only Corosync version 2 and greater is supported as the cluster layer, but if another stack is supported in the future, the stack does not need to be the same one before the upgrade.
One variation of this approach is to build a new cluster on new hosts. This allows the new version to be tested beforehand, and minimizes downtime by having the new nodes ready to be placed in production as soon as the old nodes are shut down.

### 7.2.2. Rolling (node by node)

In this scenario, each node is removed from the cluster, upgraded, and then brought back online, until all nodes are running the newest version.

**Special considerations when planning a rolling upgrade:**

- If you plan to upgrade other cluster software — such as the messaging layer — at the same time, consult that software’s documentation for its compatibility with a rolling upgrade.

- If the major version number is changing in the Pacemaker version you are upgrading to, a rolling upgrade may not be possible. Read the new version's release notes (as well the information here) for what limitations may exist.

- If the CRM feature set is changing in the Pacemaker version you are upgrading to, you should run a mixed-version cluster only during a small rolling upgrade window. If one of the older nodes drops out of the cluster for any reason, it will not be able to rejoin until it is upgraded.

- If the LRMD protocol version is changing, all cluster nodes should be upgraded before upgrading any Pacemaker Remote nodes.

See the ClusterLabs wiki's [Release Calendar](http://clusterlabs.org/wiki/ReleaseCalendar) to figure out whether the CRM feature set and/or LRMD protocol version changed between the the Pacemaker release versions in your rolling upgrade.

**To perform a rolling upgrade, on each node in turn:**

1. Put the node into standby mode, and wait for any active resources to be moved cleanly to another node. (This step is optional, but allows you to deal with any resource issues before the upgrade.)

2. Shutdown the cluster software (pacemaker and the messaging layer) on the node.

3. Upgrade the Pacemaker software. This may also include upgrading the messaging layer and/or the underlying operating system.

4. If this is the first node to be upgraded, check the configuration with the `crm_verify` tool.

5. Start the messaging layer. This must be the same messaging layer (currently only Corosync version 2 and greater is supported) that the rest of the cluster is using.

---

2 http://clusterlabs.org/wiki/ReleaseCalendar
Chapter 7. Upgrading a Pacemaker Cluster

Note

Even if a rolling upgrade from the current version of the cluster to the newest version is not directly possible, it may be possible to perform a rolling upgrade in multiple steps, by upgrading to an intermediate version first.

<table>
<thead>
<tr>
<th>Version being Installed</th>
<th>Oldest Compatible Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pacemaker 2.y.z</td>
<td>Pacemaker 1.1.11¹</td>
</tr>
<tr>
<td>Pacemaker 1.y.z</td>
<td>Pacemaker 1.0.0</td>
</tr>
<tr>
<td>Pacemaker 0.7.z</td>
<td>Pacemaker 0.6.z</td>
</tr>
</tbody>
</table>

¹Rolling upgrades from Pacemaker 1.1.z to 2.y.z are possible only if the cluster uses corosync version 2 or greater as its messaging layer, and the Cluster Information Base (CIB) uses schema 1.0 or higher in its validate-with property.

7.2.3. Detach and Reattach

The reattach method is a variant of a complete cluster shutdown, where the resources are left active and get re-detected when the cluster is restarted.

This method may not be used if the cluster contains any Pacemaker Remote nodes.

1. Tell the cluster to stop managing services. This is required to allow the services to remain active after the cluster shuts down.

   ```bash
   # crm_attribute --name maintenance-mode --update true
   ```

2. On each node, shutdown the cluster software (pacemaker and the messaging layer), and upgrade the Pacemaker software. This may also include upgrading the messaging layer. While the underlying operating system may be upgraded at the same time, that will be more likely to cause outages in the detached services (certainly, if a reboot is required).

3. Check the configuration with the `crm_verify` tool.

4. On each node, start the cluster software. Currently, only Corosync version 2 and greater is supported as the cluster layer, but if another stack is supported in the future, the stack does not need to be the same one before the upgrade.

5. Verify that the cluster re-detected all resources correctly.

6. Allow the cluster to resume managing resources again:

   ```bash
   # crm_attribute --name maintenance-mode --delete
   ```

7.3. Upgrading the Configuration

The CIB schema version can change from one Pacemaker version to another.
Upgrading the Configuration

After cluster software is upgraded, the cluster will continue to use the older schema version that it was previously using. This can be useful, for example, when administrators have written tools that modify the configuration, and are based on the older syntax. ³

However, when using an older syntax, new features may be unavailable, and there is a performance impact, since the cluster must do a non-persistent configuration upgrade before each transition. So while using the old syntax is possible, it is not advisable to continue using it indefinitely.

Even if you wish to continue using the old syntax, it is a good idea to follow the upgrade procedure outlined below, except for the last step, to ensure that the new software has no problems with your existing configuration (since it will perform much the same task internally).

If you are brave, it is sufficient simply to run `cibadmin --upgrade`.

A more cautious approach would proceed like this:

1. Create a shadow copy of the configuration. The later commands will automatically operate on this copy, rather than the live configuration.

   ```bash
   # crm_shadow --create shadow
   ```

2. Verify the configuration is valid with the new software (which may be stricter about syntax mistakes, or may have dropped support for deprecated features):

   ```bash
   # crm_verify --live-check
   ```

3. Fix any errors or warnings.

4. Perform the upgrade:

   ```bash
   # cibadmin --upgrade
   ```

5. If this step fails, there are three main possibilities:

   a. The configuration was not valid to start with (did you do steps 2 and 3?).
   b. The transformation failed - `report a bug`⁴ or `email the project`.⁵
   c. The transformation was successful but produced an invalid result.

      If the result of the transformation is invalid, you may see a number of errors from the validation library. If these are not helpful, visit the `Validation FAQ wiki page`⁶ and/or try the manual upgrade procedure described below.

6. Check the changes:

   ```bash
   # crm_shadow --diff
   ```

---

³ As of Pacemaker 2.0.0, only schema versions pacemaker-1.0 and higher are supported (excluding pacemaker-1.1, which was an experimental schema now known as pacemaker-next).
⁴ [http://bugs.clusterlabs.org/](http://bugs.clusterlabs.org/)
⁵ mailto:users@clusterlabs.org?subject=Transformation%20failed%20during%20upgrade
⁶ [http://clusterlabs.org/wiki/Validation_FAQ](http://clusterlabs.org/wiki/Validation_FAQ)
Chapter 7. Upgrading a Pacemaker Cluster

If at this point there is anything about the upgrade that you wish to fine-tune (for example, to change some of the automatic IDs), now is the time to do so:

```bash
# crm_shadow --edit
```

This will open the configuration in your favorite editor (whichever is specified by the standard $EDITOR environment variable).

7. Preview how the cluster will react:

```bash
# crm_simulate --live-check --save-dotfile shadow.dot -S
# dot -Tsvg shadow.dot -o shadow.svg
```

You can then view shadow.svg with any compatible image viewer or web browser. Verify that either no resource actions will occur or that you are happy with any that are scheduled. If the output contains actions you do not expect (possibly due to changes to the score calculations), you may need to make further manual changes. See Section 4.3, “Testing Your Configuration Changes” for further details on how to interpret the output of `crm_simulate` and `dot`.

8. Upload the changes:

```bash
# crm_shadow --commit shadow --force
```

In the unlikely event this step fails, please report a bug.

Note

It is also possible to perform the configuration upgrade steps manually:

1. Locate the `upgrade*.xsl` conversion scripts provided with the source code. These will often be installed in a location such as `/usr/share/pacemaker`, or may be obtained from the source repository.

2. Run the conversion scripts that apply to your older version, for example:

```bash
# xsltproc /path/to/upgrade06.xsl config06.xml > config10.xml
```

3. Locate the `pacemaker.rng` script (from the same location as the xsl files).

4. Check the XML validity:

```bash
# xmllint --relaxng /path/to/pacemaker.rng config10.xml
```

The advantage of this method is that it can be performed without the cluster running, and any validation errors are often more informative.

Note

It is also possible to perform the configuration upgrade steps manually:

1. Locate the `upgrade*.xsl` conversion scripts provided with the source code. These will often be installed in a location such as `/usr/share/pacemaker`, or may be obtained from the source repository.

2. Run the conversion scripts that apply to your older version, for example:

```bash
# xsltproc /path/to/upgrade06.xsl config06.xml > config10.xml
```

3. Locate the `pacemaker.rng` script (from the same location as the xsl files).

4. Check the XML validity:

```bash
# xmllint --relaxng /path/to/pacemaker.rng config10.xml
```

The advantage of this method is that it can be performed without the cluster running, and any validation errors are often more informative.

Note

It is also possible to perform the configuration upgrade steps manually:

1. Locate the `upgrade*.xsl` conversion scripts provided with the source code. These will often be installed in a location such as `/usr/share/pacemaker`, or may be obtained from the source repository.

2. Run the conversion scripts that apply to your older version, for example:

```bash
# xsltproc /path/to/upgrade06.xsl config06.xml > config10.xml
```

3. Locate the `pacemaker.rng` script (from the same location as the xsl files).

4. Check the XML validity:

```bash
# xmllint --relaxng /path/to/pacemaker.rng config10.xml
```

The advantage of this method is that it can be performed without the cluster running, and any validation errors are often more informative.
7.4. What Changed in 2.0
The main goal of the 2.0 release was to remove support for deprecated syntax, along with some small changes in default configuration behavior and tool behavior. Highlights:

• Only Corosync version 2 and greater is now supported as the underlying cluster layer. Support for Heartbeat and Corosync 1 (including CMAN) is removed.

• The Pacemaker detail log file is now stored in /var/log/pacemaker/pacemaker.log by default.

• The record-pending cluster property now defaults to true, which allows status tools such as crm_mon to show operations that are in progress.

• Support for a number of deprecated build options, environment variables, and configuration settings has been removed.

• The master tag has been deprecated in favor of using a clone tag with the new promotable meta-attribute set to true. “Master/slave” clone resources are now referred to as “promotable” clone resources, though it will take longer for the full terminology change to be completed.

• The public API for Pacemaker libraries that software applications can use has changed significantly.

For a detailed list of changes, see the release notes and the Pacemaker 2.0 Changes page on the ClusterLabs wiki.

7.5. What Changed in 1.0

7.5.1. New
• Failure timeouts.

• New section for resource and operation defaults.

• Tool for making offline configuration changes.

• Rules, instance_attributes, meta_attributes and sets of operations can be defined once and referenced in multiple places.

• The CIB now accepts XPath-based create/modify/delete operations. See the cibadmin help text.

• Multi-dimensional colocation and ordering constraints.

• The ability to connect to the CIB from non-cluster machines.

• Allow recurring actions to be triggered at known times.

7.5.2. Changed
• Syntax

• All resource and cluster options now use dashes (-) instead of underscores (_)

• master_slave was renamed to master

8 https://wiki.clusterlabs.org/wiki/Pacemaker_2.0_Changes
Chapter 7. Upgrading a Pacemaker Cluster

- The attributes container tag was removed
- The operation field pre-req has been renamed requires
- All operations must have an interval, start/stop must have it set to zero
- The stonith-enabled option now defaults to true.
- The cluster will refuse to start resources if stonith-enabled is true (or unset) and no STONITH resources have been defined
- The attributes of colocation and ordering constraints were renamed for clarity.
- resource-failure-stickiness has been replaced by migration-threshold.
- The parameters for command-line tools have been made consistent
- Switched to RelaxNG schema validation and libxml2 parser
  - id fields are now XML IDs which have the following limitations:
    - id's cannot contain colons (:)
    - id's cannot begin with a number
    - id's must be globally unique (not just unique for that tag)
  - 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.
- The CIB representation, from which a MD5 digest is calculated to verify CIBs on the nodes, has changed.
  This means that every CIB update will require a full refresh on any upgraded nodes until the cluster is fully upgraded to 1.0. This will result in significant performance degradation and it is therefore highly inadvisable to run a mixed 1.0/0.6 cluster for any longer than absolutely necessary.
- Ping node information no longer needs to be added to ha.cf.
  Simply include the lists of hosts in your ping resource(s).

7.5.3. Removed
- Syntax
  - It is no longer possible to set resource meta options as top-level attributes. Use meta attributes instead.
  - Resource and operation defaults are no longer read from crm_config.
### 8.1. Resource Agent Actions

If one resource depends on another resource via constraints, the cluster will interpret an expected result as sufficient to continue with dependent actions. This may cause timing issues if the resource agent start returns before the service is not only launched but fully ready to perform its function, or if the resource agent stop returns before the service has fully released all its claims on system resources. At a minimum, the start or stop should not return before a status command would return the expected (started or stopped) result.

### 8.2. OCF Resource Agents

#### 8.2.1. Location of Custom Scripts

OCF Resource Agents are found in `/usr/lib/ocf/resource.d/provider`

When creating your own agents, you are encouraged to create a new directory under `/usr/lib/ocf/resource.d/` so that they are not confused with (or overwritten by) the agents shipped by existing providers.

So, for example, if you choose the provider name of bigCorp and want a new resource named bigApp, you would create a resource agent called `/usr/lib/ocf/resource.d/bigCorp/bigApp` and define a resource:

```xml
<primitive id="custom-app" class="ocf" provider="bigCorp" type="bigApp"/>
```

#### 8.2.2. Actions

All OCF resource agents are required to implement the following actions.

<table>
<thead>
<tr>
<th>Action</th>
<th>Description</th>
<th>Instructions</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>start</code></td>
<td>Start the resource</td>
<td>Return 0 on success and an appropriate error code otherwise. Must not report success until the resource is fully active.</td>
</tr>
<tr>
<td><code>stop</code></td>
<td>Stop the resource</td>
<td>Return 0 on success and an appropriate error code otherwise. Must not report success until the resource is fully stopped.</td>
</tr>
</tbody>
</table>
Chapter 8. Resource Agents

<table>
<thead>
<tr>
<th>Action</th>
<th>Description</th>
<th>Instructions</th>
</tr>
</thead>
</table>
| monitor  | Check the resource's state   | Exit 0 if the resource is running, 7 if it is stopped, and anything else if it is failed.  
|          |                              | NOTE: The monitor script should test the state of the resource on the local machine only. |
| meta-data| Describe the resource        | Provide information about this resource as an XML snippet. Exit with 0.       |
|          |                              | NOTE: This is not performed as root.                                         |
| validate-all| Verify the supplied parameters | Return 0 if parameters are valid, 2 if not valid, and 6 if resource is not configured. |

Additional requirements (not part of the OCF specification) are placed on agents that will be used for advanced concepts such as clone resources.

Table 8.2. Optional Actions for OCF Resource Agents

<table>
<thead>
<tr>
<th>Action</th>
<th>Description</th>
<th>Instructions</th>
</tr>
</thead>
<tbody>
<tr>
<td>promote</td>
<td>Promote the local instance of a promotable clone resource to the master (primary) state.</td>
<td>Return 0 on success</td>
</tr>
<tr>
<td>demote</td>
<td>Demote the local instance of a promotable clone resource to the slave (secondary) state.</td>
<td>Return 0 on success</td>
</tr>
<tr>
<td>notify</td>
<td>Used by the cluster to send the agent pre- and post-notification events telling the resource what has happened and will happen.</td>
<td>Must not fail. Must exit with 0</td>
</tr>
</tbody>
</table>

One action specified in the OCF specs, recover, is not currently used by the cluster. It is intended to be a variant of the start action that tries to recover a resource locally.

**Important**

If you create a new OCF resource agent, use ocf-tester to verify that the agent complies with the OCF standard properly.

8.2.3. How are OCF Return Codes Interpreted?

The first thing the cluster does is to check 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 recovery action is initiated.

There are three types of failure recovery:

Table 8.3. Types of recovery performed by the cluster

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Action Taken by the Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>soft</td>
<td>A transient error occurred</td>
<td>Restart the resource or move it to a new location</td>
</tr>
<tr>
<td>hard</td>
<td>A non-transient error that may be specific to the current node occurred</td>
<td>Move the resource elsewhere and prevent it from being retried on the current node</td>
</tr>
</tbody>
</table>
### OCF Return Codes

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Action Taken by the Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>fatal</td>
<td>A non-transient error that will be common to all cluster nodes (e.g. a bad configuration was specified)</td>
<td>Stop the resource and prevent it from being started on any cluster node.</td>
</tr>
</tbody>
</table>

#### 8.2.4. OCF Return Codes

The following table outlines the different OCF return codes and the type of recovery the cluster will initiate when a failure code is received. Although counterintuitive, even actions that return 0 (aka. OCF_SUCCESS) can be considered to have failed, if 0 was not the expected return value.

Table 8.4. OCF Return Codes and their Recovery Types

<table>
<thead>
<tr>
<th>RC</th>
<th>OCF Alias</th>
<th>Description</th>
<th>RT</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>OCF_SUCCESS</td>
<td>Success. The command completed successfully. This is the expected result for all start, stop, promote and demote commands.</td>
<td>soft</td>
</tr>
<tr>
<td>1</td>
<td>OCF_ERR_GENERIC</td>
<td>Generic &quot;there was a problem&quot; error code.</td>
<td>soft</td>
</tr>
<tr>
<td>2</td>
<td>OCF_ERR_ARGS</td>
<td>The resource’s configuration is not valid on this machine. E.g. it refers to a location not found on the node.</td>
<td>hard</td>
</tr>
<tr>
<td>3</td>
<td>OCF_ERR_UNIMPLEMENTED</td>
<td>The requested action is not implemented.</td>
<td>hard</td>
</tr>
<tr>
<td>4</td>
<td>OCF_ERR_PERM</td>
<td>The resource agent does not have sufficient privileges to complete the task.</td>
<td>hard</td>
</tr>
<tr>
<td>5</td>
<td>OCF_ERR_INSTALLED</td>
<td>The tools required by the resource are not installed on this machine.</td>
<td>hard</td>
</tr>
<tr>
<td>6</td>
<td>OCF_ERR_CONFIGURED</td>
<td>The resource’s configuration is invalid. E.g. required parameters are missing.</td>
<td>fatal</td>
</tr>
<tr>
<td>7</td>
<td>OCF_NOT_RUNNING</td>
<td>The resource is safely stopped. The cluster will not attempt to stop a resource that returns this for any action.</td>
<td>N/A</td>
</tr>
<tr>
<td>8</td>
<td>OCF_RUNNING_MASTER</td>
<td>The resource is running in master mode.</td>
<td>soft</td>
</tr>
<tr>
<td>9</td>
<td>OCF_FAILED_MASTER</td>
<td>The resource is in master mode but has failed. The resource will be demoted, stopped and then started (and possibly promoted) again.</td>
<td>soft</td>
</tr>
<tr>
<td>other</td>
<td>N/A</td>
<td>Custom error code.</td>
<td>soft</td>
</tr>
</tbody>
</table>

Exceptions to the recovery handling described above:

- Probes (non-recurring monitor actions) that find a resource active (or in master mode) will not result in recovery action unless it is also found active elsewhere.

- The recovery action taken when a resource is found active more than once is determined by the resource’s multiple-active property.

- Recurring actions that return OCF_ERR_UNIMPLEMENTED do not cause any type of recovery.

---

1 [http://refspects.linuxfoundation.org/lsb.shtml](http://refspects.linuxfoundation.org/lsb.shtml)
8.3. LSB Resource Agents (Init Scripts)

8.3.1. LSB Compliance

The relevant part of the LSB specifications\(^1\) includes a description of all the return codes listed here.

Assuming `some_service` is configured correctly and currently inactive, the following sequence will help you determine if it is LSB-compatible:

1. Start (stopped):

   ```
   # /etc/init.d/some_service start ; echo "result: $?"
   ```
   a. Did the service start?
   b. Did the echo command print `result: 0` (in addition to the init script's usual output)?

2. Status (running):

   ```
   # /etc/init.d/some_service status ; echo "result: $?"
   ```
   a. Did the script accept the command?
   b. Did the script indicate the service was running?
   c. Did the echo command print `result: 0` (in addition to the init script's usual output)?

3. Start (running):

   ```
   # /etc/init.d/some_service start ; echo "result: $?"
   ```
   a. Is the service still running?
   b. Did the echo command print `result: 0` (in addition to the init script's usual output)?

4. Stop (running):

   ```
   # /etc/init.d/some_service stop ; echo "result: $?"
   ```
   a. Was the service stopped?
   b. Did the echo command print `result: 0` (in addition to the init script's usual output)?

5. Status (stopped):

   ```
   # /etc/init.d/some_service status ; echo "result: $?"
   ```
   a. Did the script accept the command?
   b. Did the script indicate the service was not running?
   c. Did the echo command print `result: 3` (in addition to the init script's usual output)?

6. Stop (stopped):

---
\(^1\) LSB specifications.
a. Is the service still stopped?

b. Did the echo command print **result: 0** (in addition to the init script’s usual output)?

7. Status (failed):

   a. This step is not readily testable and relies on manual inspection of the script.

   The script can use one of the error codes (other than 3) listed in the LSB spec to indicate that it is active but failed. This tells the cluster that before moving the resource to another node, it needs to stop it on the existing one first.

   If the answer to any of the above questions is no, then the script is not LSB-compliant. Your options are then to either fix the script or write an OCF agent based on the existing script.
Appendix A. Revision History

Revision 1-0  Tue Jan 23 2018  Ken Gaillot  kgaillot@redhat.com
Move administration-oriented information from Pacemaker Explained into its own book

Revision 1-1  Mon Jan 28 2019  Ken Gaillot  kgaillot@redhat.com, Jan Pokorný  jpokorny@redhat.com
Add “Troubleshooting” chapter, minor clarifications and reformatting
Index

Symbols
0
   OCF_SUCCESS, 35
1
   OCF_ERR_GENERIC, 35
2
   OCF_ERR_ARGS, 35
3
   OCF_ERR_UNIMPLEMENTED, 35
4
   OCF_ERR_PERM, 35
5
   OCF_ERR_INSTALLED, 35
6
   OCF_ERR_CONFIGURED, 35
7
   OCF_NOT_RUNNING, 35
8
   OCF_RUNNING_MASTER, 35
9
   OCF_FAILED_MASTER, 35

A
Action
demote, 34
meta-data, 34
monitor, 34
notify, 34
promote, 34
start, 33
stop, 33
validate-all, 34
Add Cluster Node, 9
   Corosync, 9

C
Changing cluster stack, 26
CIB_encrypted, 18
CIB_passwd, 18
CIB_port, 18
CIB_server, 18
CIB_user, 18
Cluster
   Querying Options, 17
   Remote administration, 18
   Remote connection, 18
   Setting Options, 17
   switching between stacks, 26
Cluster Option, 17, 17
Configuration, 28, 29
   upgrade manually, 30
   upgrading, 28
   validate XML, 30
   verify, 29
   convert, 30
   Corosync, 9, 9, 10
      Add Cluster Node, 9
      Remove Cluster Node, 9
      Replace Cluster Node, 10

D
demote, 34
   OCF Action, 34

E
Environment Variable
   CIB_encrypted, 18
   CIB_passwd, 18
   CIB_port, 18
   CIB_server, 18
   CIB_user, 18
error
   fatal, 35
   hard, 34
   soft, 34

F
fatal, 35
   OCF error, 35
feedback
   contact information for this manual, xiii

H
hard, 34
   OCF error, 34

M
meta-data, 34
   OCF Action, 34
monitor, 34
   OCF Action, 34

N
notify, 34
   OCF Action, 34

O
OCF
   Action
      demote, 34
      meta-data, 34
      monitor, 34
      notify, 34
      promote, 34
Index

start, 33
stop, 33
validate-all, 34
error
fatal, 35
hard, 34
soft, 34
OCF Action, 33, 33, 34, 33, 34, 34, 34, 34, 34, 34
OCF error, 34, 34, 35
OCF Resource Agents, 33
ocf-tester, 34
OCF_ERR_ARGS, 35, 35
OCF_ERR_CONFIGURED, 35, 35
OCF_ERR_GENERIC, 35, 35
OCF_ERR_INSTALLED, 35, 35
OCF_ERR_PERM, 35, 35
OCF_ERR_UNIMPLEMENTED, 35, 35
OCF_FAILED_MASTER, 35, 35
OCF_NOT_RUNNING, 35, 35
OCF_RUNNING_MASTER, 35, 35
OCF_SUCCESS, 35, 35
Option
remote-clear-port, 19
remote-tls-port, 18
other, 35
P
promote, 34
OCF Action, 34
Q
Querying
Cluster Option, 17
Querying Options, 17
R
reattach, 26
reattach upgrade, 26
Remote administration, 18
Remote connection, 18
Remote Connection
Option
remote-clear-port, 19
remote-tls-port, 18
Remote Connection Option, 18, 19
remote-clear-port, 19
Remote Connection Option, 19
remote-tls-port, 18
Remote Connection Option, 18
Remove Cluster Node, 9
Corosync, 9
Replace Cluster Node, 10
Corosync, 10

Return Code
0
OCF_SUCCESS, 35
1
OCF_ERR_GENERIC, 35
2
OCF_ERR_ARGS, 35
3
OCF_ERR_UNIMPLEMENTED, 35
4
OCF_ERR_PERM, 35
5
OCF_ERR_INSTALLED, 35
6
OCF_ERR_CONFIGURED, 35
7
OCF_NOT_RUNNING, 35
8
OCF_RUNNING_MASTER, 35
9
OCF_FAILED_MASTER, 35
OCF_ERR_ARGS, 35
OCF_ERR_CONFIGURED, 35
OCF_ERR_GENERIC, 35
OCF_ERR_INSTALLED, 35
OCF_ERR_PERM, 35
OCF_ERR_UNIMPLEMENTED, 35
OCF_FAILED_MASTER, 35
OCF_NOT_RUNNING, 35
OCF_RUNNING_MASTER, 35
OCF_SUCCESS, 35
other, 35
S
Setting
Cluster Option, 17
Setting Options, 17
shutdown, 26
shutdown upgrade, 26
soft, 34
OCF error, 34
start, 33
OCF Action, 33
stop, 33
OCF Action, 33
switching between stacks, 26
U
upgrade
Configuration, 28
reattach, 26
rolling, 26
  shutdown, 26
upgrade manually, 30
upgrading, 28

V
validate configuration, 30
validate XML, 30
validate-all, 34
  OCF Action, 34
verify, 29
  Configuration, 29

X
XML
  convert, 30