

SUSE Linux Enterprise High Availability Extension

11 SP1

www.novell.com

2010 年 4 月 22 日

High Availability 指南



High Availability 指南

版权所有 © 2006- 2010 Novell, Inc.

根据自由软件基金会 (Free Software Foundation) 发布的 GNU 自由文档许可证 (GNU Free Documentation License) 版本 1.2 或更高版本中的条款，在此授予您复制、分发和/或修订本文档的许可权限；本版权声明和许可证附带不可变部分。标题为“GNU 自由文档许可证”的章节中包含许可证副本。

SUSE®、openSUSE®、openSUSE® 徽标、Novell®、Novell® 徽标和 N® 徽标是 Novell, Inc. 在美国和其他国家或地区的注册商标。Linux* 是 Linus Torvalds 的注册商标。所有第三方商标均属其各自所有者的财产。商标符号（®、™ 等）代表 Novell 的商标；星号 (*) 表示第三方商标。

本指南力求涵盖所有细节。但这并不确保本指南准确无误。无论是 Novell, Inc.、SUSE LINUX 产品 GmbH、作者还是翻译人员都不对任何可能的错误或因错误造成的任何后果负责。

目录

关于本指南	ix
部分 I 安装和设置	1
1 产品概述	3
1.1 主要特征	3
1.2 优势	6
1.3 群集配置：储存	9
1.4 体系结构	11
2 入门	15
2.1 硬件要求	15
2.2 软件要求	16
2.3 共享磁盘系统需求	16
2.4 准备工作	16
2.5 概述：安装和设置群集	17
3 用 YaST 进行安装和基本设置	19
3.1 安装 High Availability Extension	19
3.2 初始群集设置	20
3.3 使群集联机	27
3.4 使用 AutoYaST 进行大批量部署	28

部分 II 配置和管理	31
4 配置和管理基础	33
4.1 全局群集选项	33
4.2 群集资源	35
4.3 资源监控	45
4.4 资源约束	45
4.5 更多信息	50
5 配置和管理群集资源 (GUI)	53
5.1 Pacemaker GUI - 概述	54
5.2 配置全局群集选项	56
5.3 配置群集资源	57
5.4 管理群集资源	76
6 配置和管理群集资源 (命令行)	83
6.1 crm 命令行工具 - 概述	83
6.2 配置全局群集选项	89
6.3 配置群集资源	90
6.4 管理群集资源	101
7 使用 Web 界面管理群集资源	105
7.1 启动 HA Web Konsole 并登录	105
7.2 使用 HA Web Konsole	106
7.3 查错	108
8 添加或修改资源代理	109
8.1 STONITH 代理	109
8.2 写入 OCF 资源代理	110
8.3 OCF 返回码和故障恢复	111
9 屏障和 STONITH	113
9.1 屏障分类	113
9.2 节点级别屏障	114
9.3 STONITH 配置	115
9.4 监视屏障设备	120
9.5 特殊的屏障设备	120
9.6 更多信息	122

10 使用 Linux Virtual Server 进行负载均衡	123
10.1 概念概述	123
10.2 使用 YaST 配置 IP 负载均衡	125
10.3 其他设置	131
10.4 更多信息	131
11 网络设备联接	133
11.1 使用 YaST 配置绑定设备	133
11.2 更多信息	135
部分 III 储存和数据复制	137
12 Oracle Cluster File System 2	139
12.1 功能和优点	139
12.2 OCFS2 包和管理实用程序	140
12.3 配置 OCFS2 服务	141
12.4 创建 OCFS2 卷	143
12.5 装入 OCFS2 卷	145
12.6 更多信息	146
13 分布式复制块设备 (DRBD)	147
13.1 概念概述	147
13.2 安装 DRBD 服务	149
13.3 配置 DRBD 服务	150
13.4 测试 DRBD 服务	153
13.5 调整 DRBD	155
13.6 DRBD 查错	155
13.7 更多信息	157
14 群集 LVM	159
14.1 概念概述	159
14.2 cLVM 配置	159
14.3 显式配置合格的 LVM2 设备	167
14.4 更多信息	168
15 储存保护	169
15.1 基于储存区的屏蔽	169
15.2 确保储存区的排它激活	174

16 Samba 群集	177
16.1 概念概述	177
16.2 基本配置	178
16.3 调试和测试群集 Samba	180
16.4 更多信息	182
 部分 IV 查错和参考	 183
 17 查错	 185
17.1 安装问题	185
17.2 “调试” HA 群集	186
17.3 常见问题解答	187
17.4 更多信息	189
 18 群集管理工具	 191
 19 HA OCF Agents	 243
 部分 V 附录	 335
 A 设置简单测试资源的示例	 337
A.1 使用 GUI 配置资源	337
A.2 手动配置资源	339
 B 将群集升级为最新产品版本	 341
B.1 从 SLES 10 升级到 SLEHA 11	341
B.2 从 SLEHA 11 升级到 SLEHA 11 SP1	345
 C 新功能?	 347
C.1 版本 10 SP3 到版本 11	347
C.2 版本 11 到版本 11 SP1	350
 D GNU 许可证	 353
D.1 GNU General Public License	353
D.2 GNU Free Documentation License	356

关于本指南

SUSE® Linux Enterprise High Availability Extension 是一个开放源代码群集技术的集成套件，使您能够实现高度可用的物理和虚拟 Linux 群集。为快速且有效地进行配置和管理，High Availability Extension 包括了一个图形用户界面 (GUI) 和一个命令行界面 (CLI)。此外，它还附带了 HA Web Konsole，用于通过 Web 界面管理 Linux 群集。

本指南适用于需要设置、配置和维护 High Availability (HA) 群集的管理员。它详细介绍了这两种界面（GUI 和 CLI），以帮助管理员选择与执行关键任务的需要匹配的相应工具。

本指南分为以下部分：

安装和设置

开始安装和配置群集之前，先熟悉群集的基础知识和体系结构，了解其关键功能和优点的概况。执行下一步之前，要先了解必须满足哪些软硬件要求以及做好哪些准备工作。使用 YaST 执行 HA 群集的安装和基本设置。

配置和管理

使用图形用户界面 (Pacemaker GUI) 或 `crm` 命令行界面添加、配置和管理资源。如果希望或需要通过 Web 界面监视群集，请使用 HA Web Konsole。了解如何利用负载平衡和屏障。如果您考虑编写自己的资源代理或修改现有的资源代理，请了解一些有关如何创建不同类型的资源代理的背景信息。

储存和数据复制

SUSE Linux Enterprise High Availability Extension 还配备群集感知文件系统和卷管理器：Oracle 群集文件系统 (OCFS2) 和群集逻辑卷管理器 (cLVM)。对于数据复制，请使用 DRBD（分布式复制块设备）将 High Availability 服务的数据从群集的活动节点镜像到其备用节点。此外，群集 Samba 服务器还可为异构环境提供 High Availability 解决方案。

查错和参考

管理群集需要执行一些查错操作。了解最常见的故障及解决方法。了解 High Availability Extension 提供的命令行工具的全面参考以便管理群集。

附录

列出上次发布以来 High Availability Extension 的新增功能和行为变化。了解如何将群集迁移到最近的发行版，并找到设置简单测试资源的示例。

本手册中的许多章节包含到附加文档资源的链接。其中包括系统上提供的附加文档以及因特网上提供的文档。

有关该产品可用文档的概述和最新文档更新，请参见 <http://www.novell.com/documentation>。

1 反馈

提供了多种反馈渠道：

Bug 和增强请求

有关产品可用的服务和支持选项，请参阅 <http://www.novell.com/services/>。

要报告产品组件的 bug，请使用 <http://support.novell.com/additional/bugreport.html>。

提交增强请求：<https://secure-www.novell.com/rms/rmsTool?action=ReqActions.viewAddPage&return=www>。

用户意见

欢迎您对本手册和本产品中包含的其他文档提出意见和建议。请使用联机文档每页底部的“用户注释”功能或转到 <http://www.novell.com/documentation/feedback.html> 并在此处输入注释。

2 文档约定

以下是本手册中使用的版式约定：

- `/etc/passwd`：目录名称和文件名
- `placeholder`：将 `placeholder` 替换为实际值
- `PATH`：环境变量 `PATH`
- `ls`、`--help`：命令、选项和参数

- `user`: 用户和组
- **Alt**、**Alt + F1**: 按键或组合键；这些键以大写形式显示，如在键盘上一样
- 文件、文件 > 另存为: 菜单项，按钮
- ▶ **amd64 em64t**: 本段只与指定的体系结构有关。箭头标记文本块的开始位置和结束位置。 ◀
- 跳舞的企鹅（企鹅一章，↑其他手册）：这是对其他手册中的某章的参考。

部分 I. 安装和设置

产品概述

SUSE® Linux Enterprise High Availability Extension 是一个开放源代码群集技术的集成套件，使您能够实现高度可用的物理和虚拟 Linux 群集，并排除单一故障点。它可确保关键网络资源的高可用性和可管理性，这些网络资源包括数据、应用程序和服务。因此，它有助于维持业务连续性、保护数据完整性及减少 Linux 关键任务工作负荷的计划外停机时间。

它随附提供必需的监视、消息交换和群集资源管理功能（支持对独立管理的群集资源进行故障转移、故障回复和迁移（负载平衡））。High Availability Extension 作为 SUSE Linux Enterprise Server 11 SP1 的外接式附件提供。

本章介绍 High Availability Extension 的主要产品功能和优点。您将在本章中找到多个示例群集并了解组成群集的组件。最后一节概述了体系结构，描述了群集内的各体系结构层和进程。

有关 High Availability 群集环境中使用的一些通用术语的解释，请参阅术语（第 361 页）。

1.1 主要特征

SUSE® Linux Enterprise High Availability Extension 有助于确保和管理网络资源的可用性。以下各节重点说明一些关键功能：

1.1.1 各种群集方案

High Availability Extension 支持以下方案：

- 主动/主动配置
- 主动/被动配置：N+1、N+M、N 到 1 和 N 到 M
- 混合物理和虚拟群集，支持将虚拟服务器和物理服务器群集在一起。这可提高服务可用性和资源利用率。

群集最多可包含 16 个 Linux 服务器。如果群集内的一台服务器发生故障，则群集内的任何其他服务器均可重启动此服务器上的资源（应用程序、服务、IP 地址和文件系统）。

1.1.2 灵活性

High Availability Extension 附带了 Corosync/OpenAIS 消息交换和成员资格层以及 Pacemaker 群集资源管理器。使用 Pacemaker，管理员可以持续监视其资源的运行状况和状态、管理依赖性以及根据高度可配置的规则和策略自动停止和启动服务。High Availability Extension 允许您将群集调整为特定的应用程序和硬件基础结构，以适应您的组织。基于时间的配置使服务可以在特定时间自动迁移回已修复的节点。

1.1.3 储存和数据复制

使用 High Availability Extension 可以根据需要动态地指派和重指派服务器储存。它支持光纤通道或 iSCSI 储存区域网络 (SAN)。它还支持共享磁盘系统，但这不是必需的。SUSE Linux Enterprise High Availability Extension 还配备群集感知文件系统和卷管理器：Oracle 群集文件系统 (OCFS2) 和群集逻辑卷管理器 (cLVM)。若要复制数据，可使用 DRBD（分布式复制块设备）将 High Availability 服务的数据从群集的主动节点镜像到其备用节点。此外，SUSE Linux Enterprise High Availability Extension 还支持 CTDB (Clustered Trivial Database)，一种 Samba 群集技术。

1.1.4 支持虚拟环境

SUSE Linux Enterprise High Availability Extension 支持物理和虚拟 Linux 服务器的混合群集。SUSE Linux Enterprise Server 11 SP1 附带了 Xen（开放源代码虚拟化超级管理程序）和 KVM（基于内核的虚拟机，是适用于 Linux 的基于硬件虚拟化扩展的虚拟化软件）。High Availability Extension 中的群集资源管理器能够识别、监视和管理正在虚拟服务器上运行的服务以及正在物理服务器上运行的服务。Guest 系统可作为服务由群集管理。

1.1.5 资源代理

SUSE Linux Enterprise High Availability Extension 包含大量资源代理来管理资源，如 Apache、IPv4、IPv6 等。它还为通用的第三方应用程序（例如 IBM WebSphere Application Server）提供了资源代理。有关产品包含的开放群集框架 (OCF) 资源代理的列表，请参见第 19 章 *HA OCF Agents*（第 243 页）。

1.1.6 用户友好的管理工具

High Availability Extension 附带了一组功能强大的工具，可用于群集的基本安装和设置及其有效配置和管理：

YaST

常规系统安装和管理的图形用户界面。按第 3.1 节“安装 High Availability Extension”（第 19 页）中所述使用它在 SUSE Linux Enterprise Server 上安装 High Availability Extension。YaST 在 High Availability 类别中还包含以下模块，可帮助您配置群集或各个组件：

- 群集：基本群集设置。有关详细信息，请参见第 3.2 节“初始群集设置”（第 20 页）。
- DRBD：配置分布式复制块设备。
- IP 负载均衡：使用 Linux Virtual Server 配置负载平衡。有关详细信息，请参见第 10 章 *使用 Linux Virtual Server 进行负载平衡*（第 123 页）。

Pacemaker GUI

可安装的图形用户界面，用于轻松配置和管理群集。指引您完成资源的创建和配置，并可用于执行启动、停止或迁移资源之类的管理任务。有关详细信息，请参见第 5 章 *配置和管理群集资源 (GUI)*（第 53 页）。

HA Web Konsole

基于 Web 的用户界面，使用此界面可从非 Linux 计算机管理 Linux 群集。如果系统未提供图形用户界面，它还是理想的解决方案。有关详细信息，请参见第 7 章 *使用 Web 界面管理群集资源*（第 105 页）。

crm

功能强大的统一命令行界面。可帮助您配置资源和执行所有监视或管理任务。有关详细信息，请参见第 6 章 *配置和管理群集资源（命令行）*（第 83 页）。

1.2 优势

High Availability Extension 允许将最多 16 台 Linux 服务器配置为一个高度可用的群集（HA 群集），在群集中可以将资源动态地切换或移动到任何服务器上。可以将资源配置为自动迁移以防服务器故障，或手动移动资源以对硬件查错或平衡工作负荷。

High Availability Extension 通过商品组件提供了高度可用性。通过将应用程序和操作合并到群集中降低了成本。High Availability Extension 还允许集中管理整个群集并调整资源以满足变化的工作负荷要求（这样就手动地使群集“负载平衡”了）。允许群集的多个（两个以上）节点共享一个“热备份”也节约了成本。

一个同样重要的好处是潜在地减少了计划外服务中断及用于软件和硬件维护和升级的计划内中断。

实施群集的理由包括：

- 提高可用性
- 改善性能
- 降低操作成本
- 可伸缩性

- 灾难恢复
- 数据保护
- 服务器合并
- 储存合并

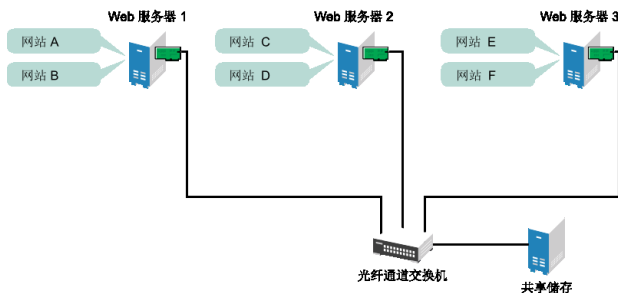
通过在共享磁盘子系统上实施 RAID 可获得共享磁盘容错。

以下方案说明了 High Availability Extension 提供的一些好处。

示例群集方案

假设您配置了一个包含三台服务器的群集，并在群集内的每台服务器上安装了 Web 服务器。群集内的每台服务器都主管两个 Web 站点。每个 Web 站点的全部数据、图形和 Web 页面内容都储存在一个连接到群集中每台服务器的共享磁盘子系统上。下图说明了该系统的结构。

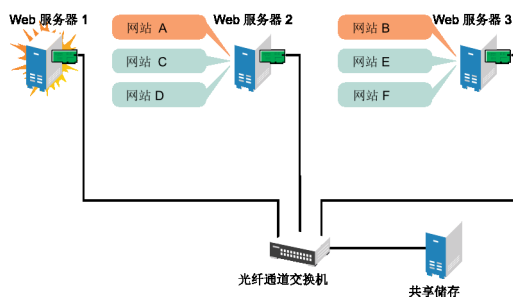
图 1.1 三台服务器的群集



在群集的正常工作状态下，每台服务器都与群集内的其它服务器持续通讯，并对所有已注册的资源进行定期巡回检测以检测故障。

假设 Web 服务器 1 出现硬件或软件故障，而依赖此 Web 服务器访问因特网、收发电子邮件和获取信息的用户失去了连接。下图说明了当万维网服务器 1 出现故障时，资源的移动情况。

图 1.2 三台服务器的群集（其中一台服务器出现故障后）



Web 站点 A 移至 Web 服务器 2，Web 站点 B 移至 Web 服务器 3。IP 地址和证书也移至 Web 服务器 2 和 Web 服务器 3。

在配置群集时，您决定了在出现故障的情况下，每台 Web 服务器上的 Web 站点将移至哪里。在上例中，您已配置将 Web 站点 A 移至 Web 服务器 2，将 Web 站点 B 移至 Web 服务器 3。这样，曾由 Web 服务器 1 处理的工作负荷继续存在且平均分配给剩余的群集成员。

当 Web 服务器 1 出现故障，High Availability Extension 软件 会执行以下操作：

- 检测到故障，并与 STONITH 确认 Web 服务器 1 确实已出现故障。STONITH 是“Shoot The Other Node In The Head”的缩写，是一种关闭行为异常节点的方式，以防止这些节点在群集中导致问题。
- 将以前安装在 Web 服务器 1 上的共享数据目录重新安装在 Web 服务器 2 和 Web 服务器 3 上。
- 在 Web 服务器 2 和 Web 服务器 3 上重新启动以前运行于 Web 服务器 1 上的应用程序。
- 将 IP 地址传送到 Web 服务器 2 和 Web 服务器 3。

在此示例中，故障转移过程迅速完成，用户在几秒钟之内就可以重新访问 Web 站点信息，而且在多数情况下无需重新登录。

现在，假设 Web 服务器 1 的故障已解决，它已恢复到正常工作状态。Web 站点 A 和 Web 站点 B 可以自动故障回复（移回）至 Web 服务器 1，或者留在当前所在的服务器上。这取决于您是如何配置它们的资源的。将服务迁移回 Web 服务

器 1 将导致一段中断期，因此 High Availability Extension 也允许您将迁移推迟到某个将极少或不会造成服务中断的时段。这两种选择都各有优缺点。

High Availability Extension 也提供了资源迁移功能。可以根据系统管理的需要将应用程序、Web 站点等资源移动到群集中的其他服务器。

例如，您可以手动将 Web 站点 A 或 Web 站点 B 从 Web 服务器 1 移至群集内的其他任何一台服务器。对万维网服务器 1 进行升级或定期维护时，或者只是要提高万维网站点的性能或可访问性，都需要执行此操作。

1.3 群集配置：储存

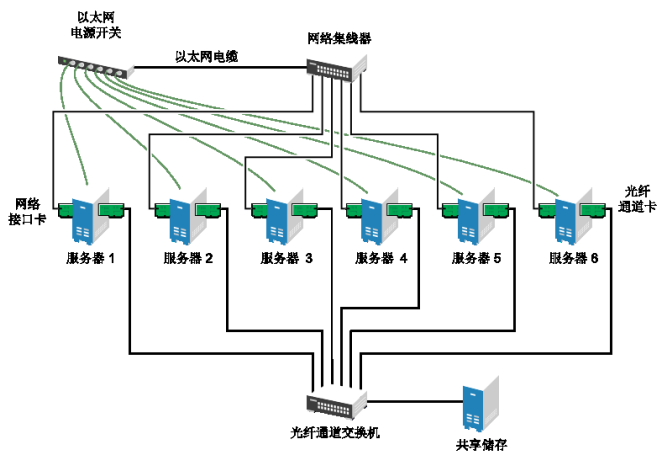
High Availability Extension 的群集配置可能包括或不包括共享磁盘子系统。共享磁盘子系统可通过高速光纤通道卡、电缆和交换机连接，也可配置为使用 iSCSI。如果服务器出现故障，群集中的另一个指定服务器将自动装入之前在故障服务器上装入的共享磁盘目录。这样，网络用户就能继续访问共享磁盘子系统上的目录。

重要：带 cLVM 的共享磁盘子系统

使用带 cLVM 的共享磁盘子系统时，此子系统必须连接到群集中所有需要访问此子系统的服务器。

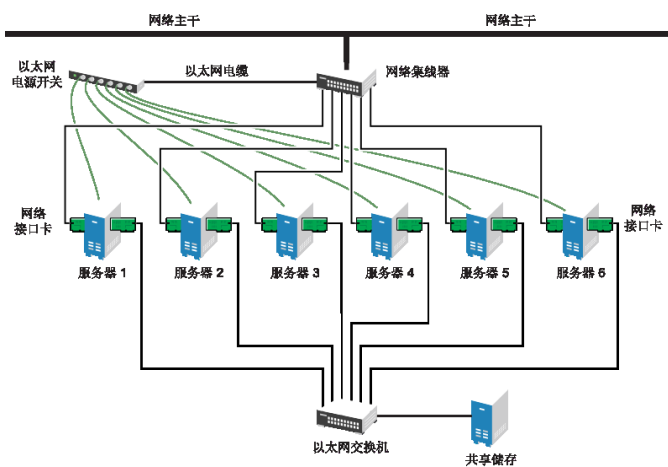
典型的资源包括数据、应用程序和服务。下图显示了一个典型的光纤通道群集配置的结构。

图 1.3 典型的光纤通道群集配置



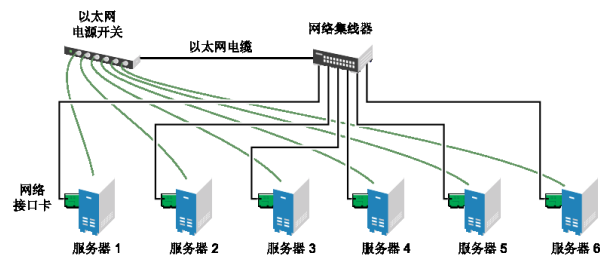
虽然光纤通道提供的性能最佳，但也可以将群集配置为使用 iSCSI。iSCSI 是除光纤通道外的另一种选择，可用于创建低成本的储存区域网络 (SAN)。下图显示了一个典型的 iSCSI 群集配置。

图 1.4 典型的 iSCSI 群集配置



虽然多数群集都包括共享磁盘子系统，但也可以创建不含共享磁盘子系统的群集。下图显示了一个不含共享磁盘子系统的群集。

图 1.5 典型的不含共享储存的群集配置



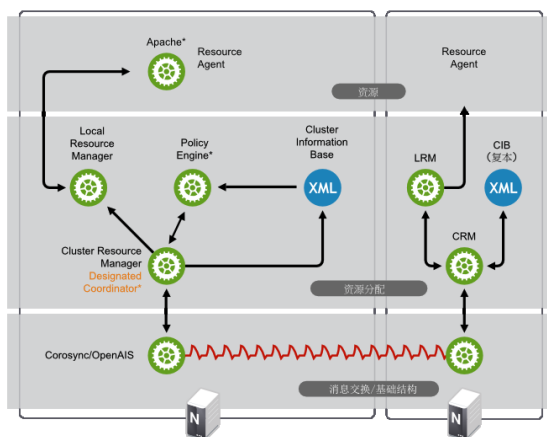
1.4 体系结构

本部分简要地概述了 High Availability Extension 体系结构。它提供了有关体系结构组件的信息，并描述了这些组件是如何协同工作的。

1.4.1 体系结构层

High Availability Extension 具有分层的体系结构。图 1.6 “体系结构”（第 12 页）说明了不同的层及其相关的组件。

图 1.6 体系结构



消息交换和基础结构层

主层或第一层是消息交换/基础结构层，也称为 Corosync/OpenAIS 层。此层包含了发送含有“我在线”信号的消息及其他信息的组件。High Availability Extension 的程序就位于此消息交换/基础结构层。

资源分配层

下一层是资源分配层。此层最复杂，它包含以下组件：

群集资源管理器 (CRM)

在资源分配层中执行的每个操作都要经过群集资源管理器。如果资源分配层的其他组件（或更高层中的组件）需要通讯，则它们通过本地 CRM 进行。

在每个节点上，CRM 维护群集信息库 (CIB)（第 13 页），包含所有群集选项、节点、资源及其关系和当前状态的定义。如果选择群集中的 CRM 为指定协调程序 (DC)，则意味着它具有主 CIB。群集中的所有其他 CIB 是此主 CIB 的副本。对 CIB 的常规读写操作通过主 CIB 进行排序。DC 是群集中唯一可以决定需要在整个群集执行更改（例如节点屏障或资源移动）的实体。

群集信息库 (CIB)

群集信息库是整个群集配置和当前状态在内存中的 XML 表示。它包含所有群集选项、节点、资源、约束及其之间的关系的定义。CIB 还将更新同步到所有群集节点。群集中有一个主 CIB，由 DC 维护。所有其他节点都包含 CIB 副本。

策略引擎 (PE)

只要指定协调程序需要进行群集范围的更改（对新 CIB 作出反应），策略引擎就会根据群集的当前状态和配置计算其下一个状态。PE 还生成一个转换图，包含用于达到下一个群集状态的（资源）操作和依赖性的列表。PE 在每个节点上都运行以加速 DC 故障转移。

本地资源管理器 (LRM)

LRM 代表 CRM 调用本地资源代理（请参见“资源层”一节（第 13 页））。因此它可以执行启动/停止/监视操作并将结果报告给 CRM。它还隐藏资源代理支持的脚本标准（OCF、LSB、Heartbeat V1）之间的区别。LRM 是其本地节点上所有资源相关信息的权威来源。

资源层

最高层是资源层。资源层包括一个或多个资源代理 (RA)。资源代理是已写入的用来启动、停止和监视某种服务（资源）的程序（通常是外壳脚本）。资源代理仅由 LRM 调用。第三方可将他们自己的代理放在文件系统中定义的位置，这样就为各自的软件提供了现成群集集成。

1.4.2 处理流程

SUSE Linux Enterprise High Availability Extension 使用 Pacemaker 作为 CRM。

CRM 作为守护程序执行 (crmd)，它在每个群集节点上都有一个实例。Pacemaker 通过选出一个 crmd 实例来充当主实例，实现所有群集决策制定的集中化。如果选定的 crmd 过程（或它所在的节点）出现故障，则将建立一个新的过程。

在每个节点上保留了一个 CIB，它反映了群集的配置和群集中所有资源的当前状态。CIB 的内容会在整个群集的同步过程中自动保留下来。

群集中执行的许多操作都将导致整个群集的更改。这些操作包括添加或删除群集资源、更改资源约束等等。了解执行这样的操作时群集中会发生的情况是很重要的。

例如，假设您要添加一个群集 IP 地址资源。为此，您可以使用一种命令行工具或 GUI 修改 CIB。您不必在 DC 上执行此操作，可以使用群集中任何节点上的任何工具，此操作会被传送到 DC 上。然后 DC 将把此 CIB 更改复制到所有群集节点。

根据 CIB 中的信息，PE 便计算群集的理想状态及如何达到此状态，并将指令列表传递给 DC。DC 通过消息交换/基础结构层发送命令，这些命令将由其他节点上的 crmd 对等体接收。每个 crmd 使用它的 LRM（作为 lrmd 实现）执行资源修改。lrmd 不是群集感知的，它直接与资源代理（脚本）交互。

所有同级节点将操作的结果报告给 DC。一旦 DC 做出所有必需操作已在群集中成功执行的结论，群集将返回至空闲状态并等待进一步事件。如果有操作未按计划执行，则会再次调用 PE，CIB 中将记录新信息。

在某些情况下，可能需要关闭节点以保护共享数据或完成资源恢复。为此，Pacemaker 附带了一个屏障子系统，stonithd。STONITH 是“Shoot The Other Node In The Head（关闭其他节点）”的首字母缩写，通常通过一个远程电源开关实施。在 Pacemaker 中，STONITH 设备已建模为资源（并在 CIB 中配置），以便轻松监视这些设备是否出现故障。而 stonithd 则负责了解 STONITH 拓扑，以便其客户端只请求受屏蔽的节点，由它来执行其余操作。

入门

在以下部分中，可了解系统要求及安装 High Availability Extension 前要做的准备。大致概览安装和设置群集的基本步骤。

2.1 硬件要求

以下列表指定了 SUSE® Linux Enterprise High Availability Extension 的群集的硬件要求。这些要求表示最低硬件配置。根据群集的用途，可能会需要其他硬件。

- 安装了第 2.2 节“软件要求”（第 16 页）中指定的软件的 1 到 16 台 Linux 服务器。服务器的硬件（内存、磁盘空间等等）无需相同。
- 至少两个 TCP/IP 通讯媒体。群集节点使用多路广播进行通讯，因此网络设备必须支持多路广播。通讯媒体应支持 100 Mbit/s 或更高的数据传送速度。最好应绑定以太网通道
- 可选：一个共享磁盘子系统，连接到群集中所有需要访问它的服务器。
- 一个 STONITH 机制。STONITH 是“Shoot the other node in the head（关闭其他节点）”的首字母缩写。STONITH 设备是电源开关，群集使用它来重设置被视为已终止或行为方式异常的节点。重设置没有检测信号的节点是确保存在但出现故障的节点未执行数据破坏的唯一可靠方法。

有关更多信息，请参考第 9 章 屏障和 *STONITH*（第 113 页）。

2.2 软件要求

请确保满足以下软件要求：

- 在所有将成为群集组成部分的节点上安装了包含全部可用联机更新的 SUSE® Linux Enterprise Server 11 SP1。
- 在所有将成为群集组成部分的节点上安装了包含全部可用联机更新的 SUSE Linux Enterprise High Availability Extension 11 SP1。

2.3 共享磁盘系统需求

如果要使数据高度可用，建议为群集使用共享磁盘系统（储存区域网络或 SAN）。如果使用共享磁盘子系统，请确保符合以下要求：

- 根据制造商的说明正确设置共享磁盘系统并且共享磁盘系统可正确运行。
- 共享磁盘系统中包含的磁盘应配置为使用镜像或 RAID，来为共享磁盘系统增加容错性。建议使用基于硬件的 RAID。所有配置都不支持基于主机的软件 RAID。
- 如果准备对共享磁盘系统访问使用 iSCSI，则请确保正确配置了 iSCSI 启动器和目标。
- 使用 DRBD 实现在两台服务器间分发数据的镜像 RAID 系统时，请确保仅访问复制的设备。请与群集的剩余节点使用相同（绑定）的 NIC 来调整此处提供的冗余。

2.4 准备工作

安装 High Availability Extension 之前，请执行以下准备步骤：

- 通过编辑群集中每个服务器上的 `/etc/hosts` 文件，配置主机名解析并使用静态主机信息。有关更多信息，请参见 <http://www.novell.com/documentation> 上的 *SUSE Linux Enterprise Server 管理指南*。请参阅章节 *基本网络 > 配置主机名和 DNS*。

群集的成员之间可按名称查找对方是基本的。如果名称不可用，则将无法进行群集内部通讯。

- 通过使群集节点与群集外的时间服务器同步来配置时间同步。有关更多信息，请参见 <http://www.novell.com/documentation> 上的 *SUSE Linux Enterprise Server 管理指南*。请参阅 *与 NTP 时间同步* 章节。

群集节点将使用此时间服务器作为它们的时间同步源。

2.5 概述：安装和设置群集

完成准备工作后，需要执行以下基本步骤才能安装和设置 SUSE® Linux Enterprise High Availability Extension 群集：

1. 将 SUSE® Linux Enterprise Server 和 SUSE® Linux Enterprise High Availability Extension 作为外接式附件安装在 SUSE Linux Enterprise Server 上。有关详细信息，请参见第 3.1 节 “安装 High Availability Extension”（第 19 页）。
2. 初始群集设置（第 20 页）
3. 使群集联机（第 27 页）
4. 配置全局群集选项并添加群集资源。

这两个操作可使用图形用户界面 (GUI) 或命令行工具完成。有关详细信息，请参见第 5 章 *配置和管理群集资源 (GUI)*（第 53 页）或第 6 章 *配置和管理群集资源 (命令行)*（第 83 页）。

5. 为保护数据免于屏障和 STONITH 可能造成的破坏，请确保将 STONITH 设备配置为资源。有关详细信息，请参见第 9 章 *屏障和 STONITH*（第 113 页）。

根据要求，您可能还想要为群集配置以下文件系统和储存相关组件：

- 在共享磁盘（储存区域网络 SAN）上创建文件系统。如有必要，将这些文件系统配置为群集资源。
- 如果需要群集感知文件系统，请使用 OCFS2。

- 要使群集可以使用逻辑卷管理器管理共享储存，请使用 cLVM，它是 LVM 的一组群集扩展。
- 要保护数据完整性，请通过使用屏蔽机制并确保排它储存访问来实施储存保护。
- 如果需要，请通过 DRBD 进行数据复制。

有关详细信息，请参见第 III 部分“储存和数据复制”（第 137 页）。

用 YaST 进行安装和基本设置

安装 High Availability 所需软件有两种方式：从命令行使用 `zypper`，或使用提供图形用户界面的 YaST 进行安装。在将作为群集一部分的所有节点上安装软件后，下一步是初始配置群集，以便使节点能够相互通讯，并启动必需的服务，使群集处于联机状态。初始群集设置既可以手动（通过编辑和复制配置文件）进行，也可以使用 YaST 群集模块进行。

本章将介绍如何从零开始对 SUSE Linux Enterprise High Availability Extension 11 SP1 进行全新安装和设置。如果要迁移运行较早版本 SUSE Linux Enterprise High Availability Extension 的现有群集，或者更新作为运行中群集一部分的节点上的任何软件包，请参阅附录 B, *将群集升级为最新产品版本*（第 341 页）一章。

3.1 安装 High Availability Extension

使用 High Availability Extension 配置和管理群集所需的包包含在 High Availability 安装模式中。此模式仅当将 SUSE® Linux Enterprise High Availability Extension 作为外接式附件安装后才可用。有关如何安装外接式附件产品的信息，请参见 <http://www.novell.com/documentation> 上的 SUSE Linux Enterprise 11 SP1 *部署指南*。请参阅 *安装外接式附件产品* 一章。

注意： 安装软件包

High Availability 群集所需的软件包不会自动复制到群集节点。

如果不想将 SUSE® Linux Enterprise Server 11 SP1 和 SUSE® Linux Enterprise High Availability Extension 11 SP1 手动安装到将作为群集一部分的所有节点

上，请使用 **AutoYaST** 克隆现有节点。有关更多信息，请参考第 3.4 节“使用 **AutoYaST** 进行大批量部署”（第 28 页）。

过程 3.1 安装 *High Availability* 模式

1 以 `root` 用户身份启动 **YaST** 并选择 **软件 > 软件管理**。

或者，以 `root` 用户身份使用 `yast2 sw_single` 从命令行启动 **YaST** 包管理器。

2 从过滤器列表中选择模式，然后在模式列表中激活高可用性模式。

3 单击接受开始安装包。

3.2 初始群集设置

安装 HA 包后，请继续执行初始群集设置。其中包括以下基本步骤：

1 定义通讯通道（第 20 页）

2 定义身份验证设置（第 23 页）

3 将配置传送到所有节点（第 24 页）

以下过程将指引您使用 **YaST** 群集模块完成每个步骤。要访问群集配置对话框，请以 `root` 用户身份启动 **YaST** 并选择 *High Availability* > 群集。或者，以 `root` 用户身份使用 `yast2 cluster` 从命令行启动 **YaST** 群集模块。

如果是首次启动群集模块，它会显示向导，指引您完成进行基本设置所需的所有步骤。否则，请单击左侧面板上的类别，以访问每个步骤的配置选项。

3.2.1 定义通讯通道

为实现群集节点间的成功通讯，请定义至少一个通讯通道。然而，建议通过两个或更多冗余路径设置通讯（通过使用网络设备绑定或使用 **Corosync** 添加第二通讯通道）。对于每个通讯通道，需要定义以下参数：

绑定网络地址 (bindnetaddr)

要绑定到的网络地址。为方便在群集间共享配置文件，OpenAIS 使用网络界面网络掩码来屏蔽仅用于路由网络的地址位。将此值设置为要用于群集多路广播的子网。

多路广播地址 (mcastaddr)

可以是 IPv4 或 IPv6 地址。

多路广播端口 (mcastport)

为 mcastaddr 指定的 UDP 端口。

群集中的所有节点通过使用同一多路广播地址和同一端口号相互感知。对于不同的群集，请使用不同的多路广播地址。

要使用 Corosync 配置冗余通讯，需要在 `/etc/corosync/corosync.conf` 中定义多个接口部分，每个接口部分具有不同的环号。通过冗余环协议 (RRP) 告知群集如何使用这些接口。RRP 可具有三种模式 (rrp_mode)：如果设置为 active，则 Corosync 将主动使用所有接口。如果设置为 passive，则 Corosync 只会在第一个环发生故障时才使用第二个接口。如果将 rrp_mode 设置为 none，则将禁用 RRP。有了 RRP，就可以使用两个物理位置分开的网络进行通讯。如果一个网络发生故障，群集节点仍可通过另一个网络进行通讯。

如果配置了多个环，每个节点都可具有多个 IP 地址。一旦启用了 rrp_mode，则默认将使用流控制传送协议 (SCTP)（而非 TCP）进行节点间通讯。

过程 3.2 定义通讯通道

- 1 在 YaST 群集模块中，切换到通讯通道类别。
- 2 定义用于所有群集节点的绑定网络地址、多路广播地址和多路广播端口。

Cluster - 通讯通道

通信通道

绑定网络地址: 192.168.8.0

组播地址: 192.168.0.254

组播端口: 5405

☐ 冗余通道

绑定网络地址:

组播地址:

组播端口:

节点 ID

☐ Auto Generate Node ID

节点 ID: 2

rrp mode: none

帮助 取消 完成

3 如果要定义第二通道：

3a 请激活冗余通道。

3b 定义冗余通道的绑定网络地址、多路广播地址和多路广播端口。

3c 选择要使用的 *rrp_mode*。要禁用 RRP，请选择 *None*。有关模式的更多信息，请单击帮助。

使用 RRP 时，主环（已配置的第一通道）和次环（冗余通道）在 `/etc/corosync/corosync.conf` 中的环号分别为 0 和 1。

4 激活自动生成节点 ID 以自动为每个群集节点生成唯一的 ID。

5 如果只想修改现有群集的通讯通道，请单击完成将配置写入 `/etc/corosync/corosync.conf` 并关闭 YaST 群集模块。YaST 随后还会自动调整防火墙设置并打开用于多路广播的 UDP 端口。

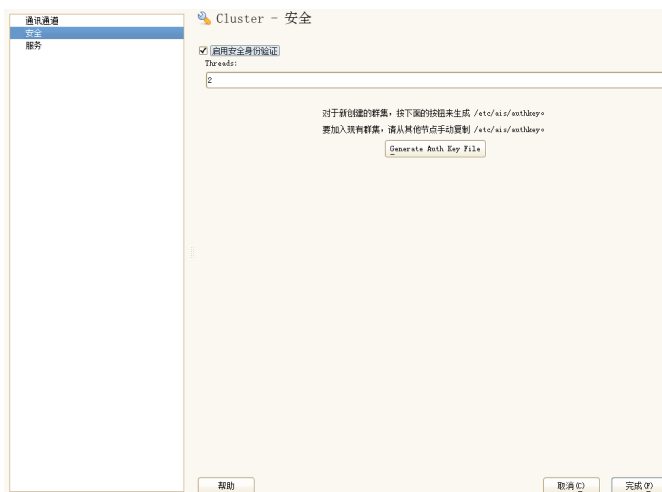
6 要进一步配置群集，请继续执行过程 3.3，“启用安全性身份验证”（第 23 页）。

3.2.2 定义身份验证设置

下一步，为群集定义身份验证设置。可使用需要共享机密的、用于保护消息并对其进行身份验证的 HMAC/SHA1 身份验证。指定的身份验证密钥（密码）将用于群集中的所有节点。

过程 3.3 启用安全性身份验证

- 1 在 YaST 群集模块中，切换到安全性类别。
- 2 激活启用安全性身份验证。
- 3 对于新创建的群集，请单击生成身份验证密钥文件。这将创建被写入 `/etc/corosync/authkey` 的身份验证密钥。



- 4 如果只想修改身份验证设置，请单击完成将配置写入 `/etc/corosync/corosync.conf` 并关闭 YaST 群集模块。
- 5 要进一步配置群集，请继续执行第 3.2.3 节“将配置传送到所有节点”（第 24 页）。

3.2.3 将配置传送到所有节点

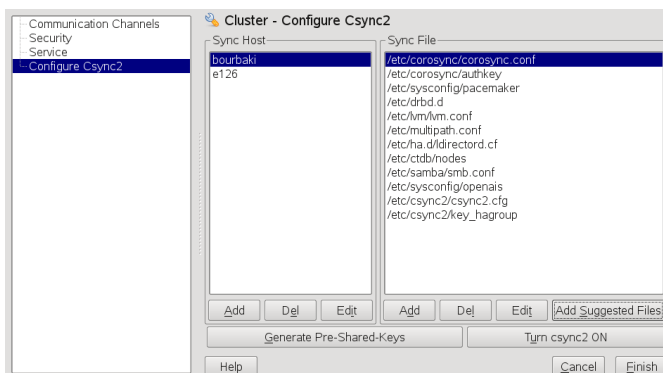
如果不想将生成的配置文件手动复制到所有节点，可使用 `csync2` 工具在群集中的所有节点间进行复制。`Csync2` 可处理排入同步组的任意数量的主机。每个同步组都有自己的成员主机列表及其包含/排除模式，包含/排除模式定义了在同步组中应同步哪些文件。同步组、属于每个组的主机名以及每个组的包含/排除规则均在 `Csync2` 配置文件 `/etc/csync2/csync2.cfg` 中指定。

对于身份验证，`Csync2` 使用 IP 地址和同步组中的预共享密钥。需要为每个同步组生成一个密钥文件，并将其复制到所有组成员。

有关 `Csync2` 的更多信息，请参阅 <http://oss.linbit.com/csync2/paper.pdf>。

过程 3.4 使用 YaST 配置 `Csync2`

- 1 在 YaST 群集模块中，切换到 `Csync2` 类别。
- 2 要指定同步组，请在*同步主机组*中单击*添加*，然后输入群集中所有节点的本地主机名。对于每个节点，必须使用 `hostname` 命令返回的确切字符串。
- 3 单击*生成预共享密钥*以创建同步组的密钥文件。密钥文件将写入 `/etc/csync2/key_hagroup`。创建后，必须将其手动复制到群集的所有成员。
- 4 要使用通常需要在所有节点间同步的文件填充*同步文件列表*，请单击*添加建议的文件*。



- 5 如果要从待同步的文件列表编辑、添加或删除文件，请使用相应按钮。必须为每个文件输入绝对路径名。
- 6 通过单击 *打开 Csync2* 激活 Csync2。这将在引导时自动启动 Csync2。
- 7 根据意愿设置所有选项后，请单击完成以关闭 YaST 群集模块。YaST 随即 *将 Csync2 配置写入 /etc/csync2/csync2.cfg*。

配置完 Csync2 后，按下述方式从命令行启动同步进程。

过程 3.5 使用 Csync2 同步配置文件

要使用 Csync2 成功同步文件，请确保满足以下先决条件：

- 同一 Csync2 配置在所有节点上均可用。将 */etc/csync2/csync2.cfg* 包含到要使用 Csync2 同步的文件列表中，或者按过程 3.4，“使用 YaST 配置 Csync2”（第 24 页）中所述配置此文件后，将其手动复制到所有节点。
- 将步骤 3（第 24 页）中在一个节点上生成的 */etc/csync2/key_hagroup* 文件复制到群集中的所有节点，因为它是 Csync2 进行身份验证时所必需的。但是，不要尝试在其他节点上重新生成文件，因为所有节点上的文件必须是同一文件。
- 确保 *xinetd* 正运行于所有节点上，因为 Csync2 依赖于此守护程序。以 *root* 用户身份使用以下命令启动 *xinetd*：

```
rcxinetd start
```

注意：引导时启动服务

如果希望 **Csync2** 和 **xinetd** 在引导时自动启动，请在所有节点上执行以下命令：

```
chkconfig csync2 on
chkconfig xinetd on
```

1 通过在一个节点上执行以下命令启动文件同步：

```
csync2 -xv
```

这将一次性同步所有文件。如果所有文件均可成功同步，则 **Csync2** 将完成，不会报错。

如果在其他节点（不仅在当前节点）上对要同步的一个或多个文件也进行了修改，则 **Csync2** 将报告冲突。您将得到类似以下内容的输出：

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

2 如果确信当前节点上的文件版本是“最佳”版本，可以通过强制使用此文件并重新同步来解决冲突：

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

有关 **Csync2** 选项的更多信息，请运行 `csync2 -help`。

注意：触发同步

Csync2 不会在节点间连续同步文件。每次更新了需要同步的任何文件后，都必须手动重新同步文件。

将密钥文件同步到群集中的所有节点后，请按第 3.3 节“使群集联机”（第 27 页）中所述启动基本服务，使群集处于联机状态。

3.2.4 启动服务

（可选）使用 YaST 可定义在引导时是否要在节点上启动特定服务。也可以用此模块手动启动和停止服务（如果您不想使用命令行来实现操作）。为使群集节点处于联机状态并启动群集资源管理器，必须将 OpenAIS 作为服务启动。

过程 3.6 启动或停止服务

- 1 在 YaST 群集模块中，切换到*服务类别*。
- 2 要在每次引导此群集节点时启动 OpenAIS，请在*引导组*中选择相应选项。
- 3 如果要使用 Pacemaker GUI 配置、管理和监视群集资源，请激活*同时启动 mgmt*。GUI 需要此守护程序。
- 4 要立即启动或停止 OpenAIS，请单击相应按钮。
- 5 单击*完成*以关闭 YaST 群集模块。

如果在*引导组*中选择了关，则必须在每次引导此节点时手动启动 OpenAIS。要手动启动 OpenAIS，请使用 `rcopenais start` 命令。

3.3 使群集联机

完成初始群集配置后，即可启动使堆栈处于联机状态所需的服务。

过程 3.7 启动 OpenAIS/Corosync 并检查状态

- 1 在每个群集节点上运行以下命令以启动 OpenAIS/Corosync:

```
rcopenais start
```

- 2 在任一节点上，用以下命令检查群集状态:

```
crm_mon
```

如果所有节点都联机，则输出应类似于如下内容:

```
=====  
Last updated: Tue Mar  2 18:35:34 2010  
Stack: openais
```

```
Current DC: e229 - partition with quorum
Version: 1.1.1-530add2a3721a0eccc24660a97dbfdaa3e68f51
2 Nodes configured, 2 expected votes
0 Resources configured.
=====

Online: [ e231 e229 ]
```

此输出表示群集资源管理器已启动，可以管理资源了。

完成基本配置后并使节点处于联机状态后，即可开始配置群集资源。使用 `crm` 命令行工具或图形用户界面。有关更多信息，请参见第 5 章 *配置和管理群集资源 (GUI)*（第 53 页）或第 6 章 *配置和管理群集资源（命令行）*（第 83 页）。

3.4 使用 AutoYaST 进行大批量部署

AutoYaST 是自动安装一个或多个 SUSE Linux Enterprise 系统而无需用户干预的系统。使用 SUSE Linux Enterprise 可创建包含安装和配置数据的 AutoYaST 配置文件。此配置文件将告知 AutoYaST 要安装的内容以及如何配置安装好的系统，以最终获得一个即用型的系统。此配置文件随即可用于以不同方式进行大批量部署。

有关在各种情境下如何利用 AutoYaST 的详细指示信息，请参见 <http://www.novell.com/documentation> 上的 SUSE Linux Enterprise 11 SP1 *部署指南*。请参阅 *自动安装* 一章。

过程 3.8 使用 AutoYaST 克隆群集节点

以下过程适合用于部署作为已存在节点的克隆的群集节点。克隆节点会安装相同的包，并具有相同的系统配置。

如果需要在不同硬件上部署群集节点，请参阅 <http://www.novell.com/documentation> 上 SUSE Linux Enterprise 11 SP1 *部署指南* 中的 *基于规则的自动安装* 部分。

重要：相同硬件

此情境假定您要将 SUSE Linux Enterprise High Availability Extension 11 SP1 部署到具有完全相同的硬件配置的一组计算机上。

- 1 确保已按第 3.1 节“安装 High Availability Extension”（第 19 页）和第 3.2 节“初始群集设置”（第 20 页）中所述对要克隆的节点进行了正确的安装和配置。
- 2 按 SUSE Linux Enterprise 11 SP1 *部署指南* 中所述进行简单的大批量安装。其中包括以下基本步骤：
 - 2a 创建 AutoYaST 配置文件。使用 AutoYaST GUI 从现有系统配置创建和修改配置文件。在 AutoYaST 中选择 *High Availability* 模块并单击克隆按钮。如果需要，调整其他模块中的配置，并将生成的控制文件另存为 XML 格式的文件。
 - 2b 确定 AutoYaST 配置文件的来源以及要传递到其他节点的安装例程的参数。
 - 2c 确定 SUSE Linux Enterprise Server 和 SUSE Linux Enterprise High Availability Extension 安装数据的来源。
 - 2d 确定并设置自动安装的引导方案。
 - 2e 通过手动添加参数或创建 `info` 文件，将命令行传递到安装例程。
 - 2f 启动并监视自动安装进程。

成功安装克隆节点后，执行以下步骤将克隆节点加入群集中：

过程 3.9 使克隆节点处于联机状态

- 1 按第 3.2.3 节“将配置传送到所有节点”（第 24 页）中所述使用 Csync2 将密钥配置文件从已配置的节点传送到克隆节点。
- 2 按第 3.3 节“使群集联机”（第 27 页）中所述在克隆节点上启动 OpenAIS 服务以使节点处于联机状态。

现在克隆节点将加入群集，因为 `/etc/corosync/corosync.config` 文件已通过 Csync2 应用到克隆节点。CIB 将在群集节点间自动同步。

部分 II. 配置和管理

配置和管理基础

HA 群集的主要目的是管理用户服务。用户服务的典型示例是 Apache Web 服务器或数据库。从用户角度来看，服务就是在客户的要求下执行某些操作。但对群集来说，服务则是可以启动或停止的资源 — 服务的本质与群集无关。

在本章中，我们将介绍一些配置资源和管理群集时需要了解的基本概念。以下章节介绍如何使用 High Availability Extension 提供的每种管理工具执行主配置和管理任务。

4.1 全局群集选项

全局群集选项控制群集在遇到特定情况时的行为方式。它们组成集合，并可使用 Pacemaker GUI 和 `crm` 外壳之类的群集管理工具进行查看和修改。在大多数情况下可保留预定义值。但为了使群集的关键功能正常工作，需要在进行基本群集设置后调整以下参数：

- 选项 `no-quorum-policy`（第 34 页）
- 选项 `stonith-enabled`（第 35 页）

要了解如何使用 GUI 调整这些参数，请参见过程 5.1, “修改全局群集选项”（第 57 页）。如果想要使用命令行方法，请参见第 6.2 节 “配置全局群集选项”（第 89 页）。

4.1.1 选项 no-quorum-policy

此全局选项定义在群集无仲裁人数（大多数节点不是分区的一部分）时应执行的操作。

允许的值有：

`ignore`

仲裁人数状态不会对群集行为产生任何影响，将继续进行资源管理。

此设置在以下情况中非常有用：

- 双节点群集：由于单个节点故障总是会导致大多数节点丢失通讯，而您通常希望群集能够继续运行。使用屏蔽可确保资源完整性，还可防止出现节点分裂的情况。
- 资源驱动型群集：对于具有冗余通讯通道的本地群集，节点分裂的情况只存在一定的可能性。因此，节点的通讯丢失很可能表示此节点已崩溃，剩余节点应恢复并重新为资源服务。

如果 `no-quorum-policy` 设置为 `ignore`，则 4 节点群集可以在服务丢失之前承受三个节点的并发故障，但如果使用其他设置，此群集将在两个节点发生并发故障后丢失仲裁人数。

`freeze`

如果仲裁人数丢失，群集将冻结。继续进行资源管理：正在运行的资源不会停止（但可能重新启动以响应监视事件），但不会启动受影响分区中的任何其他资源。

如果群集中的某些资源依赖于与其他节点的通讯（例如，OCFS2 装入），建议对此类群集使用此设置。在这种情况下，默认设置

`no-quorum-policy=stop` 没有任何作用，因为它将导致以下情况：在对等节点不可到达时将无法停止这些资源。反之，尝试停止这些资源最终将超时并导致 `stop failure`，进而触发升级恢复和屏蔽。

`stop`（默认值）

如果仲裁人数丢失，受影响群集分区中的所有资源都将以一种有序的方式停止。

suicide

屏蔽受影响群集分区中的所有节点。

4.1.2 选项 `stonith-enabled`

此全局选项定义是否要应用屏蔽，以允许 STONITH 设备关闭故障节点以及无法停止其资源的节点。默认情况下，此全局选项设置为 `true`，因为对于常规的群集操作，有必要使用 STONITH 设备。根据默认值，如果未定义 STONITH 资源，则群集将拒绝启动任何资源。

如果由于某些原因需要禁用屏蔽，请将 `stonith-enabled` 设置为 `false`。

有关所有全局群集选项及其默认值的概述，请参见 <http://clusterlabs.org/wiki/Documentation> 上的 *Pacemaker 1.0 - 配置说明*。请参阅 *可用的群集选项* 一节。

4.2 群集资源

作为群集管理员，您需要在群集中为服务器上运行的每个资源或应用程序创建群集资源。群集资源可以包括 Web 站点、电子邮件服务器、数据库、文件系统、虚拟机和任何其他基于服务器的应用程序或在任意时间对用户都可用的服务。

4.2.1 资源管理

必须先设置群集中的资源，然后才能使用它。例如，如果要使用 Apache 服务器作为群集资源，请先设置 Apache 服务器并完成 Apache 配置，然后才能启动群集中的相应资源。

如果资源有特定环境要求，请确保这些要求已得到满足并且在所有群集节点上均相同。这种配置不由 High Availability Extension 管理。您必须自行管理。

注意：不要处理由群集管理的服务

使用 High Availability Extension 管理资源时，不得以其他方式（在群集外，例如手动或者引导时或重引导时）启动或停止同一资源。High Availability Extension 软件负责所有服务的启动或停止操作。

但是，如果要检查服务是否正确配置，可手动启动该服务，但请确定在 High Availability 接管前再次停止该服务。

配置群集中的资源后，请使用群集管理工具手动启动、停止、清理、删除或迁移资源。有关如何执行此操作的详细信息，请参阅第 5 章 *配置和管理群集资源 (GUI)*（第 53 页）或第 6 章 *配置和管理群集资源 (命令行)*（第 83 页）。

4.2.2 支持的资源代理类

对于添加的每个群集资源，需要定义资源代理需遵守的标准。资源代理提取它们提供的服务并显示群集的确切状态，以使群集对其管理的资源不作确答。群集依赖于资源代理在收到启动、停止或监视命令时作出相应反应。

通常，资源代理的形式为外壳脚本。High Availability Extension 支持以下各种资源代理：

旧版 Heartbeat 1 资源代理

Heartbeat 版本 1 附带自己的资源代理样式。由于很多用户已根据约定编写了自己的代理，所以同样也支持这些资源代理。但是，建议如有可能请将配置迁移到 High Availability OCF RA。

Linux Standards Base (LSB) 脚本

LSB 资源代理一般由操作系统/分发包提供，并可在 `/etc/init.d` 中找到。要用于群集，它们必须遵守 LSB `init` 脚本规范。例如，它们必须实施了多个操作，至少包括 `start`、`stop`、`restart`、`reload`、`force-reload` 和 `status`。有关详细信息，请参见<http://ldn.linuxfoundation.org/lsb/lsb4-resource-page%23Specification>。

这些服务的配置没有标准化。如果要将 LSB 脚本用于 High Availability，请确保您了解如何配置相关脚本。通常，您可以在 `/usr/share/doc/packages/PACKAGENAME` 中的相关包文档中找到配置此类脚本的信息。

Open Cluster Framework (OCF) 资源代理

OCF RA 代理最适合用于 High Availability，尤其是当您需要主资源或特殊监视功能时。这些代理通常位于 `/usr/lib/ocf/resource.d/provider/` 中。其功能与 LSB 脚本的功能相似。但是，始终使用环境变量进行配置，这样可轻松接受和处理参数。OCF 规范（由于它与资源代理相关）可在 <http://www.opencf.org/cgi-bin/viewcvs.cgi/specs/ra/resource-agent-api.txt?rev=HEAD&content-type=text/vnd.viewcvs-markup> 中找到。OCF 规范有严格的定义，其中包括操作必须返回退出码，请参见第 8.3 节“OCF 返回码和故障恢复”（第 111 页）。群集严格遵循这些规范。有关所有可用的 OCF RA 的详细列表，请参见第 19 章 *HA OCF Agents*（第 243 页）。

要求所有 OCF 资源代理至少包含操作 `start`、`stop`、`status`、`monitor` 和 `meta-data`。`meta-data` 操作可检索有关如何配置代理的信息。例如，如果要了解提供程序 `heartbeat` 的 `IPaddr` 代理的更多信息，请使用以下命令：

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

输出是 XML 格式的信息，包括多个部分（代理的常规描述、可用参数和可用操作）。

STONITH 资源代理

此类仅用于与屏障相关的资源。有关详细信息，参见第 9 章 *屏障和 STONITH*（第 113 页）。

随 High Availability Extension 提供的代理已写入 OCF 规范。

4.2.3 资源类型

可创建以下类型的资源：

原始资源

原始资源是最基本的资源类型。

要了解如何使用 GUI 创建原始资源，请参见过程 5.2，“添加原始资源”（第 57 页）。如果想要使用命令行方法，请参见第 6.3.1 节“创建群集资源”（第 90 页）。

组

组包含一组需要放在一起、按顺序启动和按相反顺序停止的资源。有关更多信息，请参考“组”一节（第 38 页）。

克隆资源

克隆是可以在多个主机上处于活动状态的资源。如果各个资源代理支持，则任何资源均可克隆。有关更多信息，请参考“克隆资源”一节（第 40 页）。

主资源

主资源是一种特殊类型的克隆资源，它们可以有多个节点。有关更多信息，请参考“主资源”一节（第 40 页）。

4.2.4 高级资源类型

原始资源是最简单的一种资源，易于配置，您可能还需要更多高级资源类型进行群集配置，如组、克隆资源或主资源。

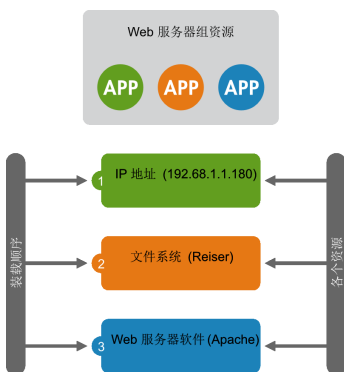
组

某些群集资源依赖于其他组件或资源，并且要求每个组件或资源都按特定顺序启动并在同一服务器上一起运行。要简化此配置，可以使用组。

例 4.1 Web 服务器的资源组

资源组示例可以是需要 IP 地址和文件系统的 Web 服务器。在本例中，每个组件都是组成群集资源组的单独的群集资源。资源组在一台或多台服务器上运行，如果软件或硬件有故障，故障转移至群集中的另一台服务器上，这与单个群集资源相同。

图 4.1 组资源



组具有以下属性：

启动和停止

资源按其显示顺序启动并按相反顺序停止。

相关性

如果组中某个资源在某处无法运行，则该组中位于其之后的任何资源都不允许运行。

内容

组可能仅包含一些原始群集资源。组必须包含至少一个资源，否则配置无效。要引用组资源的子级，请使用子级 ID 而不是组 ID。

限制

尽管在约束中可以引用组的子代，但通常倾向于使用组的名称。

黏性

黏性在组中可以累加。每个活动的组成员可以将其黏性值累加到组的总分中。因此，如果默认 `resource-stickiness` 是 100，且组有 7 个成员（其中 5 个成员处于活动状态），那么整个组当前位置分数将是 500。

资源监控

要为组启用资源监视，必须为组中每个要监视的资源分别配置监视。

要了解如何使用 GUI 创建组，请参过程 5.12, “添加资源组”（第 72 页）。如果想要使用命令行方法，请参见第 6.3.9 节 “配置群集资源组”（第 100 页）。

克隆资源

您可能希望某些资源在群集的多个节点上同时运行。为此，必须将资源配置为克隆资源。可以配置为克隆资源的资源示例包括 STONITH 和群集文件系统（如 OCFS2）。可以克隆提供的任何资源。资源的资源代理支持此操作。克隆资源的配置甚至也有不同，具体取决于资源驻留的节点。

资源克隆有三种类型：

匿名克隆

这是最简单的克隆类型。这种克隆类型在所有位置上的运行方式都相同。因此，每台计算机上只能有一个匿名克隆实例是活动的。

全局唯一克隆

这些资源各不相同。一个节点上运行的克隆实例与另一个节点上运行的实例不同，同一个节点上运行的任何两个实例也不同。

状态克隆

这些资源的活动实例分为两种状态：主动和被动。有时也称为主要和辅助，或主和从。状态克隆可以是匿名克隆也可以是全局唯一克隆。另请参见“主资源”一节（第 40 页）。

克隆资源必须正好包含一组或一个常规资源。

配置资源监视或约束时，主资源与简单资源具有不同的要求。有关细节，请参见 <http://clusterlabs.org/wiki/Documentation> 上的 *Pacemaker 1.0 - 配置说明*。请参阅 *克隆资源 - 应在多个主机上处于活动状态的资源* 一节。

要了解如何使用 GUI 创建克隆资源，请参见过程 5.14，“添加或修改克隆”（第 76 页）。如果想要使用命令行方法，请参见第 6.3.10 节“配置克隆资源”（第 100 页）。

主资源

主资源是克隆资源的特殊化，允许实例为两种操作模式中的一种（master 或 slave）。主资源必须只能包含一个组或一个常规资源。

配置资源监视或约束时，主资源与简单资源具有不同的要求。有关细节，请参见 <http://clusterlabs.org/wiki/Documentation> 上的 *Pacemaker 1.0 - 配置说明*。请参阅 *多状态 - 具有多个节点的资源* 一节。

4.2.5 资源选项（元属性）

您可以为添加的每个资源定义选项。群集使用这些选项来决定资源的行为方式，它们会告知 CRM 如何对待特定的资源。可以使用 `crm_resource --meta` 命令或 GUI 按过程 5.3, “添加或修改元属性和实例属性”（第 59 页）中所述设置资源选项。

表 4.1 原始资源选项

选项	描述
<code>priority</code>	如果不允许所有的资源都处于活动状态，群集会停止优先级较低的资源以便保持较高优先级资源处于活动状态。
<code>target-role</code>	群集应在哪种状态下尝试保留此资源？允许的值有： <code>stopped</code> 和 <code>started</code> 。
<code>is-managed</code>	是否允许群集启动和停止资源？允许的值： <code>true</code> 和 <code>false</code> 。
<code>resource-stickiness</code>	资源留在所处位置的自愿程度如何？默认为 <code>default- resource-stickiness</code> 的值。
<code>migration-threshold</code>	节点上的此资源应发生多少故障后才能确定该节点没有资格主管此资源？默认值： <code>none</code> 。
<code>multiple-active</code>	如果群集发现资源在多个节点上处于活动状态，应执行什么操作？允许的值： <code>block</code> （将资源标记为未受管）、 <code>stop_only</code> 和 <code>stop_start</code> 。
<code>failure-timeout</code>	在恢复为如同未发生故障一样正常工作（并允许资源返回它发生故障的节点）之前，需要等待几秒钟？默认值： <code>never</code> 。
<code>allow-migrate</code>	允许对支持 <code>migrate_to/migrate_from</code> 操作的资源进行资源迁移。

4.2.6 实例属性

可为所有资源类的脚本指定参数，这些参数可确定脚本的行为方式和所控制的服务实例。如果资源代理支持参数，则可使用 `crm_resource` 命令或 GUI 按过程 5.3, “添加或修改元属性和实例属性”（第 59 页）中所述添加这些参数。在 `crm` 命令行实用程序中，实例属性称为 `params`。通过以 `root` 身份执行以下命令可以找到 OCF 脚本支持的实例属性列表：

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

或更简短：

```
crm ra info resource_agent
```

输出列出了所有支持的属性及其用途和默认值。

例如，命令

```
crm ra info Ipaddr
```

返回以下输出：

```
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

```

注意：组、克隆资源或主资源的实例属性

请注意，组、克隆资源和主资源没有实例属性。但是，任何实例属性集都将由组、克隆资源或主资源的子级继承。

4.2.7 资源操作

默认情况下，群集将不会确保您的资源一直正常。要指示群集执行此操作，需要将监视操作添加到资源定义中。可为所有类或资源代理添加监视操作。有关更多信息，请参考第 4.3 节“资源监控”（第 45 页）。

表 4.2 资源操作

操作	描述
ID	您的操作名称。必须是唯一的。（不会显示 ID）。
name	要执行的操作。常见值：monitor、start 和 stop。
interval	执行操作的频率。单位：秒
timeout	需要等待多久才能声明操作失败。
requires	需要满足哪些条件才会发生此操作。允许的值： nothing、quorum 和 fencing。默认值取决于是否 启用屏障和资源的类是否为 stonith。对于 STONITH 资源，默认值为 nothing。
on-fail	此操作失败时执行的操作。允许的值： <ul style="list-style-type: none"> • ignore：假装资源没有失败。 • block：不对资源执行任何进一步操作。 • stop：停止资源并且不在其他位置启动该资源。 • restart：停止资源并（可能在不同的节点上）重启动。 • fence：关闭资源失败的节点 (STONITH)。 • standby：将所有资源从资源失败的节点上移走。
enabled	如果值为 false，将操作视为不存在。允许的值： true、false。
role	仅当资源具有此角色时才运行操作。

操作	描述
record-pending	可全局设置或为单独资源设置。使 CIB 反映资源上“正在进行中的”操作的状态。
description	操作描述。

4.3 资源监控

如果要确保资源正在运行，必须为其配置资源监视。

如果资源监视程序检测到故障，将发生以下情况：

- 根据 `/etc/corosync/corosync.conf` 中 `logging` 部分指定的配置生成日志文件消息。默认情况下，日志将写入系统日志，通常为 `/var/log/messages`。
- 故障将反映在群集管理工具（Pacemaker GUI、HA Web Konsole 和 `crm_mon`）中和 CIB 状态部分中。
- 群集将启动重要的恢复操作，可包括停止资源以修复故障状态以及在本地或
在其他节点上重新启动资源。资源也可能根本不会重新启动，具体取决于配置和群集状态。

如果不配置资源监视，则不会告知成功启动的资源故障，且群集始终显示资源状况正常。

要了解如何使用 GUI 将监视操作添加到资源中，请参见过程 5.11, “添加或修改监视操作”（第 70 页）。如果想要使用命令行方法，请参见第 6.3.8 节 “配置资源监视”（第 99 页）。

4.4 资源约束

配置好所有资源只是完成了该任务的一部分。即便群集熟悉所有必需资源，它可能还无法进行正确处理。资源约束允许您指定在哪些群集节点上运行资源、以何种顺序装载资源，以及特定资源依赖于哪些其他资源。

4.4.1 约束类型

提供三种不同的约束：

Resource Location（资源位置）

位置约束定义资源可以、不可以或首选在哪些节点上运行。

Resource Collocation（资源排列）

排列约束告诉群集资源可以或不可以在某个节点上一起运行。

Resource Order（资源顺序）

排序约束定义操作的顺序。

有关配置约束的更多信息以及顺序和排列基本概念的详细背景信息，请参考 <http://clusterlabs.org/wiki/Documentation> 上提供的以下文档：

- *Pacemaker 1.0 - 配置说明，资源约束一章*
- *排列说明*
- *排序说明*

要了解如何使用 GUI 添加各种约束，请参见第 5.3.3 节“配置资源约束”（第 62 页）。如果想要使用命令行方法，请参见第 6.3.4 节“配置资源约束”（第 94 页）。

4.4.2 分数和无限值

定义约束时，还需要指定分数。各种分数是群集工作方式的重要组成部分。其实，从迁移资源到决定在已降级群集中停止哪些资源的整个过程是通过以某种方式操纵分数来实现的。分数按每个资源来计算，资源分数为负的任何节点都无法运行该资源。计算资源的分数后，群集会选择分数最高的节点。

INFINITY（无穷大）目前定义为 1,000,000。提高或降低分数需遵循以下三个基本规则：

- 任何值 + 无穷大 = 无穷大
- 任何值 - 无穷大 = -无穷大

- 无穷大 - 无穷大 = -无穷大

定义资源约束时，需为每个约束指定一个分数。分数表示您指派给此资源约束的值。分数较高的约束先应用，分数较低的约束后应用。通过使用不同的分数为既定资源创建更多位置约束，可以指定资源要故障转移至的目标节点的顺序。

4.4.3 故障转移节点

资源在出现故障时会自动重启动。如果在当前节点上无法实现此操作，或者此操作在当前节点上失败了 N 次，它将尝试故障转移到其他节点。每次资源失败时，其失败计数都会增加。您可以多次定义资源的故障次数

(`migration-threshold`)，在该值之后资源会迁移到新节点。如果群集中存在两个以上的节点，特定资源故障转移的节点由 **High Availability** 软件选择。

但可以通过为资源配置一个或多个位置约束和一个 `migration-threshold` 来指定此资源将故障转移到的节点。有关如何使用 **GUI** 实现此操作的详细指示信息，请参阅第 5.3.4 节“指定资源故障转移节点”（第 65 页）。如果想要使用命令行方法，请参见第 6.3.5 节“指定资源故障转移节点”（第 96 页）。

例 4.2 迁移阈值 - 进程流

例如，假设您已经为 `r1` 资源配制了一个首选在 `node1` 节点上运行的位置约束。如果那里失败了，系统会检查 `migration-threshold` 并与故障计数进行比较。如果故障计数 \geq `migration-threshold`，会将资源迁移到下一个自选节点。

默认情况下，一旦达到阈值，节点将不再能运行失败资源，直到重设置资源的失败计数为止。这可以由群集管理员手动执行或通过设置资源的 `failure-timeout` 选项执行。

例如，`migration-threshold=2` 和 `failure-timeout=60s` 设置将导致资源在两次失败后迁移到新节点，并可能在一分钟后回退（取决于粘性和约束分数）。

迁移阈值概念有两个异常，发生在资源启动失败或停止失败时：

- 启动失败将失败计数设置为 `INFINITY`，因此总是会导致立即迁移。
- 停止故障会导致屏障（`stonith-enabled` 设置为 `true` 时，这是默认设置）。

如果未定义 STONITH 资源（或 stonith-enabled 设置为 false），则该资源根本不会迁移。

有关使用迁移阈值和重置失败计数的细节，请参阅第 5.3.4 节“指定资源故障转移节点”（第 65 页）。如果想要使用命令行方法，请参见第 6.3.5 节“指定资源故障转移节点”（第 96 页）。

4.4.4 故障回复节点

当原始节点恢复联机并位于群集中时，资源可能会故障回复到该节点。如果要防止资源在故障转移前故障回复到之前运行的节点，或者要指定此资源故障回复到的其他节点，必须更改其资源粘性值。可以在创建资源时指定资源粘性或稍后指定。

指定资源粘性值时请考虑以下含义：

值为 0：

这是默认选项。资源放置在系统中的最适合位置。这意味着当负载能力“较好”或较差的节点变得可用时才转移资源。此选项的作用几乎等同于自动故障回复，只是资源可能会转移到非之前活动的节点上。

值大于 0：

资源更愿意留在当前位置，但是如果有更合适的节点可用时会移动。值越高表示资源更愿意留在当前位置。

值小于 0：

资源更愿意移离当前位置。绝对值越高表示资源越愿意离开当前位置。

值为 INFINITY：

如果不是因节点不适合运行资源（节点关机、节点待机、达到 migration-threshold 或配置更改）而强制资源转移，资源总是留在当前位置。此选项的作用几乎等同于完全禁用自动故障回复。

值为 -INFINITY：

资源总是移离当前位置。

4.4.5 根据资源负载影响放置资源

并非所有资源都相等。某些资源（如 Xen guest）需要托管它们的节点满足其容量要求。如果所放置资源的总需求超过了提供的容量，则资源性能将降低（或甚至失败）。

要考虑此情况，可使用 High Availability Extension 指定以下参数：

1. 特定节点提供的容量。
2. 特定资源需要的容量。
3. 资源放置整体策略。

当前这些设置为静态，且必须由管理员配置 - 它们不能动态发现或调整。

要了解如何使用 GUI 配置这些设置，请参见第 5.3.6 节“根据负载影响配置资源放置”（第 67 页）。如果想要使用命令行方法，请参见第 6.3.7 节“根据负载影响配置资源放置”（第 97 页）。

如果节点有充足的可用容量来满足资源要求，则此节点将被视为此资源的有效节点。所需或所提供容量的性质对 High Availability Extension 而言完全无关紧要，它只是确保在将资源移动到节点上之前满足资源的所有容量要求。

要配置资源要求和节点提供的容量，请使用利用率属性。可根据个人喜好命名利用率属性，并根据配置需要定义多个名称/值对。但是，属性值必须是整数。

可以使用 placement-strategy 属性（在全局群集选项中）指定放置策略。可用值如下：

default（默认值）

完全不考虑利用率值。根据位置得分分配资源。如果分数相等，资源将均匀分布在节点中。

utilization

在确定节点是否有足够的可用容量来满足资源要求时考虑利用率值。但仍会根据分配给节点的资源数执行负载平衡。

minimal

在确定节点是否有足够的可用容量来满足资源要求时考虑利用率值。尝试将资源集中到尽可能少的节点上（以便在剩余节点上实现节电）。

balanced

在确定节点是否有足够的可用容量来满足资源要求时考虑利用率值。尝试均匀分布资源，从而优化资源性能。

注意：配置资源优先级

可用的放置策略是最佳方法 - 它们不使用复杂的启发式解析程序即可始终实现最佳分配结果。因此，设置资源优先级时应确保首先调度最重要的资源。

例 4.3 负载均衡放置配置示例

以下示例演示了配有四台虚拟机、节点数相等的三节点群集。

```
node node1 utilization memory="4000"
node node2 utilization memory="4000"
node node3 utilization memory="4000"
primitive xenA ocf:heartbeat:Xen utilization memory="3500" \
    meta priority="10"
primitive xenB ocf:heartbeat:Xen utilization memory="2000" \
    meta priority="1"
primitive xenC ocf:heartbeat:Xen utilization memory="2000" \
    meta priority="1"
primitive xenD ocf:heartbeat:Xen utilization memory="1000" \
    meta priority="5"
property placement-strategy="minimal"
```

如果三个节点都处于正常状态，那么资源 xenA 将首先放置到一个节点上，然后是 xenD。xenB 和 xenC 将分配在一起或者其中一个与 xenD 分配在一起。

如果一个节点出现故障，可用的总内存将不足以托管所有资源。将确保分配 xenA，xenD 同样如此。但是，只能再放置剩余资源 xenB 和 xenC 中的一个。由于它们的优先级相同，结果未定。要解决这种不确定性，需要为其中一个资源设置更高的优先级。

4.5 更多信息

<http://clusterlabs.org/>

Pacemaker 主页，随 High Availability Extension 提供的群集资源管理器。

<http://linux-ha.org>

高可用性 Linux 项目的主页。

<http://clusterlabs.org/wiki/Documentation>

CRM 命令行界面：crm 命令行工具介绍。

<http://clusterlabs.org/wiki/Documentation>

Pacemaker 1.0—Configuration Explained（Pacemaker 1.0 - 配置说明）：说明用于配置 Pacemaker 的概念。包含全面而非常详细的信息供参考。

配置和管理群集资源 (GUI)

要配置和管理群集资源，可使用图形用户界面 (Pacemaker GUI) 或 `crm` 命令行实用程序。有关命令行方法，请参见第6章 *配置和管理群集资源（命令行）*（第83页）。

本章介绍了 Pacemaker GUI，并包含配置和管理群集资源时所需的基本任务：创建基本和高级类型的资源（组和克隆资源）、配置约束、指定故障转移节点和故障回复节点、配置资源监视以及手动启动、清理、删除和迁移资源。

通过以下两个包提供 GUI 支持：`pacemaker-mgmt` 包，它包含 GUI 后端（`mgmtd` 守护程序）。它必须安装在要使用 GUI 连接到的所有群集节点上。在要运行 GUI 的任何计算机上，安装 `pacemaker-mgmt-client` 包。

5.1 Pacemaker GUI - 概述

要启动 Pacemaker GUI，请在命令行输入 `crm_gui`。要访问配置和管理选项，需要登录到群集。

5.1.1 连接到群集

注意：用户身份验证

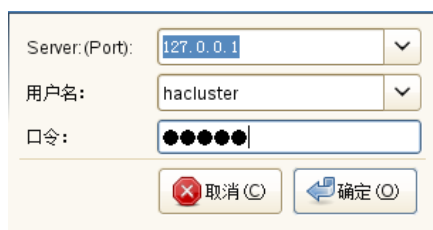
要从 Pacemaker GUI 登录到群集，相应用户必须是 `haclient` 组的成员。安装将创建名为 `hacluster` 的 `linux` 用户，他/她是 `haclient` 组的成员。

使用 Pacemaker GUI 之前，为 `hacluster` 用户设置密码，或创建作为 `haclient` 组成员的新用户。

对要使用 Pacemaker GUI 连接的每个节点执行此操作。

要连接到群集，请选择 *Connection*（连接）> 登录。默认情况下，*Server*（服务器）字段会显示本地主机的 IP 地址，*User Name*（用户名）字段会显示 `hacluster`。输入该用户的密码以继续。

图 5.1 连接群集

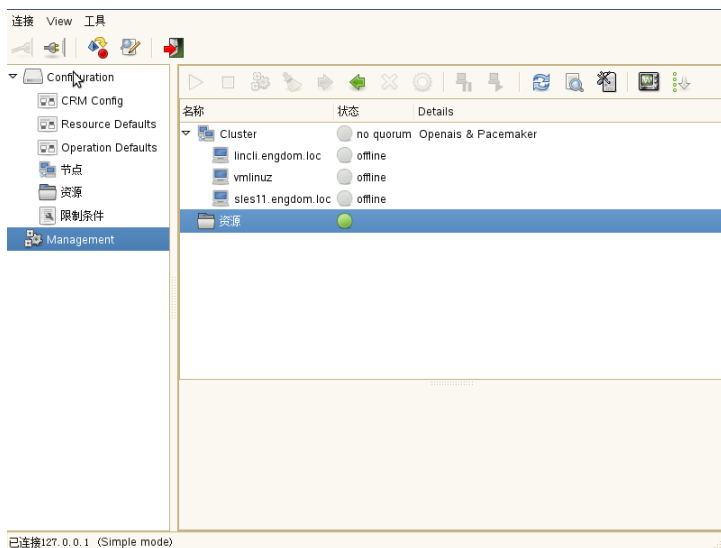
A screenshot of the Pacemaker GUI login dialog box. It features three input fields: 'Server:(Port):' with a dropdown menu showing '127.0.0.1', '用户名:' with a dropdown menu showing 'hacluster', and '口令:' with a password field containing six black dots. At the bottom, there are two buttons: '取消 (C)' (Cancel) with a red 'X' icon and '确定 (O)' (OK) with a blue arrow icon.

如果是远程运行 Pacemaker GUI，请为 *Server*（服务器）字段输入群集节点的 IP 地址。对于 *用户名*，您还可以使用属于 `haclient` 组的任何其他用户连接到群集。

5.1.2 主窗口

连接后，系统会打开主窗口：

图 5.2 Pacemaker GUI - 主窗口



要查看或修改 CRM、资源、节点或约束之类的群集组件，可在左窗格中选择配置类别的相应子条目，然后使用右窗格中的可用选项。此外，使用 Pacemaker GUI 可轻松查看、编辑、导入和导出以下子项 CIB 的 XML 片段：资源默认值、操作默认值、节点、资源和约束。选择任意配置子项，然后在窗口右上角选择显示 > XML 模式。

如果已配置资源，请单击左窗格中的管理类别以显示群集及其资源的状态。使用此视图还可将节点设置为 standby 以及修改节点的管理状态（它们当前是否由群集管理）。要访问资源的主要功能（启动、停止、清理或迁移资源），可在右窗格中选择资源并使用工具栏中的图标。或者右键单击资源并从上下文菜单中选择相应的菜单项。

使用 Pacemaker GUI 还可在不同视图模式间切换，从而影响软件行为以及隐藏或显示特定方面的信息：

简单模式

可用于以类似于向导的模式添加资源。创建和修改资源时，显示子对象的常用选项卡，使您能够通过选项卡直接添加此类型的对象。

可通过在左窗格中选择 *CRM 配置* 项查看和更改所有可用的全局群集选项。右窗格随即显示其当前设置的值。如果没有为选项设置任何特定值，它将显示默认值。

专家方式

可用于以类似于向导的模式或通过对话框窗口添加资源。创建和修改资源时，如果 *CIB* 中已存在特定的子对象类型，将仅显示相应的选项卡。添加新的子对象时，系统将提示您选择对象类型，从而可添加所有受支持的子对象类型。

在左窗格中选择 *CRM 配置* 项时，仅显示实际已设置的全局群集选项的值。隐藏将自动使用默认值的所有群集选项（因为未设置任何值）。在此模式中，只能使用各个配置对话框修改全局群集选项。

黑客模式

与专家模式具有相同功能。可用于添加包括特定规则的其他属性集，以使配置更加动态。例如，可以使节点根据承载它的节点而具有不同的实例属性。此外，还可以为元属性集添加基于时间的规则，以确定属性的生效时间。

窗口的状态栏还会显示当前的活动模式。

以下各节将指引您完成配置群集选项和资源时需要执行的主要任务，并介绍如何使用 *Pacemaker GUI* 管理资源。除非另有说明，否则逐步指示信息反映了在简单模式下执行的过程。

5.2 配置全局群集选项

全局群集选项控制群集在遇到特定情况时的行为方式。它们组成集合，并可使用 *Pacemaker GUI* 和 *crm* 外壳之类的群集管理工具进行查看和修改。在大多数情况下可保留预定义值。但为了使群集的关键功能正常工作，需要在进行基本群集设置后调整以下参数：

- 选项 *no-quorum-policy*（第 34 页）
- 选项 *stonith-enabled*（第 35 页）

过程 5.1 修改全局群集选项

- 1 按第 5.1.1 节“连接到群集”（第 54 页）中所述，启动 Pacemaker GUI 并登录到群集。
- 2 选择 **视图 > 简单模式**。
- 3 在左窗格中，选择 **CRM 配置** 可查看全局群集选项及其当前值。
- 4 根据群集要求，将 **无仲裁人数策略** 设置为适当值。
- 5 如果由于某些原因需要禁用屏蔽，请取消选择 **Stonith Enabled**。
- 6 单击 **应用** 确认更改。

通过在左窗格中选择 **CRM 配置** 并单击 **默认值** 可随时切换回所有选项的默认值。

5.3 配置群集资源

作为群集管理员，您需要在群集中为服务器上运行的每个资源或应用程序创建群集资源。群集资源可以包括 Web 站点、电子邮件服务器、数据库、文件系统、虚拟机和任何其他基于服务器的应用程序或在任意时间对用户都可用的服务。

有关可创建的资源类型的概述，请参阅第 4.2.3 节“资源类型”（第 37 页）。

5.3.1 创建简单群集资源

要创建最基本的资源类型，请按如下操作：

过程 5.2 添加原始资源

- 1 按第 5.1.1 节“连接到群集”（第 54 页）中所述，启动 Pacemaker GUI 并登录到群集。
- 2 在左窗格中，选择 **资源** 并单击 **添加 > Primitive（原始）**。
- 3 在下一个对话框中，为资源设置以下参数：

- 3a** 为资源输入唯一的 *ID*。
- 3b** 从 *Class*（类）列表中，选择要用于该资源的资源代理类：*heartbeat*、*lsb*、*ocf* 或 *stonith*。有关详细信息，参见第 4.2.2 节“支持的资源代理类”（第 36 页）。
- 3c** 如果选择了 *ocf* 作为类，请同时指定 OCF 资源代理的 *Provider*（提供程序）。OCF 规范允许多个供应商供应相同的资源代理。
- 3d** 从 *Type*（类型）列表中，选择要使用的资源代理（例如 *IPaddr* 或 *Filesystem*）。该资源代理的简短描述显示在下方。
- Type*（类型）列表中提供的选项取决于您选择的 *Class*（类）（对于 OCF 资源还取决于 *Provider*（提供程序）中选择的内容）。
- 3e** 在 *Options*（选项）下面，设置 *Initial state of resource*（资源的初始状态）。
- 3f** 如果希望群集监视资源状况是否仍然正常，请激活 *Add monitor operation*（添加监视操作）。

添加 普通资源 - Basic Settings

必填

标识符:

my_primitive

类别:

ocf

提供者:

heartbeat

类型:

IPaddr

描述

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:

Stopped

☒

Add monitor operation

取消(C)

前进(E)

- 4 单击 *Forward*（前进）。下一个窗口将显示已为该资源定义的参数摘要。系统会列出该资源的所有必需的 *Instance Attributes*（实例属性）。若要设置相应的值，需要对实例属性进行编辑。可能还需要添加更多的属性，具体取决于您的部署和设置。有关如何操作的细节，请参考过程 5.3, “添加或修改元属性和实例属性”（第 59 页）。
- 5 如果所有参数都按您的需要进行了设置，请单击 *应用* 完成该资源的配置。配置对话框关闭，主窗口显示新添加的资源。

在创建资源的过程中或创建资源后，可以添加或修改资源的以下参数：

- Instance attributes - 确定资源控制的服务实例。有关更多信息，请参考第 4.2.6 节 “实例属性”（第 42 页）。
- Meta attributes - 告知 CRM 如何处理特定资源。有关更多信息，请参考第 4.2.5 节 “资源选项（元属性）”（第 41 页）。
- Operations - 资源监视需要用到它们。有关更多信息，请参考第 4.2.7 节 “资源操作”（第 43 页）。

过程 5.3 添加或修改元属性和实例属性

- 1 在 Pacemaker GUI 主窗口中，单击左窗格中的资源以查看已配置的群集资源。
- 2 在右窗格中，选择要修改的资源并单击 *Edit*（编辑）（或双击资源）。下一个窗口将显示已为该资源定义的基本资源参数和 *Meta Attributes*（元属性）、*Instance Attributes*（实例属性）或 *Operations*（操作）。

Show: List Mode

必填
 标识符: my_primitive
 类别: ocf
 提供者: heartbeat
 类型: IPAddr

▸ 选填

描述
 Manages virtual IPv4 addresses.
 This script manages IP alias IP addresses
 It can add an IP alias, or remove one.

Meta Attributes Instance Attributes **操作**

名称	值
ip	192.168.8.212

标识符: nvpair-e2f36987-795f459c-b445-7a3d7ba1924f
 名称: ip
 值: 192.168.8.212

添加 (+) 编辑 (E) 删除 (D)

取消 (C) 重置 确定 (O)

- 要添加新的元属性或实例属性，请选择相应的选项卡并单击 *Add*（添加）。
- 选择要添加的属性名称。将显示简短的 *描述*。
- 如果需要，请指定属性 *值*。否则，使用该属性的默认值。
- 单击 *OK*（确定）确认更改。新添加或新修改的属性便显示在选项卡上。
- 如果所有参数都按您的需要进行了设置，请单击 *OK*（确定）完成该资源的配置。配置对话框关闭，主窗口显示已修改的资源。

提示：资源的 XML 源代码

使用 Pacemaker GUI 可查看从已定义的参数生成的 XML 片段。对于单独资源，可在资源配置对话框右上角选择显示 > *XML 模式*。

要访问已配置的所有资源的 XML 表示，请在左窗格中单击资源，然后在主窗口右上角选择显示 > XML 模式。

编辑器将显示 XML 代码，允许您 *Import*（导入）或 *Export*（导出）XML 元素或手动编辑 XML 代码。

5.3.2 创建 STONITH 资源

要配置屏障，需要配置一个或多个 STONITH 资源。

过程 5.4 添加 STONITH 资源

- 1 按第 5.1.1 节“连接到群集”（第 54 页）所述，启动 Pacemaker GUI 并登录到群集。
- 2 在左窗格中，选择资源并单击添加 > *Primitive*（原始）。
- 3 在下一个对话框中，为资源设置以下参数：
 - 3a 为资源输入唯一的 *ID*。
 - 3b 从 *Class*（类）列表，选择资源代理类 *stonith*。
 - 3c 从 *Type*（类型）列表，选择 STONITH 插件以控制 STONITH 设备。该插件的简短描述显示在下方。
 - 3d 在 *Options*（选项）下面，设置 *Initial state of resource*（资源的初始状态）。
 - 3e 如果希望群集监视屏障设备，请激活 *Add monitor operation*（添加监视操作）。有关更多信息，请参考第 9.4 节“监视屏障设备”（第 120 页）。
- 4 单击 *Forward*（前进）。下一个窗口将显示已为该资源定义的参数摘要。系统会列出所选 STONITH 插件的所有必需的实例属性。若要设置相应的值，需要对实例属性进行编辑。可能还需要添加更多的属性或监视操作，具体取决于您的部署和设置。有关如何操作的细节，请参考过程 5.3，“添加或修改元属性和实例属性”（第 59 页）和第 5.3.7 节“配置资源监视”（第 70 页）。

- 5 如果所有参数都按您的需要进行了设置，请单击 *Apply*（应用）完成该资源的配置。配置对话框关闭，主窗口显示新添加的资源。

要完成屏障配置，请添加约束和/或使用克隆。有关详细信息，请参考第9章 屏障和 *STONITH*（第 113 页）。

5.3.3 配置资源约束

配置好所有资源只是完成了该任务的一部分。即便群集熟悉所有必需资源，它可能还无法进行正确处理。资源约束允许您指定在哪些群集节点上运行资源、以何种顺序装载资源，以及特定资源依赖于哪些其他资源。

有关可用约束类型的概述，请参阅第 4.4.1 节“约束类型”（第 46 页）。定义约束时，还需要指定分数。有关分数及其在群集中的含义的更多信息，请参见第 4.4.2 节“分数和无限值”（第 46 页）。

通过以下过程了解如何创建不同类型的约束。

过程 5.5 添加或修改位置约束

- 1 按第 5.1.1 节“连接到群集”（第 54 页）中所述，启动 Pacemaker GUI 并登录到群集。
- 2 在 Pacemaker GUI 主窗口中，单击左窗格中的 *Constraints*（约束）以查看群集已配置的约束。
- 3 在左窗格中，选择 *Constraints*（约束）并单击 *Add*（添加）。
- 4 选择 *Resource Location*（资源位置）并单击 *OK*（确定）。
- 5 为约束输入唯一的 *ID*。修改现有约束时，*ID* 已经定义并显示在配置对话框中。
- 6 选择要定义约束的资源。列表中显示群集已配置的所有资源的 *ID*。
- 7 设置约束的分数。正值表示资源可以在以下指定的节点上运行。负值表示资源不能在此节点上运行。值 *+/- INFINITY* 则由“可以”变为 *must*。
- 8 选择约束的节点。



- 9 如果将 *Node*（节点）和 *Score*（分数）字段留为空白，也可以通过单击 *Add*（添加）> *Rule*（规则）来添加规则。要添加有效期，只需单击 *Add*（添加）> *Lifetime*（有效期）即可。
- 10 如果所有参数都按您的需要进行了设置，请单击 *OK*（确定）完成约束配置。配置对话框关闭，主窗口显示新添加或新修改的约束。

过程 5.6 添加或修改排列约束

- 1 按第 5.1.1 节“连接到群集”（第 54 页）中所述，启动 Pacemaker GUI 并登录到群集。
- 2 在 Pacemaker GUI 主窗口中，单击左窗格中的 *Constraints*（约束）以查看群集已配置的约束。
- 3 在左窗格中，选择 *Constraints*（约束）并单击 *Add*（添加）。
- 4 选择 *Resource Collocation*（资源排列）并单击 *OK*（确定）。
- 5 为约束输入唯一的 *ID*。修改现有约束时，*ID* 已经定义并显示在配置对话框中。
- 6 选择要作为排列源的资源。列表中显示群集已配置的所有资源的 *ID*。

如果约束无法满足，群集可以决定根本不允许运行资源。

- 7 如果将 *Resource*（资源）和 *With Resource*（排列资源）字段都保留为空，也可以通过单击 *Add*（添加）> *Resource Set*（资源集）来添加资

源集。要添加有效期，只需单击 *Add*（添加） > （有效期） *Lifetime* 即可。

- 8 在 *With Resource*（排列资源）中，定义排列目标。群集先决定将此资源放置在什么位置，再决定将此资源放置在 *Resource*（资源）字段的什么位置。
- 9 定义 *Score*（分数）可确定两个资源的位置关系。正值表示两个资源应在相同的节点上运行。负值表示两个资源不应在相同的节点上运行。值 $\pm \text{INFINITY}$ 则由 *should* 变为 *must*。分数将与其他因数结合使用，以确定放置资源的位置。
- 10 如果需要，请指定进一步参数，如 *Resource Role*（资源角色）。

根据选择的参数和选项，显示简短描述，解释您正在配置的排列约束的效果。

- 11 如果所有参数都按您的需要进行了设置，请单击 *OK*（确定）完成约束的配置。配置对话框关闭，主窗口显示新添加或新修改的约束。

过程 5.7 添加或修改顺序约束

- 1 按第 5.1.1 节“连接到群集”（第 54 页）中所述，启动 Pacemaker GUI 并登录到群集。
- 2 在 Pacemaker GUI 主窗口中，单击左窗格中的 *Constraints*（约束）以查看群集已配置的约束。
- 3 在左窗格中，选择 *Constraints*（约束）并单击 *Add*（添加）。
- 4 选择 *Resource Order*（资源顺序）并单击 *OK*（确定）。
- 5 为约束输入唯一的 *ID*。修改现有约束时，*ID* 已经定义并显示在配置对话框中。
- 6 使用 *首先* 定义必须先于使用 *然后* 指定的资源启动的资源。
- 7 使用 *Then*（然后）定义必须后于 *First*（首先）资源启动的资源。

根据选择的参数和选项，显示简短描述，解释您正在配置的顺序约束的效果。

8 如果需要，定义更多参数，例如：

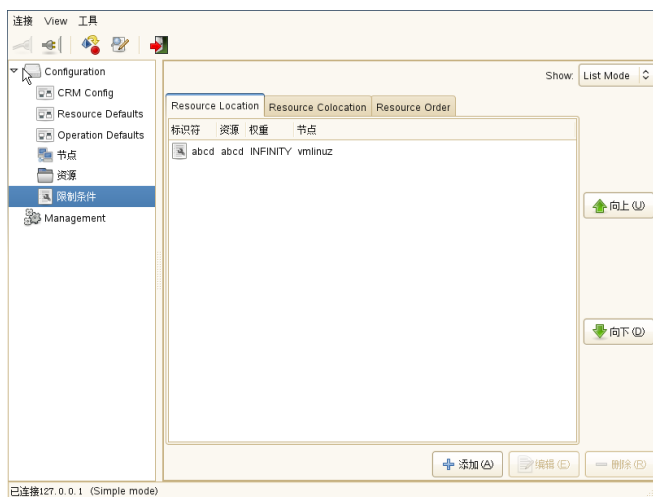
8a 指定分数。如果分数大于零，则约束为强制性的，否则只是建议。默认值为 INFINITY。

8b 为对称指定值。如果为 true，则资源将按相反顺序停止。默认值为 true。

9 如果所有参数都按您的需要进行了设置，请单击 **OK**（确定）完成约束配置。配置对话框关闭，主窗口显示新添加或新修改的约束。

您可以访问和修改在 Pacemaker GUI 的 *Constraints*（约束）视图下配置的所有约束。

图 5.3 Pacemaker GUI - 约束



5.3.4 指定资源故障转移节点

资源在出现故障时会自动重启动。如果在当前节点上无法实现重启动，或如果在当前节点上发生 N 次故障，则资源会试图故障转移到其他节点。您可以多次定义资源的故障次数（migration-threshold），在该值之后资源会迁移到新节点。如果群集中存在两个以上的节点，特定资源故障转移的节点由 High Availability 软件选择。

但您可以按如下操作指定资源将故障转移到的节点：

- 1 按过程 5.5, “添加或修改位置约束”（第 62 页）中所述，为该资源配置位置约束。
- 2 按过程 5.3, “添加或修改元属性和实例属性”（第 59 页）中所述，为该资源添加 `migration-threshold` 元属性，并输入迁移阈值的值。值应该是小于 INFINITY 的正值。
- 3 如果希望资源的故障计数自动失效，请按过程 5.3, “添加或修改元属性和实例属性”（第 59 页）中所述为该资源添加 `failure-timeout` 元属性，并输入故障超时的值。
- 4 如果希望为资源指定更多的首选故障转移节点，请创建更多的位置约束。

有关群集中与迁移阈值和失败计数相关的进程流示例，请参见例 4.2 “迁移阈值 - 进程流”（第 47 页）。

您可以随时手动清理资源的失败计数，而不是让资源的失败计数自动失效。有关细节，请参阅第 5.4.2 节 “清理资源”（第 78 页）。

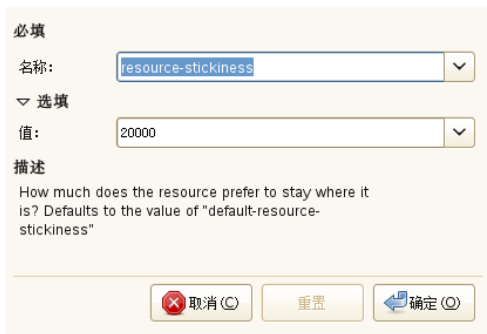
5.3.5 指定资源故障回复节点（资源黏性）

当原始节点恢复联机并位于群集中时，资源可能会故障回复到该节点。如果要防止资源在故障转移前故障回复到之前运行的节点，或者要指定此资源故障回复到的其他节点，必须更改其资源粘性值。可以在创建资源时指定资源粘性或稍后指定。

有关不同资源粘性值的含义，请参阅第 4.4.4 节 “故障回复节点”（第 48 页）。

过程 5.8 指定资源黏性

- 1 按过程 5.3, “添加或修改元属性和实例属性”（第 59 页）中所述为资源添加 resource-stickness 元属性。



必填

名称: resource-stickness

选填

值: 20000

描述

How much does the resource prefer to stay where it is? Defaults to the value of "default-resource-stickness"

取消 重置 确定

- 2 在资源黏性值中，指定介于 $-\text{INFINITY}$ 和 INFINITY 之间的值。

5.3.6 根据负载影响配置资源放置

并非所有资源都相等。某些资源（如 Xen guest）需要托管它们的节点满足其容量要求。如果所放置资源的总需求超过了提供的容量，则资源性能将降低（或甚至失败）。

要考虑此情况，可使用 High Availability Extension 指定以下参数：

1. 特定节点提供的容量。
2. 特定资源需要的容量。
3. 资源放置整体策略。

有关参数的详细背景信息和配置示例，请参阅第 4.4.5 节“根据资源负载影响放置资源”（第 49 页）。

要配置资源要求和节点提供的容量，请按过程 5.9, “添加或修改利用率属性”（第 68 页）中所述使用利用率属性。可根据个人喜好命名利用率属性，并根据配置需要定义多个名称/值对。

过程 5.9 添加或修改利用率属性

在下例中，我们假定您已有群集节点和资源的基本配置，现在想要配置特定节点提供的容量以及特定资源需要的容量。添加利用率属性的过程基本相同，仅步骤 2（第 68 页）和步骤 3（第 68 页）有所不同。

- 1 按第 5.1.1 节“连接到群集”（第 54 页）中所述，启动 Pacemaker GUI 并登录到群集。
- 2 指定节点提供的容量：
 - 2a 在左窗格中，单击节点。
 - 2b 在右窗格中，选择要配置其容量的节点，然后单击编辑。
- 3 指定资源需要的容量：
 - 3a 在左窗格中，单击资源。
 - 3b 在右窗格中，选择要配置其容量的资源，然后单击编辑。
- 4 选择利用率选项卡，然后单击添加以添加利用率属性。
- 5 为新属性输入名称。可以根据个人喜好命名利用率属性。
- 6 为属性输入值，然后单击确定。属性值必须是整数。
- 7 如果需要更多利用率属性，请重复步骤 5（第 68 页）到步骤 6（第 68 页）。

利用率选项卡显示已为此节点或资源定义的利用率属性的摘要。
- 8 根据意愿设置所有参数后，单击确定关闭配置对话框。

图 5.4 “节点容量配置示例”（第 69 页）显示了将向运行在此节点上的资源提供 8 个 CPU 单元和 16 GB 内存的节点配置：

图 5.4 节点容量配置示例

Show: List Mode

Required

ID: bourbaki

Uname: bourbaki

Type: normal

Optional

Instance Attributes

Utilization

Name	Value
cpu	8
memory	16384

Up

Down

ID: nodes-bourbaki-cpu

Name: cpu

Value: 8

Add

Edit

Remove

Cancel

Reset

OK

需要使用节点 4096 个内存单元和 4 个 CPU 单元的资源配置示例如下所示：

图 5.5 资源容量配置示例

Show: List Mode

Required

ID: xen1

Class: ocf

Provider: heartbeat

Type: Xen

Optional

Description

Manages Xen unprivileged domains (DomUs).

Resource Agent for the Xen Hypervisor.

Manages Xen virtual machine instances by managing cluster

Meta Attributes

Instance Attributes

Operations

Utilization

Name	Value
cpu	4
memory	4096

Up

Down

ID: xen1-utilization-cpu

Name: cpu

Value: 4

Add

Edit

Remove

Cancel

Reset

OK

配置了节点提供的容量和资源需要的容量后，需要设置全局群集选项中的放置策略，否则容量配置将不会生效。可使用多个策略来调度负载：例如，可以将负载集中到尽可能少的节点上，或使其均匀分布在所有可用节点上。有关更多信息，请参考第 4.4.5 节“根据资源负载影响放置资源”（第 49 页）。

过程 5.10 设置放置策略

- 1 按第 5.1.1 节“连接到群集”（第 54 页）中所述，启动 Pacemaker GUI 并登录到群集。
- 2 选择 *视图 > 简单模式*。
- 3 在左窗格中，选择 *CRM 配置* 可查看全局群集选项及其当前值。
- 4 根据要求，将 *放置策略* 设置为适当值。
- 5 如果由于某些原因需要禁用屏蔽，请取消选择 *Stonith Enabled*。
- 6 单击 *应用* 确认更改。

5.3.7 配置资源监视

虽然 High Availability Extension 可以检测节点故障，但也能够检测节点上的各个资源何时发生故障。如果要确保资源正在运行，必须为其配置资源监视。资源监视包括指定超时和/或启动延迟值以及间隔。间隔告诉 CRM 检查资源状态的频率。您还可以设置特定参数，如为 *start* 或 *stop* 操作设置 *timeout*。

过程 5.11 添加或修改监视操作

- 1 按第 5.1.1 节“连接到群集”（第 54 页）中所述，启动 Pacemaker GUI 并登录到群集。
- 2 在 Pacemaker GUI 主窗口，单击左窗格中的 *Resources*（资源）以查看群集已配置的资源。
- 3 在右窗格中，选择要修改的资源并单击 *Edit*（编辑）。下一个窗口将显示为该资源定义的基本资源参数、元属性、实例属性和操作。
- 4 要添加新的监视操作，请选择各自选项卡并单击 *Add*（添加）。

要修改现有操作，请选择各自条目并单击 *Edit*（编辑）。
- 5 在 *Name*（名称）中，选择要执行的操作，例如 *monitor*、*start* 或 *stop*。

如下所示的参数取决于您在此处所作的选择。



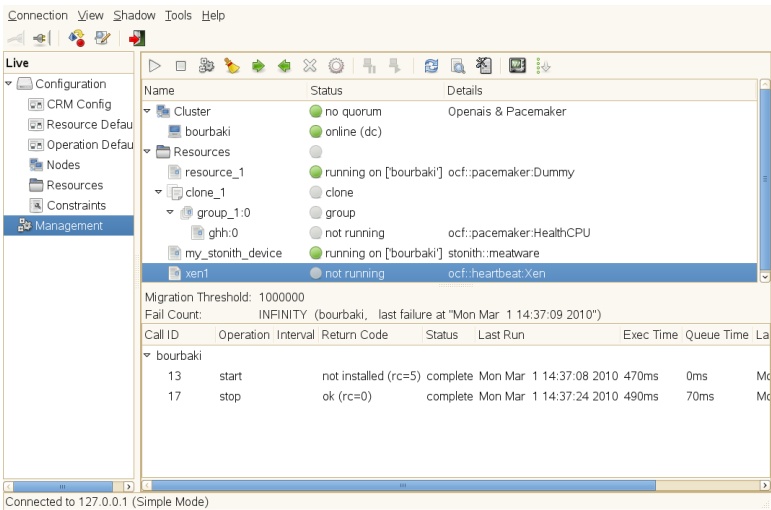
标识符: my_primitive-op-monitor-5s
名称: monitor
间隔: 5s
超时: 20s
▷ 选填
+ 添加 (+) ↻ 编辑 (E) - 删除 (D)
✕ 取消 (C) 重置 ↵ 确定 (O)

- 6 在 *timeout*（超时）字段中，输入以秒表示的值。在指定的超时期间后，操作会被视为 *failed*。PE 会决定如何做或执行您在监视操作的 *On Fail*（失败时）字段中指定的操作。
- 7 如果需要，可展开 *可选部分* 并添加参数，如 *失败时*（如果此操作失败将执行什么操作？）或 *Requires*（要求）（发生此操作前需要满足哪些条件？）。
- 8 如果所有参数都按您的需要进行了设置，请单击 *OK*（确定）完成该资源的配置。配置对话框关闭，主窗口显示已修改的资源。

有关在资源监视程序检测到故障时将发生的操作，请参阅第 4.3 节“资源监控”（第 45 页）。

要在 Pacemaker GUI 中查看资源故障，请在左窗格中单击 *管理*，然后选择要在右窗格中查看其细节的资源。对于失败的资源，右窗格中间将显示资源的失败计数和上次失败信息（在 *迁移阈值* 项下面）。

图 5.6 查看资源的失败计数



5.3.8 配置群集资源组

某些群集资源依赖于其他组件或资源，并且要求每个组件或资源都按特定顺序启动并在同一服务器上一起运行。为了简化此配置，我们支持组的概念。

有关资源组的示例以及组及其属性的更多信息，请参阅“组”一节（第 38 页）。

注意：空组

组必须包含至少一个资源，否则配置无效。

过程 5.12 添加资源组

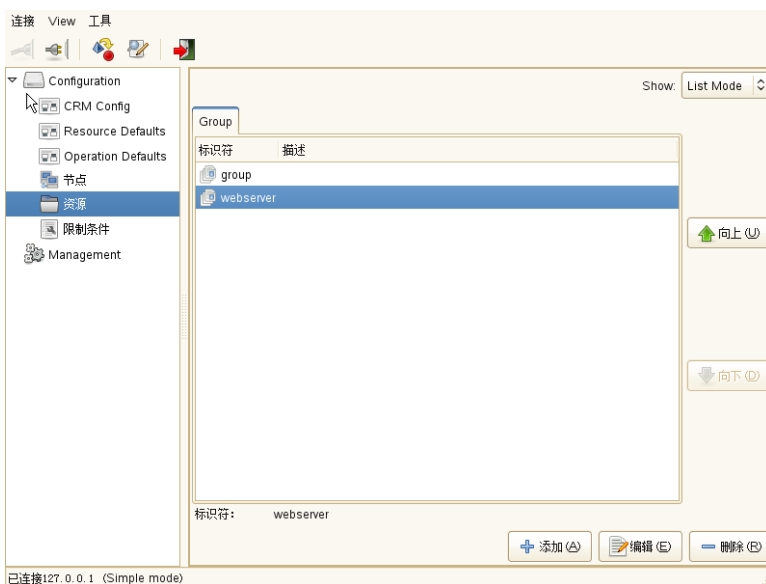
- 按第 5.1.1 节“连接到群集”（第 54 页）中所述，启动 Pacemaker GUI 并登录到群集。
- 在左窗格中，选择 *Resources*（资源）并单击 *Add*（添加）> *Group*（组）。
- 为组输入唯一的 *ID*。

- 4 在 *Options*（选项）下面，设置 *Initial state of resource*（资源的初始状态）并单击 *Forward*（前进）。
- 5 在下一个步骤中，可以添加原始资源作为组的子资源。它们的创建方式与过程 5.2，“添加原始资源”（第 57 页）中描述的步骤类似。
- 6 如果所有参数都按您的需要进行了设置，请单击 *Apply*（应用）完成原始资源的配置。
- 7 在下一个窗口中，可以通过再次选择 *Primitive*（原始）并单击 *OK*（确定）来继续为组添加子资源。

当不希望再向组中添加原始资源时，单击 *Cancel*（取消）。下一个窗口将显示您为该组定义的参数摘要。系统会列出组的 *Meta Attributes*（元属性）和 *Primitives*（原始）资源。资源在 *Primitive*（原始）选项卡上的位置代表资源在群集中的启动顺序。

- 8 由于资源在组中的顺序很重要，可使用*向上*和*向下*按钮对组中的原始资源进行排序。
- 9 如果所有参数都按您的需要进行了设置，请单击 *OK*（确定）完成该组的配置。配置对话框关闭，主窗口显示新创建或新修改的组。

图 5.7 Pacemaker GUI - 组



假定您已按过程 5.12, “添加资源组”（第 72 页）中所述创建资源组。以下过程说明了如何修改组以与例 4.1 “Web 服务器的资源组”（第 38 页）匹配。

过程 5.13 向现有组添加资源

- 1 按第 5.1.1 节“连接到群集”（第 54 页）中所述，启动 Pacemaker GUI 并登录到群集。
- 2 在左窗格中，切换到 *Resources*（资源）视图；在右窗格中，选择要修改的组并单击 *Edit*（编辑）。下一个窗口将显示为该资源定义的基本组参数以及元属性和原始资源。
- 3 单击 *Primitives*（原始）选项卡并单击 *Add*（添加）。
- 4 在下一个对话框中，若要将 IP 地址添加为组的子资源，请设置以下参数：
 - 4a 输入唯一的 *ID*（例如，*my_ipaddress*）。
 - 4b 从 *Class*（类）列表中，选择 *ocf* 作为资源代理类。
 - 4c 在 OCF 资源代理的 *Provider*（提供程序）中，选择 *heartbeat*。

- 4d** 从 *Type*（类型）列表中，选择 *IPaddr* 作为资源代理。
- 4e** 单击 *Forward*（前进）。
- 4f** 在 *Instance Attribute*（实例属性）选项卡中，选择 *IP* 条目并单击 *Edit*（编辑）（或双击 *IP* 条目）。
- 4g** 在 *Value*（值）中输入所需的 IP 地址，例如 192.168.1.1。
- 4h** 单击 *OK*（确定）并单击 *Apply*（应用）。组配置对话框将显示新添加的原始资源。
- 5** 再次单击 *Add*（添加）可添加下一个子资源（文件系统和 Web 服务器）。
- 6** 设置每个子资源各自的参数，其过程类似于从步骤 4a（第 74 页）到步骤 4h（第 75 页）的步骤，直到为组配置完所有子资源为止。



- 由于已按子资源在群集中所需的启动顺序对子资源进行了配置，所以 *Primitives*（原始）选项卡上的顺序已是正确的。
- 7** 如果需要更改组的资源顺序，请使用 *向上* 和 *向下* 按钮对原始资源选项卡上的资源排序。
- 8** 要从组中删除资源，请在 *Primitive*（原始）选项卡上选择资源，并单击 *Remove*（删除）。
- 9** 单击 *OK*（确定）完成该组的配置。配置对话框关闭，主窗口显示修改后的组。

5.3.9 配置克隆资源

您可能希望某些资源在群集的多个节点上同时运行。为此，必须将资源配置为克隆资源。可以配置为克隆资源的资源示例包括 STONITH 和群集文件系统（如 OCFS2）。可以克隆提供的任何资源。资源的资源代理支持此操作。克隆资源的配置甚至也有不同，具体取决于资源驻留的节点。

有关可用的资源克隆类型的概述，请参阅“克隆资源”一节（第 40 页）。

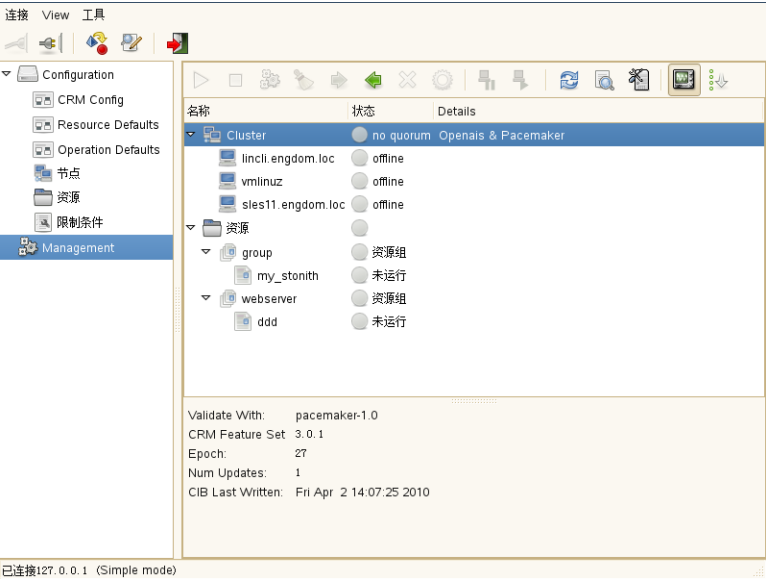
过程 5.14 添加或修改克隆

- 1 按第 5.1.1 节“连接到群集”（第 54 页）中所述，启动 Pacemaker GUI 并登录到群集。
- 2 在左窗格中，选择 *Resources*（资源）并单击 *Add*（添加） > *Clone*（克隆）。
- 3 为克隆资源输入唯一的 *ID*。
- 4 在 *Options*（选项）下面，设置 *Initial state of resource*（资源的初始状态）。
- 5 为克隆资源激活要设置的各个选项，并单击 *Forward*（前进）。
- 6 在下一个步骤中，可以添加 *Primitive*（原始）或 *Group*（组）作为克隆资源的子资源。创建方式与过程 5.2，“添加原始资源”（第 57 页）或过程 5.12，“添加资源组”（第 72 页）中描述的过程类似。
- 7 如果所有参数都按您的需要进行了设置，请单击 *Apply*（应用）完成复制配置。

5.4 管理群集资源

除可用于配置群集资源外，Pacemaker GUI 还可用于管理现有资源。要切换到管理视图并访问可用选项，请在左窗格中单击 *管理*。

图 5.8 Pacemaker GUI - 管理



5.4.1 启动资源

启动群集资源之前，应确保资源设置正确。例如，如果要使用 Apache 服务器作为群集资源，请先设置 Apache 服务器并完成 Apache 配置，然后才能启动群集中的相应资源。

注意：不要处理由群集管理的服务

使用 High Availability Extension 管理资源时，不得以其他方式（在群集外，例如手动或者引导时或重引导时）启动或停止同一资源。High Availability Extension 软件负责所有服务的启动或停止操作。

但是，如果要检查服务是否正确配置，可手动启动该服务，但请确定在 High Availability 接管前再次停止该服务。

要对群集当前管理的资源进行干预，请按第 5.4.5 节“更改资源的管理模式”（第 81 页）中所述先将资源设置为 unmanaged mode。

在使用 Pacemaker GUI 创建资源的过程中，可以使用 `target-role` 元属性设置资源的初始状态。如果其值已设置为 `stopped`，则资源不会在创建后自动启动。

过程 5.15 启动新资源

- 1 按第 5.1.1 节“连接到群集”（第 54 页）中所述，启动 Pacemaker GUI 并登录到群集。
- 2 在左窗格中单击 *管理*。
- 3 在右窗格中，右键单击资源并从上下文菜单中选择 *启动*（或使用工具栏中的 *启动资源* 图标）。

5.4.2 清理资源

如果资源失败，它会自动重新启动，但每次失败都会增加资源的失败计数。要使用 Pacemaker GUI 查看资源的失败计数，请在左窗格中单击 *管理*，然后在右窗格中选择资源。如果资源失败，其失败计数将显示在右窗格中间（*迁移阈值* 项下面）。

如果已为此资源设置 `migration-threshold`，那么一旦失败计数达到迁移阈值，节点将不再能运行此资源。

可自动重设置资源的失败计数（通过设置资源的 `failure-timeout` 选项），也可如下所述手动重设置。

过程 5.16 清理资源

- 1 按第 5.1.1 节“连接到群集”（第 54 页）中所述，启动 Pacemaker GUI 并登录到群集。
- 2 在左窗格中单击 *管理*。
- 3 在右窗格中，右键单击相应资源，然后从上下文菜单中选择 *清理资源*（或使用工具栏中的 *清理资源* 图标）。

这将对指定节点上的指定资源执行命令 `crm_resource -C` 和 `crm_failcount -D`。

有关详细信息，另请参见`crm_resource(8)`（第 217 页）和`crm_failcount(8)`（第 208 页）。

5.4.3 删除群集资源

如果需要从群集中删除资源，请遵循以下过程以免出现配置错误：

注意：删除引用的资源

如果群集资源的 ID 由任何约束引用，则无法删除该群集资源。如果您无法删除某个资源，请检查引用资源 ID 的位置，然后先从该约束删除资源。

过程 5.17 删除群集资源

- 1 按第 5.1.1 节“连接到群集”（第 54 页）中所述，启动 Pacemaker GUI 并登录到群集。
- 2 在左窗格中单击 *管理*。
- 3 在右窗格中选择相应的资源。
- 4 按过程 5.16, “清理资源”（第 78 页）中所述清理所有节点上的资源。
- 5 停止资源。
- 6 删除与资源相关的所有约束，否则将无法删除资源。

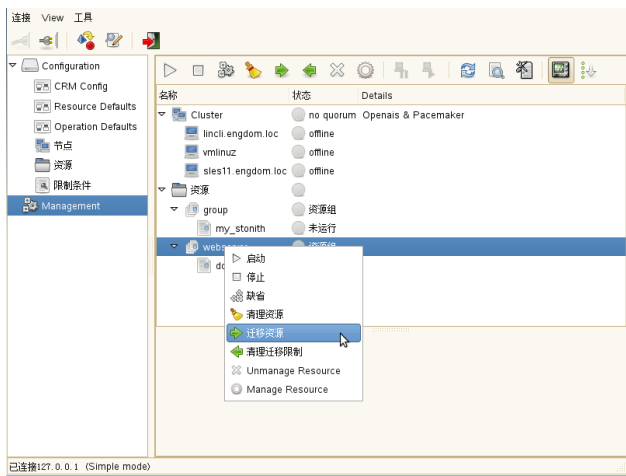
5.4.4 迁移群集资源

如第 5.3.4 节“指定资源故障转移节点”（第 65 页）中所述，如果软件或硬件发生故障，群集会进行资源故障转移（迁移）——根据定义的特定参数（例如，迁移阈值或资源黏性）。除此之外，还可以手动将资源迁移到群集资源中的其他节点。

过程 5.18 手动迁移资源

- 1 按第 5.1.1 节“连接到群集”（第 54 页）中所述，启动 Pacemaker GUI 并登录到群集。

- 2 在左窗格中单击**管理**。
- 3 在右窗格中右键单击相应资源，然后选择**迁移资源**。



- 4 在新窗口的 *To Node*（目标节点）中，选择资源要移动到的节点。此操作将创建目标节点分数为 INFINITY 的位置约束。
- 5 如果只是希望临时迁移资源，请激活 *Duration*（持续时间）并输入资源迁移到新节点的时间范围。在持续时间到期后，资源可以移回到其原始位置，也可以留在当前位置（取决于资源黏性）。
- 6 如果资源无法迁移（若资源在当前节点上的黏性和约束总分数大于 INFINITY），请激活 *Force*（强制）选项。它通过创建当前位置规则和 -INFINITY 的分数强制资源移动。

注意

这将阻止资源在使用清除迁移约束删除约束或持续时间到期之前在此节点上运行。

- 7 单击 **OK**（确定）确认迁移。

要使资源重新移回，请按如下操作：

过程 5.19 清除迁移约束

- 1 按第 5.1.1 节“连接到群集”（第 54 页）中所述，启动 Pacemaker GUI 并登录到群集。
- 2 在左窗格中单击管理。
- 3 在右窗格中右键单击相应资源，然后选择清除迁移约束。

这将使用 `crm_resource -U` 命令。资源可以移回到其原始位置，也可以留在当前位置（取决于资源黏性）。

有关更多信息，请参见 `crm_resource(8)`（第 217 页）或 <http://clusterlabs.org/wiki/Documentation> 上的 *Pacemaker 1.0 - 配置说明*。请参阅 资源迁移一章。

5.4.5 更改资源的管理模式

由群集管理资源时，不得以其他方式（在群集外）处理资源。要维护各个资源，可将相应资源设置为 `unmanaged mode`，在此模式下可在群集外修改资源。

- 1 按第 5.1.1 节“连接到群集”（第 54 页）中所述，启动 Pacemaker GUI 并登录到群集。
- 2 在左窗格中单击管理。
- 3 在右窗格中右键单击相应资源，然后从上下文菜单中选择不受管资源。
- 4 完成此资源的维护任务后，在右窗格中再次右键单击相应资源，然后选择管理资源。

此后，资源将再次由 High Availability Extension 软件管理。

配置和管理群集资源（命令行）

要配置和管理群集资源，可使用图形用户界面 (Pacemaker GUI) 或 `crm` 命令行实用程序。有关 GUI 方法，请参阅第 5 章 *配置和管理群集资源 (GUI)*（第 53 页）。

本章介绍了命令行工具 `crm`，并包含此工具的概述以及如何使用模板，主要介绍如何配置和管理群集资源：创建基本和高级类型的资源（组和克隆资源）、配置约束、指定故障转移节点和故障回复节点、配置资源监视以及手动启动、清理、删除和迁移资源。

6.1 `crm` 命令行工具 - 概述

安装后，通常只需要 `crm` 命令。此命令有多个子命令，这些子命令用于管理资源、CIB、节点和资源代理等。运行 `crm help` 可以获取所有可用命令的概述。它提供了全面的帮助系统，并嵌入了示例。

可按如下方式使用 `crm` 命令：

- **直接** 将所有子命令添加到 `crm` 中，按 **Enter**，您将立即看到输出。例如，输入 `crm help ra` 可获取有关 `ra` 子命令（资源代理）的信息。
- **作为外壳脚本** 使用 `crm` 和包含 `crm` 命令的脚本。这可通过以下两种方式实现：

```
crm -f script.cli  
crm < script.cli
```

脚本可包含任何 `crm` 命令。例如：

```
# A small example
status
node list
```

以哈希符号 (#) 开头的所有行都是注释，可忽略。如果行过长，可在结尾处插入反斜杠 (\)，然后在下一行继续。

- **作为内壳交互** 输入 `crm` 以进入内壳。提示符变为 `crm(live)#`。使用 `help` 可获取可用子命令的概述。由于内壳具有不同的子命令级别，您可以仅输入一个子命令“进入”此级别，然后按 **Enter**。

例如，如果输入 `resource`，则进入资源管理级别。提示符将更改为 `crm(live)resource#`。如果要退出内壳，可使用命令 `quit`、`bye` 或 `exit`。如果需要返回上一个级别，可使用 `up`、`end` 或 `cd`。

可以直接进入级别，方法是输入 `crm`、相应的子命令，不带选项。

内壳还支持使用 `Tab` 键完成子命令和资源。输入命令的开头，按 `→` 和 `crm` 完成相应对象。

注意：管理和配置子命令之间的区别

`crm` 工具有管理功能（子命令 `resource` 和 `node`），可用于配置（`cib` 和 `configure`）。

以下小节概述了 `crm` 工具的一些重要方面。

6.1.1 显示有关 OCF 资源代理的信息

由于在群集配置中必须一直应对资源代理，`crm` 工具包含 `ra` 命令以获取有关资源代理的信息并对其进行管理（有关其他信息，另请参见第 4.2.2 节“支持的资源代理类”（第 36 页））：

```
# crm ra
crm(live)ra#
```

命令 `classes` 可提供所有类和提供程序的列表：

```
crm(live)ra# classes
heartbeat
```



```
lsb
ocf / heartbeat linbit lvm2 ocfs2 pacemaker
stonith
```

要获取有关某个类（和提供程序）的所有可用资源的概述，可使用 `list` 命令：

```
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
...
```

可使用 `info` 查看有关资源代理的概述：

```
crm(live)ra# info ocf:drbd:linbit
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
```

按 **Q** 退出查看器。可在附录 A, 设置简单测试资源的示例（第 337 页）中查找配置示例。

提示：直接使用 **crm**

在之前的示例中，我们使用了 `crm` 命令的内壳。但是您不必使用它。将相应子命令添加到 `crm` 中可以得到相同的结果。例如，在外壳中输入 `crm ra list ocf` 可以列出所有 OCF 资源代理。

6.1.2 使用模板

模板是现成的群集配置。只需稍作更改即可满足特定用户的需要。每次使用模板创建配置时，都会出现警告消息，提示您哪些可以稍后编辑以供来自定义。

以下步骤显示了如何创建简单有效的 Apache 配置：

1 以 `root` 身份登录。

2 启动 `crm` 工具：

```
# crm configure
```

3 从模板创建一个新配置：

3a 切换到 `template` 子命令：

```
crm(live)configure# template
```

3b 列出可用模板：

```
crm(live)configure template# list templates  
gfs2-base    filesystem  virtual-ip  apache      clvm        ocfs2       gfs2
```

3c 确定需要的模板。由于我们需要 Apache 配置，因此选择 `apache` 模板：

```
crm(live)configure template# new intranet apache  
INFO: pulling in template apache  
INFO: pulling in template virtual-ip
```

4 定义参数：

4a 列出刚创建的配置：

```
crm(live)configure template# list
intranet
```

4b 显示必填的最少所需更改：

```
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
```

4c 调用首选的文本编辑器，填写显示为错误（如步骤 4b（第 87 页）中所示）的所有行：

```
crm(live)configure template# edit
```

5 显示配置并检查配置是否有效（粗体文本取决于您在步骤 4c（第 87 页）中进入的配置）：

```
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 intranet \
    apache virtual-ip
```

6 应用配置：

```
crm(live)configure template# apply
crm(live)configure# cd ..
crm(live)configure# show
```

7 将更改提交到 CIB：

```
crm(live)configure# commit
```

如果知道细节，可以更加简化命令。上述过程可汇总为外壳上的以下命令：

```
crm configure template \
    new intranet apache params \
    configfile="/etc/apache2/httpd.conf" ip="192.168.1.101"
```

如果在 crm 内壳中，可使用以下命令：

```
crm(live)configure template# new intranet apache params \
    configfile="/etc/apache2/httpd.conf" ip="192.168.1.101"
```

但是，上面的命令仅从模板创建其配置。它不会将其应用或提交到 CIB。

6.1.3 使用阴影配置进行测试

阴影配置可用于测试不同的配置方案。如果创建了多个阴影配置，则可逐一测试这些配置，以查看更改的影响。

一般的流程显示如下：

- 1 打开壳层并成为 root。

- 2 使用以下命令启动 crm 外壳：

```
crm configure
```

- 3 创建新的阴影配置：

```
crm(live)configure# cib new myNewConfig  
INFO: myNewConfig shadow CIB created
```

- 4 如果要将当前的活动配置复制到阴影配置中，可使用以下命令，否则请跳过此步骤：

```
crm(myNewConfig)# cib reset myNewConfig
```

使用上面的命令便于稍后修改现有资源。

- 5 照常进行更改。创建阴影配置后，会应用所有更改。要保存所有更改，请使用以下命令：

```
crm(myNewConfig)#
```

- 6 如果再次需要活动群集配置，可使用以下命令切换回此配置：

```
crm(myNewConfig)configure# cib use live  
crm(live)#
```

6.1.4 调试配置更改

将配置更改装载回群集之前，建议使用 `ptest` 复查更改。使用 `ptest` 可显示提交更改时将产生的操作图。需要 `graphviz` 包才能显示这些图。以下示例是一个抄本，添加了监视操作：

```
# crm configure
crm(live)configure# show fence-node2
primitive fence-node2 stonith:apcsmart \
    params hostlist="node2"
crm(live)configure# monitor fence-node2 120m:60s
crm(live)configure# show changed
primitive fence-node2 stonith:apcsmart \
    params hostlist="node2" \
    op monitor interval="120m" timeout="60s"
crm(live)configure# ptest
crm(live)configure# commit
```

6.2 配置全局群集选项

全局群集选项控制群集在遇到特定情况时的行为方式。可以使用 `crm` 工具查看和修改这些选项。在大多数情况下可保留预定义值。但为了使群集的关键功能正常工作，需要在进行基本群集设置后调整以下参数：

- 选项 `no-quorum-policy`（第 34 页）
- 选项 `stonith-enabled`（第 35 页）

过程 6.1 使用 `crm` 修改全局群集选项

- 1 打开壳层并成为 `root`。
- 2 输入 `crm configure` 打开内壳。
- 3 使用以下命令仅设置双节点群集的选项：

```
crm(live)configure# property no-quorum-policy=ignore
crm(live)configure# property stonith-enabled=false
```

- 4 显示更改：

```
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="false"
```

5 提交更改并退出：

```
crm(live)configure# commit  
crm(live)configure# exit
```

6.3 配置群集资源

作为群集管理员，您需要在群集中为服务器上运行的每个资源或应用程序创建群集资源。群集资源可以包括 Web 站点、电子邮件服务器、数据库、文件系统、虚拟机和任何其他基于服务器的应用程序或在任意时间对用户都可用的服务。

有关可创建的资源类型的概述，请参阅第 4.2.3 节“资源类型”（第 37 页）。

6.3.1 创建群集资源

有三种 RA（资源代理）类型可用于群集（有关背景信息，请参见第 4.2.2 节“支持的资源代理类”（第 36 页））。要创建群集资源，请使用 `crm` 工具。要将新资源添加到群集，请按如下操作：

- 1 打开壳层并成为 `root`。
- 2 输入 `crm configure` 打开内壳。
- 3 配置原始 IP 地址：

```
crm(live)configure# primitive myIP ocf:heartbeat:IPaddr \  
    params ip=127.0.0.99 op monitor interval=60s
```

上一命令配置了名称为 `myIP` 的“原始资源”。需要选择一个类（此处为 `ocf`）、提供程序（`heartbeat`）和类型（`IPaddr`）。此外，此原始资源还需要其他参数，如 IP 地址。根据设置更改地址。

- 4 显示您所做的更改并进行复查：

```
crm(live)configure# show
```

5 提交更改使其生效:

```
crm(live)configure# commit
```

6.3.2 NFS 服务器的示例配置

要设置 NFS 服务器，需要完成以下操作：

- 1 配置 DRBD。
- 2 设置文件系统资源。
- 3 设置 NFS 服务器并配置 IP 地址。

以下小节介绍了如何实现这些操作。

配置 DRBD

开始 DRBD High Availability配置之前，请手动设置 DRBD 设备。这主要是配置 DRBD 并使其同步。确切步骤如第 13 章 分布式复制块设备 (*DRBD*) (第 147 页) 中所述。现在，假定在两个群集节点上均配置了可从设备 `/dev/drbd_r0` 访问的资源 `r0`。

DRBD 资源是 OCF 主/从资源。这可在 DRBD 资源代理的元数据描述中找到。但是，重要的是元数据的 `actions` 部分中存在操作 `promote` 和 `demote`。它们对于主/从资源是必需的，且通常不可用于其他资源。

对于 High Availability，主/从资源可以在不同节点上具有多个主资源。甚至有可能主资源和从属资源在同一个节点上。因此，请以这样的方式配置此资源：只有一个主资源和一个从属资源，且各自在不同的节点上运行。可使用 `master` 资源的元属性来执行此操作。主/从资源是克隆资源在 High Availability 中的特殊类型。每个主资源和每个从属资源计为一个克隆。

请按如下操作以配置 DRBD 资源：

- 1 打开壳层并成为 `root`。

2 输入 `crm configure` 打开内壳。

3 如果有双节点群集，为每个 `ms` 资源设置以下属性：

```
crm(live)configure# primitive my-stonith stonith:external/ipmi ...
crm(live)configure# ms ms_drbd_r0 res_drbd_r0 meta \
    globally-unique=false ...
crm(live)configure# property no-quorum-policy=ignore
crm(live)configure# property stonith-enabled=true
```

4 创建原始 DRBD 资源：

```
crm(live)configure# primitive drbd_r0 ocf:linbit:drbd params \
    drbd drbd_resource=r0 op monitor interval="30s"
```

5 创建主/从资源：

```
crm(live)configure# ms ms_drbd_r0 res_drbd_r0 meta master-max=1 \
    master-node-max=1 clone-max=2 clone-node-max=1 notify=true
```

6 指定组配和顺序约束：

```
crm(live)configure# colocation fs_on_drbd_r0 inf: res_fs_r0
ms_drbd_r0:Master
crm(live)configure# order fs_after_drbd_r0 inf: ms_drbd_r0:promote
res_fs_r0:start
```

7 使用 `show` 命令显示更改。

8 使用 `commit` 命令提交更改。

设置文件系统资源

`filesystem` 资源已配置为 DRBD 的 OCF 原始资源。它的任务是在收到启动和停止请求时将设备装入和卸载到目录。在本例中，设备为 `/dev/drbd_r0`，要用作安装点的目录为 `/srv/f ilover`。使用的文件系统为 `xfs`。

在 `crm` 外壳中使用以下命令配置文件系统资源：

```
crm(live)# configure
crm(live)configure# primitive filesystem_resource \
    ocf:linbit:drbd \
    params device=/dev/drbd_r0 directory=/srv/failover fstype=xfs
```


NFS 服务器和 IP 地址

要使 NFS 服务器在同一 IP 地址上始终可用，请使用其他 IP 地址以及计算机正常运行时所使用的 IP 地址。除系统的 IP 地址以外，此地址随后会指派到活动的 NFS 服务器。

NFS 服务器和 NFS 服务器的 IP 地址在同一台计算机上应始终是活动的。在这种情况下，启动顺序就不是很重要了。它们甚至可以同时启动。这些是组资源的典型要求。

在启动 High Availability RA 配置之前，先使用 YaST 配置 NFS 服务器。不要让系统启动 NFS 服务器。只需要设置配置文件。如果要手动执行该操作，请参阅手册页导出 `exports(5)` (`man 5 exports`)。配置文件为 `/etc/exports`。将 NFS 服务器配置为 LSB 资源。

通过 High Availability RA 配置可完全配置 IP 地址。系统中不需要任何其他修改。IP 地址 RA 为 OCF RA。

```
crm(live)# configure
crm(live)configure# primitive nfs_resource ocf:nfsserver \
    params nfs_ip=10.10.0.1 nfs_shared_infodir=/shared
crm(live)configure# primitive ip_resource ocf:heartbeat:IPaddr \
    params ip=10.10.0.1
crm(live)configure# group nfs_group nfs_resource ip_resource
crm(live)configure# show
primitive ip_res ocf:heartbeat:IPaddr \
    params ip="192.168.1.10"
primitive nfs_res ocf:heartbeat:nfsserver \
    params nfs_ip="192.168.1.10" nfs_shared_infodir="/shared"
group nfs_group nfs_res ip_res
crm(live)configure# commit
crm(live)configure# end
crm(live)# quit
```

6.3.3 创建 STONITH 资源

从 `crm` 透视图来看，STONITH 设备只是另一种资源。要创建 STONITH 资源，请执行如下步骤：

- 1 打开壳层并成为 `root`。
- 2 输入 `crm` 打开内壳。

3 使用以下命令获取所有 STONITH 类型的列表：

```
crm(live)# ra list stonith
apcmaster                apcsmart                baytech
cyclades                 drac3                  external/drac5
external/hmchttp         external/ibmrsa        external/ibmrsa-telnet
external/ipmi            external/kdumpcheck    external/rackpdu
external/riloe           external/sbd           external/ssh
external/vmware          external/xen0          external/xen0-ha
ibmhmc                  ipmilan               meatware
null                    nw_rpc100s            rcd_serial
rps10                   ssh                   suicide
```

4 从以上列表中选择 STONITH 类型并查看可用的选项列表。使用以下命令：

```
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. If 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.
...
```

5 使用 stonith 类（您在步骤 4 中选择的类型）和相应参数（如果需要）创建 STONITH 资源，例如：

```
crm(live)# configure
crm(live)configure# primitive my-stonith stonith:external/ipmi \
    params hostname="node1"
    ipaddr="192.168.1.221" \
    userid="admin" passwd="secret" \
    op monitor interval=60m timeout=120s
```

6.3.4 配置资源约束

配置所有资源只是任务的一部分。即使群集了解所有需要的资源，它仍然不能正确处理它们。例如，尝试在 drbd 的从节点上不装入文件系统（事实上，这将导致 drbd 出现故障）。定义约束以使这些信息可用于群集。

有关约束的更多信息，请参见第 4.4 节“资源约束”（第 45 页）。

位置约束

每个资源可多次添加此类约束。对于给定资源，将评估所有 location 约束。下面是首选在名为 earth 的节点上将 ID 为 fs1-loc 的资源运行到 100 的简单示例：

```
crm(live)configure# location fs1-loc fs1 100: earth
```

另一个示例是使用 pingd 的位置：

```
crm(live)configure# primitive pingd pingd \  
    params name=pingd dampen=5s multiplier=100 host_list="r1 r2"  
crm(live)configure# location node_pref internal_www \  
    rule 50: #uname eq node1 \  
    rule pingd: defined pingd
```

组合约束

collocation 命令用于定义哪些资源应在相同或不同的主机上运行。

只能设置 +inf 或 -inf 的分数，定义必须始终或不得在相同节点上运行的资源。还可以使用有限分数。在这种情况下，组配将称为*建议*，群集可决定不遵循它们，从而在出现冲突时不停止其他资源。

例如，要使 ID 为 filesystem_resource 和 nfs_group 的资源始终在同一主机上，可使用以下约束：

```
crm(live)configure# colocation nfs_on_filesystem inf: nfs_group  
filesystem_resource
```

对于主从属配置，除在本地运行资源以外，还有必要了解当前节点是否为主节点。

排序约束

有时必需提供资源操作顺序。例如，在设备可用于系统之前，您不能装入文件系统。使用排序约束可在另一个资源满足某个特殊条件之前或之后启动或停止某项服务，如已启动、已停止或已升级到主资源。在 crm 外壳中使用以下命令配置顺序约束：

```
crm(live)configure# order nfs_after_filesystem mandatory: group_nfs
filesystem_resource
```

示例配置约束

本章中使用的示例必须与其他约束结合使用。其中最基本的就是让所有资源在同一台计算机上作为 drbd 资源的主资源运行。drbd 资源必须是主资源，其他资源才能启动。在 DRBD 设备不是主资源时尝试装入 DRBD 只会失败。必须实现以下约束：

- 文件系统必须始终与 DRBD 资源的主资源位于同一节点上。

```
crm(live)configure# colocation filesystem_on_master inf: \
filesystem_resource drbd_resource:Master
```

- NFS 服务器及 IP 地址必须与文件系统位于相同的节点上。

```
crm(live)configure# colocation nfs_with_fs inf: \
nfs_group filesystem_resource
```

- NFS 服务器及 IP 地址在装入文件系统后启动：

```
crm(live)configure# order nfs_second mandatory: \
filesystem_resource:start nfs_group
```

- 必须在 DRBD 资源提升为节点上的主资源后才能在此节点上装入文件系统。

```
crm(live)configure# order drbd_first inf: \
drbd_resource:promote filesystem_resource
```

6.3.5 指定资源故障转移节点

要确定资源故障转移，可使用元属性 migration-threshold。例如：

```
crm(live)configure# location r1-node1 r1 100: node1
```

通常，r1 首选在 node1 上运行。如果失败，将检查 migration-threshold 并与它与故障计数进行比较。如果故障计数 \geq migration-threshold，则会将该资源迁移到具有下一个最佳自选设置的节点。

根据 start-failure-is-fatal 选项，启动失败会将失败计数设置为 inf。停止故障可导致屏障。如果未定义 STONITH，将不会迁移资源。

有关概述，请参阅第 4.4.3 节“故障转移节点”（第 47 页）。

6.3.6 指定资源故障回复节点（资源黏性）

当原始节点恢复联机并位于群集中时，资源可能会故障回复到该节点。如果要防止资源在故障转移前故障回复到之前运行的节点，或者要指定此资源故障回复到的其他节点，必须更改其资源粘性值。可以在创建资源时指定资源粘性或稍后指定。

有关概述，请参阅第 4.4.4 节“故障回复节点”（第 48 页）。

6.3.7 根据负载影响配置资源放置

根据负载影响配置资源放置

并非所有资源都相等。某些资源（如 Xen guest）需要托管它们的节点满足其容量要求。如果所放置资源的总需求超过了提供的容量，则资源性能将降低（或甚至失败）。

要考虑此情况，可使用 High Availability Extension 指定以下参数：

1. 特定节点提供的容量。
2. 特定资源需要的容量。
3. 资源放置整体策略。

有关参数的详细背景信息和配置示例，请参阅第 4.4.5 节“根据资源负载影响放置资源”（第 49 页）。

要配置资源要求和节点提供的容量，请按过程 5.9,“添加或修改利用率属性”（第 68 页）中所述使用利用率属性。可根据个人喜好命名利用率属性，并根据配置需要定义多个名称/值对。

在下例中，我们假定您已有群集节点和资源的基本配置，现在想要配置特定节点提供的容量以及特定资源需要的容量。

过程 6.2 使用 *crm* 添加或修改利用率属性

- 1 使用以下命令启动 *crm* 外壳：

```
crm configure
```

- 2 要指定节点提供的容量，请使用以下命令并将占位符 *NODE_1* 替换为节点名称：

```
crm(live)configure# node NODE_1 utilization memory=16384 cpu=8
```

上例中的这些值将假定 *NODE_1* 向资源提供 16 GB 内存和 8 个 CPU 核心。

- 3 要指定资源需要的容量，请使用：

```
crm(live)configure# primitive xenl ocf:heartbeat:Xen ... \  
utilization memory=4096 cpu=4
```

这会使资源消耗 nodeA 的 4096 个内存单元以及 4 个 cpu 单元。

- 4 使用 *property* 命令配置放置策略：

```
crm(live)configure# property ...
```

有四个值可用于放置策略：

```
propertyplacement-strategy=default
```

根据默认设置，完全不考虑利用率值。根据位置得分分配资源。如果分数相等，资源将均匀分布在节点中。

```
propertyplacement-strategy=utilization
```

在根据节点是否有足够的可用容量来满足资源要求以确定其是否为有效节点时考虑利用率值。但仍会根据分配给节点的资源数执行负载平衡。

```
propertyplacement-strategy=minimal
```

在确定节点是否可以为资源提供服务时考虑利用率值；尝试将资源集中到尽可能少的节点上，从而在剩余节点上实现节电。

```
propertyplacement-strategy=balanced
```

在确定节点是否可以为资源提供服务时考虑利用率值；尝试均匀分布资源，从而优化资源性能。

放置策略是最佳方法，不使用复杂的启发式解析程序即可始终实现最佳分配结果。确保正确设置资源优先级，以便首选调度最重要的资源。

5 退出 crm 外壳之前提交更改：

```
crm(live)configure# commit
```

以下示例演示了配有四台虚拟机、节点数相等的三节点群集：

```
crm(live)configure# node node1 utilization memory="4000"
crm(live)configure# node node2 utilization memory="4000"
crm(live)configure# node node3 utilization memory="4000"
crm(live)configure# primitive xenA ocf:heartbeat:Xen \
    utilization memory="3500" meta priority="10"
crm(live)configure# primitive xenB ocf:heartbeat:Xen \
    utilization memory="2000" meta priority="1"
crm(live)configure# primitive xenC ocf:heartbeat:Xen \
    utilization memory="2000" meta priority="1"
crm(live)configure# primitive xenD ocf:heartbeat:Xen \
    utilization memory="1000" meta priority="5"
crm(live)configure# property placement-strategy="minimal"
```

如果三个节点都处于正常状态，那么 `xenA` 将首先放置到一个节点上，然后是 `xenD`。`xenB` 和 `xenC` 将分配在一起或者其中一个与 `xenD` 分配在一起。

如果一个节点出现故障，可用的总内存将不足以托管所有资源。将确保分配 `xenA`，`xenD` 同样如此；但是，只能再放置 `xenB` 和 `xenC` 中的一个，由于它们的优先级相同，结果未定。要解决这种不确定性，需要为其中一个资源设置更高的优先级。

6.3.8 配置资源监视

要监视资源，有两种可能性：使用 `op` 关键字或 `monitor` 命令定义监视操作。以下示例使用 `op` 关键字配置 `Apache` 资源并每 30 分钟监视一次：

```
crm(live)configure# primitive apache apache \
    params ... \
    op monitor interval=60s timeout=30s
```

同样也可以使用以下方式来实现：

```
crm(live)configure# primitive apache apache \
    params ...
crm(live)configure# monitor apache 60s:30s
```

有关概述，请参阅第 4.3 节“资源监控”（第 45 页）。

6.3.9 配置群集资源组

群集的一个最常见元素是需要放置在一起的一组资源。按顺序启动，并按相反顺序停止。为了简化此配置，我们支持组的概念。以下示例创建了两个原始资源（一个 IP 地址和一个电子邮件资源）：

1 以系统管理员的身份运行 `crm` 命令。提示符更改为 `crm(live)`。

2 配置这两个原始资源：

```
crm(live)# configure
crm(live)configure# primitive Public-IP ocf:IPaddr:heartbeat \
    params ip=1.2.3.4
crm(live)configure# primitive Email lsb:exim
```

3 以正确顺序按其相关标识符对原始资源进行分组：

```
crm(live)configure# group shortcut Public-IP Email
```

有关概述，请参阅“组”一节（第 38 页）。

6.3.10 配置克隆资源

最初将克隆构想成便于启动一个 IP 地址的 N 个实例并使它们分布在群集各处以保持负载平衡的一种方法。它们可用于许多其他用途，包括与 DLM、屏蔽子系统和 OCFS2 集成。您可以克隆资源代理支持的任何资源。

要了解有关克隆资源的更多信息，请参见“克隆资源”一节（第 40 页）。

创建匿名克隆资源

要创建匿名克隆资源，首先要创建一个原始资源，然后使用 `clone` 命令来引用它。执行下列操作：

1 打开壳层并成为 `root`。

2 输入 `crm configure` 打开内壳。

3 配置原始资源，例如：

```
crm(live)configure# primitive Apache lsb:apache
```

4 克隆原始资源：

```
crm(live)configure# clone apache-clone Apache
```

创建有状态/多状态克隆资源

要创建有状态的克隆资源，请先创建原始资源，然后创建主/从资源。

1 打开壳层并成为 root。

2 输入 `crm configure` 打开内壳。

3 配置原始资源。必要时更改时间间隔：

```
crm(live)configure# primitive myRSC ocf:myCorp:myAppl \  
    op monitor interval=60 \  
    op monitor interval=61 role=Master
```

4 创建主从属资源：

```
crm(live)configure# clone apache-clone Apache
```

6.4 管理群集资源

除可用于配置群集资源外，`crm` 工具还可用于管理现有资源。以下小节进行了概述。

6.4.1 启动新的群集资源

要启动新的群集资源，您需要相应的标识符。按如下所示继续：

1 打开外壳，并以 `root` 用户身份登录。

2 输入 `crm` 打开内壳。

3 切换到资源级别：

```
crm(live)# resource
```

4 使用 start 启动资源，然后按 →| 键显示所有已知资源：

```
crm(live)resource# start start ID
```

6.4.2 清理资源

如果资源失败，它会自动重启动，但每次失败都会增加资源的失败计数。如果已为此资源设置 migration-threshold，那么一旦失败计数达到迁移阈值，节点将不再能运行此资源。

1 打开外壳，并以 root 用户身份登录。

2 获取所有资源的列表。

```
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 如果资源正在运行，那么必须先停止。将 RSC 替换为资源名称。

```
crm resource stop RSC
```

例如，如果要停止 DLM 资源，请从 dlm-clvm 资源组将 RSC 替换为 dlm。

4 删除资源本身：

```
crm configure delete ID
```

6.4.3 删除群集资源

要删除群集资源，需要相关标识符。按如下所示继续：

1 打开壳层并成为 root。

- 2 运行以下命令来获取您的资源列表：

```
crm(live)# resource status
```

例如，输出可能类似于以下内容（其中 **myIP** 是资源的相关标识符）：

```
myIP      (ocf::IPaddr:heartbeat) ...
```

- 3 删除具有相关标识符的资源（也暗指 `commit`）：

```
crm(live)# configure delete YOUR_ID
```

- 4 提交更改：

```
crm(live)# configure commit
```

6.4.4 迁移群集资源

虽然资源已配置为在发生硬件或软件故障时自动故障转移（或迁移）到群集的其他节点，您也可以使用 Pacemaker GUI 或命令行将资源手动迁移到群集中的其他节点。

- 1 打开外壳，并以 `root` 用户身份登录。
- 2 输入 `crm` 打开内壳。
- 3 要将名为 `ipaddress1` 的资源迁移到名为 `node2` 的群集节点，请输入：

```
crm(live)# resource
crm(live)resource# migrate ipaddress1 node2
```


使用 Web 界面管理群集资源

除了 `crm` 命令行工具和 Pacemaker GUI 外，High Availability Extension 还附带了 HA Web Konsole，一个用于管理任务的基于 Web 的用户界面。它还可用于从非 Linux 计算机监视和管理 Linux 群集。此外，如果系统未提供或不支持图形用户界面，它还是理想的解决方案。

此 Web 界面包含在 `hawk` 包中。它必须安装在要使用 HA Web Konsole 连接到的所有群集节点上。在要使用 HA Web Konsole 访问群集节点的计算机上，只需启用了 JavaScript 和 cookie 的（图形）Web 浏览器即可建立连接。

注意：用户身份验证

要从 HA Web Konsole 登录到群集，相应用户必须是 `haclient` 组的成员。安装将创建名为 `hacluster` 的 Linux 用户，他/她是 `haclient` 组的成员。

使用 HA Web Konsole 之前，为 `hacluster` 用户设置密码，或创建作为 `haclient` 组成员的新用户。

在将使用 HA Web Konsole 连接到的所有节点上执行此操作。

7.1 启动 HA Web Konsole 并登录

过程 7.1 启动 HA Web Konsole

要使用 HA Web Konsole，必须在要使用 Web 界面连接到的节点上启动相应的 Web 服务。对于通讯，将使用标准 HTTP(s) 协议和端口 7630。

1 在要连接到的节点上，打开外壳并以 root 用户身份登录。

2 通过输入以下命令，检查服务的状态

```
rchawk status
```

3 如果服务未在运行，请使用以下命令启动服务

```
rchawk start
```

如果希望 HA Web Konsole 在引导时自动启动，请执行以下命令：

```
chkconfig hawk on
```

4 在任何计算机上，启动 Web 浏览器并确保 JavaScript 和 cookie 已启用。

5 将其指向任何群集节点的 IP 地址或主机名，或指向可能已配置的任何 IPaddr(2) 资源的地址：

```
https://IPaddress:7630/main/status
```

注意：证书警告

根据浏览器和浏览器选项，在首次尝试访问 URL 时可能会收到证书警告。这是因为 HA Web Konsole 使用默认情况下未被视为可信的自签名证书。

要继续，可在浏览器中添加例外，以绕过警告。要从根本上避免警告，还可将自签名证书替换为官方证书颁发机构签名的证书。有关如何执行此操作的信息，请参阅替换自签名证书（第 108 页）。

6 在 HA Web Konsole 登录屏幕上，输入 hacluster 用户（或作为 haclient 组成员的任何其他用户）的用户名和密码，然后单击登录。

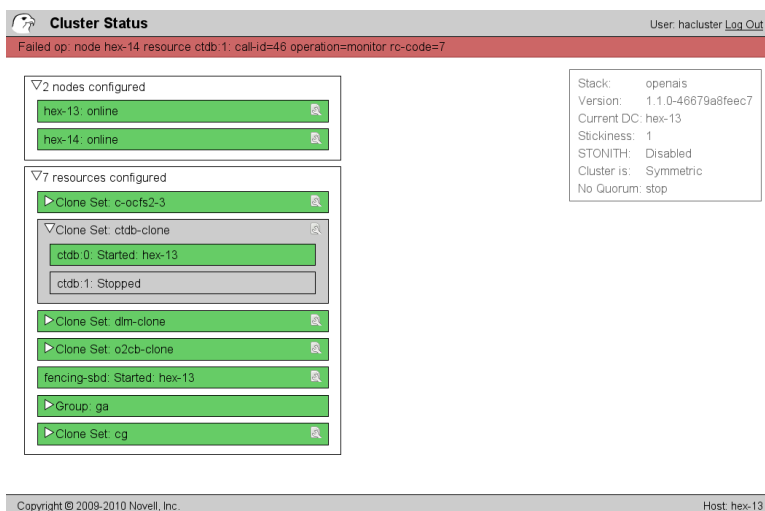
随即出现群集状态屏幕，显示群集节点和资源的状态，类似于 crm_mon 的输出。

7.2 使用 HA Web Konsole

登录后，HA Web Konsole 将显示最重要的全局群集参数以及群集节点和资源的状态。以下颜色代码用于状态显示：

- 绿色：正常。例如，资源正在运行或节点处于联机状态。
- 红色：故障，不正常。例如，资源失败或节点未正常关闭。
- 黄色：正在转换。例如，节点当前正在关闭。
- 灰色：未运行，但群集认为它还在运行。例如，由管理员停止或进入 standby 模式的节点。处于脱机状态的节点也显示为灰色（如果已正常关闭）。

图 7.1 HA Web Konsole - 群集状态



单击节点和资源组中的箭头符号可展开和折叠树视图。

如果资源失败，屏幕顶部将出现一条以红色显示的失败消息，其中包含失败细节。

单击节点或资源右侧的扳手图标，以访问可用于执行某些操作的上下文菜单，这些操作包括启动、停止或清理资源（或使节点进入 online 或 standby 模式或屏蔽某个节点）。

目前，HA Web Konsole 仅可用于执行基本的操作员任务，但以后将添加更多功能，例如配置资源和节点的功能。

7.3 查错

HA Web Konsole 日志文件

在 `/srv/www/hawk/log` 中找到 HA Web Konsole 日志文件。由于某些原因无法访问 HA Web Konsole 时，检查这些日志文件很有用。

如果使用 HA Web Konsole 启动或停止资源时遇到问题，请检查 Pacemaker 将记录到 `/var/log/messages`（默认位置）的日志文件。

身份验证失败

如果不能以添加到 `haclient` 组的新用户身份登录到 HA Web Konsole（或者如果在 HA Web Konsole 接受此用户的登录之前遇到延迟），请使用 `rcnsd stop` 停止 `rcnsd` 守护程序，然后重试。

替换自签名证书

要避免在首次启动 HA Web Konsole 时收到有关自签名证书的警告，请将自动创建的证书替换为您自己的证书或官方证书颁发机构 (CA) 签名的证书。

证书储存在 `/etc/lighttpd/certs/hawk-combined.pem` 中，并包含密钥和证书。创建或接收新密钥和证书后，请执行以下命令将它们合并：

```
cat keyfile certificationfile > /etc/lighttpd/certs/hawk-combined.pem
```

更改许可权限使文件只能由 `root` 访问：

```
chown root.root /etc/lighttpd/certs/hawk-combined.pem  
chmod 600 /etc/lighttpd/certs/hawk-combined.pem
```


添加或修改资源代理

需由群集管理的所有任务都必须可用作资源。在此处需要考虑两个主要组：资源代理和 STONITH 代理。对于这两个类别，您都可以添加自己的代理，根据需要扩展群集的功能。

8.1 STONITH 代理

群集有时会检测到某个节点行为异常，需要删除此节点。这称为屏障，通常使用 STONITH 资源来实现。所有 STONITH 资源都驻留在每个节点上的 `/usr/lib/stonith/plugins` 中。

警告：不支持 SSH 和 STONITH

由于无法了解 SSH 可能对其他系统问题如何做出反应。因此，生产环境不支持 SSH 和 STONITH 代理。

要（从软件端）获取所有当前可用的 STONITH 设备列表，请使用 `stonith -L` 命令。

目前尚无有关写入 STONITH 代理的文档。如果要写入新的 STONITH 代理，请参考 `heartbeat-common` 包的源中提供的示例。

8.2 写入 OCF 资源代理

所有 OCF 资源代理 (RA) 都可在 `/usr/lib/ocf/resource.d/` 中找到, 请参见第 4.2.2 节“支持的资源代理类”(第 36 页) 了解更多信息。每个资源代理都必须支持以下操作才能进行控制:

`start`

启动或启用资源

`stop`

停止或禁用资源

`status`

返回资源状态

`monitor`

与 `status` 类似, 但还会检查是否存在意外状态

`validate`

验证资源配置

`meta-data`

返回有关资源代理的 XML 格式的信息

创建 OCF RA 的常规过程大概如下:

- 1 将文件 `/usr/lib/ocf/resource.d/pacemaker/Dummy` 装载为模板。
- 2 为每个新资源代理创建新的子目录, 以避免发生命名冲突。例如, 如果您的一个资源组 `kitchen` 具有资源 `coffee_machine`, 可将此资源添加到目录 `/usr/lib/ocf/resource.d/kitchen/`。要访问此资源代理, 请执行命令 `crm`:

```
configure
```

```
primitive coffee_1 ocf:coffee_machine:kitchen ...
```

- 3 实施其他外壳功能, 并用不同名称保存文件。

可在 http://linux-ha.org/wiki/Resource_Agents 中找到有关写入 OCF 资源代理的更多细节。在第 1 章 产品概述（第 3 页）中可以找到有关若干概念的特殊信息。

8.3 OCF 返回码和故障恢复

根据 OCF 规范，有一些关于操作必须返回的退出代码的严格定义。群集会始终检查返回代码与预期结果是否相符。如果结果与预期值不匹配，则将操作视为失败，并将启动恢复操作。有三种类型的故障恢复：

表 8.1 故障恢复类型

恢复类型	描述	群集执行的操作
软	发生临时错误。	重新启动资源或将它移到新位置。
硬	发生非临时错误。错误可能特定于当前节点。	将资源移到别处，避免在当前节点上重试该资源。
致命	发生所有群集节点共有的非临时错误。这表示指定了错误配置。	停止资源，避免在任何群集节点上启动该资源。

假定将某个操作视为已失败，下表概括了不同的 OCF 返回代码以及收到相应的错误代码时群集将启动的恢复类型。

表 8.2 OCF 返回代码

OCF 返回代码	OCF 别名	描述	恢复类型
0	OCF_SUCCESS	成功。命令成功完成。这是所有启动、停止、升级和降级命令的所需结果。	软
1	OCF_ERR_GENERIC	通用“出现问题”错误代码。	软

OCF 返回代码	OCF 别名	描述	恢复类型
2	OCF_ERR_ARGS	资源配置在此计算机上无效（例如，它引用了在节点上找不到的位置/工具）。	硬
3	OCF_ERR_UNIMPLEMENTED	请求的操作未实现。	硬
4	OCF_ERR_PERM	资源代理没有足够的特权，不能完成此任务。	硬
5	OCF_ERR_INSTALLED	资源所需的工具未安装在此计算机上。	硬
6	OCF_ERR_CONFIGURED	资源配置无效（例如，缺少必需的参数）。	致命
7	OCF_NOT_RUNNING	<p>资源未运行。群集将不会尝试停止为任何操作返回此代码的资源。</p> <p>此 OCF 返回代码可能需要或不需要资源恢复，这取决于所需的资源状态。如果出现意外，则执行软恢复。</p>	不适用
8	OCF_RUNNING_MASTER	资源正在主节点中运行。	软
9	OCF_FAILED_MASTER	资源在主节点中，但已失败。资源将再次被降级、停止再重新启动（然后也可能升级）。	软
其他	不适用	自定义错误代码。	软

屏障和 STONITH

屏障在 HA（高可用性）计算机群集中是一个非常重要的概念。群集有时会检测到某个节点行为异常，需要删除此节点。这就称为屏障，它通常通过 STONITH 资源完成。屏障可以定义为一种使 HA 群集具有已知状态的方法。

群集中的每个资源均具有状态。例如：“资源 r1 已在 node1 上启动”。在 HA 群集中，这种状态暗示了“资源 r1 在除节点 1 的所有节点上都是停止的”，因为 HA 群集必须确保每个资源最多只能在一个节点上启动。每个节点都必须报告资源发生的每个更改。这样群集状态就是资源状态和节点状态的集合。

如果某些节点或资源的状态无法确定地建立（不论何种原因），就会采取屏障。即使在群集未感知到给定节点上发生的事件时，屏蔽也可确保此节点不会运行任何重要资源。

9.1 屏障分类

有两类屏障：资源级别屏障和节点级别屏障。后者是本章的主题。

资源级别屏障

通过使用资源级屏蔽，群集可确保节点不能访问一个或多个资源。SAN 就是一个典型的示例，屏障操作更改了 SAN 交换机上的规则从而拒绝节点的访问。

通过使用要保护的资源所依赖的常规资源可以完成资源级别屏障。这种资源当然会拒绝在此节点上启动，所以依赖它的资源将不会在同一节点上运行。

节点级别屏障

节点级屏蔽可确保节点绝对不会运行任何资源。这通常以一种非常简单但行之有效的方式执行：使用电源开关重设置节点。这在节点变得无响应时很有必要。

9.2 节点级别屏障

在 SUSE® Linux Enterprise High Availability Extension 中，实现屏障的是 STONITH。它提供节点级别屏障。High Availability Extension 包括 `stonith` 命令行工具，一个能远程关闭群集中节点的可扩展界面。有关可用选项的概述，请运行 `stonith --help` 或参见 `stonith` 的手册页了解更多信息。

9.2.1 STONITH 设备

要使用节点级屏蔽，首先需要有屏蔽设备。要获取 High Availability Extension 支持的 STONITH 设备的列表，请以 `root` 在任何节点上运行以下命令：

```
stonith -L
```

STONITH 设备可分为以下类别：

电源分配单元 (PDU)

电源分发单元是管理关键网络、服务器和数据中心设备的电源容量和功能的基本元素。它可以提供对已连接设备的远程负载监视和独立电源出口控制，以实现远程电源循环。

不间断电源 (UPS)

稳定的电源可在公用电源故障时通过从单独源供电向连接的设备提供应急电源。

刀片电源控制设备

如果是在刀片组上运行群集，则刀片外壳中的电源控制设备就是提供屏障的唯一候选。当然，此设备必须能够管理单个刀片计算机。

无人值守设备

无人值守设备（IBM RSA、HP iLO 和 Dell DRAC）正变得越来越普遍，在未来它们甚至可能成为现成可用计算机上的标准配置。然而，它们相比 UPS 设备有一点不足，因为它们与主机（群集节点）共享一个电源。如果节点没

有电源，则设想为控制节点的设备就等于没用。在这种情况下，CRM 将继续无限期地尝试屏蔽节点，而所有其他资源操作都将等待屏蔽/STONITH 操作完成。

测试设备

测试设备仅用于测试目的。它们通常对硬件更加友好。一旦群集进入生产阶段，它们必须替换为真实的屏障设备。

对 STONITH 设备的选择主要取决于您的预算和所用硬件的种类。

9.2.2 STONITH 实现

SUSE® Linux Enterprise High Availability Extension 的 STONITH 实现由两个组件组成：

stonithd

stonithd 是可由本地进程或通过网络访问的守护程序。它接受与屏蔽操作（重设置、关闭电源和打开电源）对应的命令。它还可以检查屏障设备的状态。

stonithd 守护程序在 CRM HA 群集中的每个节点上运行。在 DC 节点上运行的 stonithd 实例从 CRM 接收屏障请求。它会对请求作出响应，其他 stonithd 程序将执行所需的屏障操作。

STONITH 插件

对于每个受支持的屏蔽设备，都有一个能够控制所述设备的 STONITH 插件。STONITH 插件是屏障设备的界面。所有 STONITH 插件都位于每个节点上的 `/usr/lib/stonith/plugins` 中。所有 STONITH 插件看上去都与 stonithd 一样，但显著区别在于反映了屏障设备的性质。

某些插件支持多个设备。`ipmilan`（或 `external/ipmi`）就是一个典型的示例，它实施 IPMI 协议并可以控制任何支持此协议的设备。

9.3 STONITH 配置

要设置屏蔽，需要配置一个或多个 STONITH 资源 - stonithd 守护程序不需要配置。所有配置都储存在 CIB 中。STONITH 资源就是 stonith 类的资源（请参见第 4.2.2 节“支持的资源代理类”（第 36 页））。STONITH 资源是 STONITH 插件在 CIB 中的代表。除了屏障操作，还可以启动、停止和镜像 STONITH 资

源，就像任何其他资源一样。在这种情况下，启动或停止 STONITH 资源意味着启用或禁用 STONITH。启动和停止仅仅是管理操作，不会转换为屏蔽设备本身上的任何操作。但是，监视会转换成设备状态。

STONITH 资源可像任何其他资源一样进行配置。有关配置资源的更多信息，请参见第 5.3.2 节“创建 STONITH 资源”（第 61 页）或第 6.3.3 节“创建 STONITH 资源”（第 93 页）。

参数（属性）列表取决于相应的 STONITH 类型。要查看特定设备的参数列表，请使用 `stonith` 命令：

```
stonith -t stonith-device-type -n
```

例如，要查看 `ibmhmc` 设备类型的参数，请输入以下命令：

```
stonith -t ibmhmc -n
```

要获取设备的简短帮助文本，请使用 `-h` 选项：

```
stonith -t stonith-device-type -h
```

9.3.1 STONITH 资源配置示例

在以下部分中，可了解一些用 `crm` 命令行工具的语法编写的示例配置。要应用这些配置，请将示例放进文本文件（例如 `sample.txt`）并运行：

```
crm < sample.txt
```

有关使用 `crm` 命令行工具配置资源的更多信息，请参见第 6 章 *配置和管理群集资源（命令行）*（第 83 页）。

警告：测试配置

以下一些示例仅用于演示和测试目的。请勿将任何测试配置示例用于真实的群集方案。

例 9.1 测试配置

```
configure
primitive st-null stonith:null \
params hostlist="node1 node2"
clone fencing st-null
commit
```


例 9.2 测试配置

备用配置：

```
configure
primitive st-node1 stonith:null \
params hostlist="node1"
primitive st-node2 stonith:null \
params hostlist="node2"
location l-st-node1 st-node1 -inf: node1
location l-st-node2 st-node2 -inf: node2
commit
```

考虑到群集软件，此配置示例是完全正确的。与真实配置的唯一区别是没有发生屏障操作。

例 9.3 测试配置

一个更实际的示例（但仍只能用于测试）是以下 `external/ssh` 配置：

```
configure
primitive st-ssh stonith:external/ssh \
params hostlist="node1 node2"
clone fencing st-ssh
commit
```

此配置也可以重设置节点。此配置非常类似于针对空 STONITH 设备的第一个示例。在此示例中使用了克隆。这是 CRM/Pacemaker 的一个功能。克隆基本上是一种快捷方式：无需定义 n 个相同但名称不同的资源，定义单个克隆资源即可。到目前为止，只要可从所有节点访问 STONITH 设备，那么最常用的克隆资源就是 STONITH 资源。

例 9.4 IBM RSA 无人值守设备的配置

真实的设备配置没有太大区别，尽管某些设备可能要求更多属性。可以如下配置 IBM RSA 无人值守设备：

```
configure
primitive st-ibmrsa-1 stonith:external/ibmrsa-telnet \
params nodename=node1 ipaddr=192.168.0.101 \
userid=USERID passwd=PASSWORD
primitive st-ibmrsa-2 stonith:external/ibmrsa-telnet \
params nodename=node2 ipaddr=192.168.0.102 \
userid=USERID passwd=PASSWORD
location l-st-node1 st-ibmrsa-1 -inf: node1
location l-st-node2 st-ibmrsa-2 -inf: node2
commit
```

在本例中，由于以下原因使用了位置约束：STONITH 操作始终有失败的可能性。因此，STONITH 操作（在也是执行程序节点上）不可靠。如果重设置节点，则它将无法发送有关屏障操作结果的通知。唯一的方法是假设操作会成功并提前发送通知。但如果操作失败，还是会出现问题。因此，stonithd 拒绝停止其主机。

例 9.5 UPS 屏障设备的配置

UPS 类型屏蔽设备的配置与上面的示例类似。细节留给读者（作为练习）自行研究。所有 UPS 设备都采用相同的屏蔽结构，但设备本身的访问方式各不相同。旧的 UPS 设备过去只有一个串行端口，在大多数情况下使用特殊的串行电缆以 1200 波特进行连接。许多新的 UPS 设备仍有一个串行端口，但它们通常还使用 USB 或以太网接口。您可使用的连接类型取决于插件所支持的连接类型。

例如，通过使用 `stonith -t stonith 设备类型 -n` 命令比较 `apcmaster` 与 `apcsmart` 设备：

```
stonith -t apcmaster -h
```

返回以下信息：

```
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
```

使用

```
stonith -t apcsmart -h
```

得到以下输出：

```
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
```

第一个插件支持带有一个网络端口的 APC UPS 和 telnet 协议。第二种插件使用通过串行线路的 APC SMART 协议，许多不同的 APC UPS 产品线都支持此协议。

9.3.2 约束与克隆

在第 9.3.1 节“STONITH 资源配置示例”（第 116 页）中，您了解了可使用多种方式配置 STONITH 资源：使用约束和/或克隆。选择何种构造用于配置取决于多个因素（屏蔽设备的性质、设备管理的主机数、群集节点数或个人喜好）。

简而言之：如果将克隆用于配置是安全的且如果它们减少了配置，则使用克隆的 STONITH 资源。

9.4 监视屏障设备

与任何其他资源一样，STONITH 类代理还支持用于检查状态的监视操作。

注意：监视 STONITH 资源

强烈建议监视 STONITH 资源。定期但谨慎地进行监视。

屏蔽设备是 HA 群集不可或缺的一部分，但您越少需要使用它们越好。电源管理设备在通讯方面非常脆弱是众所周知的。如果广播通讯量过大，一些设备会停止工作。某些设备无法处理每分钟多于十个连接的情况。如果两个客户端同时尝试进行连接，一些设备会分辨不清。大多数设备不能同时处理多个会话。

在大多数情况下，每两小时检查一次屏蔽设备应已足够。在这几个小时内需要执行屏障操作及电源开关出现故障的可能性通常很小。

有关如何配置监视操作的详细信息，对于 GUI 方法请参过程 5.3，“添加或修改元属性和实例属性”（第 59 页），对于命令行方法请参见第 6.3.8 节“配置资源监视”（第 99 页）。

9.5 特殊的屏障设备

除了用于处理实际 STONITH 设备的插件外，一些 STONITH 插件还需要附加说明。

警告：仅供测试

下面提到的一些 **STONITH** 插件仅供演示和测试之用。不要在实际情境中使用以下任何设备，因为这可能导致数据损坏和无法预料的结果：

- `external/ssh`
- `ssh`
- `null`

`external/kdumpcheck`

此插件对于检查节点上是否正在进行内核转储很有用。如果正在进行内核转储，它将返回 `true`，如同此节点已屏蔽一样（此时，此节点不能运行任何资源）。这可避免屏蔽已关闭但正在进行转储的节点，从而节省屏蔽所需时间。此插件必须与另一个实际 **STONITH** 设备一同使用。有关更多细节，请参见 `/usr/share/doc/packages/cluster-glue/README_kdumpcheck.txt`。

`external/sbd`

这是一个自屏障设备。它对可以插入共享磁盘的所谓的“毒药”作出反应。在共享储存连接丢失时，它还可使节点停止运行。要了解如何使用此 **STONITH** 代理实施基于储存的屏蔽，请参见第 15 章 **储存保护**（第 169 页）。有关更多细节，另请参见 http://www.linux-ha.org/wiki/SBD_Fencing。

`external/ssh`

另一个基于软件的“屏蔽”机制。节点必须能够以 `root` 用户身份相互登录，且无需密码。它使用一个参数 `hostlist` 指定它将指向的目标节点。由于不能重设置已确实失败的节点，它不得用于实际群集 - 仅供测试和演示之用。将其用于共享储存将导致数据损坏。

`meatware`

`meatware` 需要人为帮助才能运行。调用 `meatware` 时，它会记录一条 **CRIT** 严重性消息，显示在节点的控制台上。然后操作员将确认节点已关闭，并发出 `meatclient(8)` 命令。此命令会告诉 `meatware` 可以通知群集认为此节点已出现故障。有关更多信息，请参见 `/usr/share/doc/packages/cluster-glue/README.meatware`。

`null`

这是一个用于各种测试方案的假设备。它始终以似乎已关闭节点的行为方式运行，并如此声明，但从不执行任何操作。除非您了解您所执行的操作，否则请勿使用它。

`suicide`

这是一个仅有软件的设备，它可以使用 `reboot` 命令重引导它运行所处的节点。这需要节点的操作系统的操作，在某些情况下可能失败。因此，如果可能，请避免使用此设备。然而，在单节点群集上使用此设备是很安全的。

`suicide` 和 `null` 是“do not shoot my host（不关闭我的主机）”规则的唯一例外。

9.6 更多信息

`/usr/share/doc/packages/cluster-glue`

在已安装系统中，此目录包含许多 STONITH 插件和设备的自述文件。

<http://www.linux-ha.org/wiki/STONITH>

高可用性 Linux 项目主页上有关 STONITH 的信息。

http://www.clusterlabs.org/doc/crm_fencing.html

Pacemaker 项目主页上有关屏蔽的信息。

http://www.clusterlabs.org/doc/en-US/Pacemaker/1.0/html/Pacemaker_Explained

说明用于配置 Pacemaker 的概念。包含全面而非常详细的信息供参考。

http://techthoughts.typepad.com/managing_computers/2007/10/split-brain-quo.html

说明 HA 群集中节点分裂、仲裁人数和屏障的概念的文章。

使用 Linux Virtual Server 进行 负载均衡

10

Linux Virtual Server (LVS) 的目标是提供一个基本框架，将网络连接定向到共享其工作负载的多台服务器。Linux Virtual Server 是服务器群集（一个或多个负载均衡器及多台用于运行服务的真实的服务器），对于外部客户端显示为一台大型的快速服务器。这种看上去像是单台服务器的服务器被称为*虚拟服务器*。Linux Virtual Server 可用于构建可灵活缩放并具有高可用性的网络服务，如 Web、缓存、邮件、FTP、媒体和 VoIP 服务。

真实的服务器和负载均衡器可通过高速 LAN 或地理位置分散的 WAN 互相连接。负载均衡器可将请求发送到不同的服务器。它们使群集的并行服务看似单个 IP 地址（虚拟 IP 地址，即 VIP）上的虚拟服务。发送请求时可使用 IP 负载均衡技术或应用程序级的负载均衡技术。系统的可伸缩性通过在群集中透明地添加或去除节点来实现。高可用性通过检测节点或守护程序故障并相应地重配置系统来提供。

10.1 概念概述

以下部分概述了 LVS 主要组件和概念。

10.1.1 Director

LVS 主要组件是 ip_vs（或 IPVS）内核代码。它在 Linux 内核中（第 4 层交换）实施传输层负载均衡。运行包含 IPVS 代码的 Linux 内核的节点称为*定向器*。控制器上运行的 IPVS 代码是 LVS 的基本功能。

当客户端连接到定向器时，进来的请求在所有群集节点间是负载平衡的：定向器使用一组能使 LVS 正常工作的修改过的路由规则，将包转发到真实服务器。例如，连接不会在定向器上发起或终止，它也不会发送确认。定向器相当于将包从最终用户转发到真实服务器（运行用于处理请求的应用程序的主机）的专用路由器。

默认情况下，内核未安装 IPVS 模块。IPVS 内核模块包含在 `cluster-network-kmp-default` 包中。

10.1.2 用户空间控制器和守护程序

`ldirectord` 守护程序是一个用户空间守护程序，用于管理 Linux Virtual Server 和监视负载平衡的虚拟服务器的 LVS 群集中的真实服务器。配置文件 `/etc/ha.d/ldirectord.cf` 指定虚拟服务及其相关真实服务器，并告知 `ldirectord` 如何将此服务器配置为 LVS 重定向器。守护程序初始化时，将为群集创建虚拟服务。

`ldirectord` 守护程序通过定期请求已知的 URL 并检查响应来监视真实服务器的运行状况。如果真实服务器发生故障，它将从负载平衡器的可用服务器列表中删除。如果设备监视器检测到发生故障的服务器已恢复并重新工作，则它会将此服务器重新添加到可用服务器列表中。如果出现所有真实服务器均宕机的情况，可以指定一台将 Web 服务重定向到的备用服务器。备用服务器通常是本地主机，它会显示一个应急页面，说明 Web 服务暂时不可用。

10.1.3 包的转发

定向器可采用三种不同方法将包从客户端发送到真实服务器：

网络地址转换 (NAT)

进来的请求到达虚拟 IP，然后通过将目标 IP 地址和端口更改为所选真实服务器的 IP 地址和端口将进来的请求转发到真实服务器。真实服务器向负载平衡器发送响应，负载平衡器再更改目标 IP 地址并将响应发回客户端，这样最终用户就从预期的源收到回复了。由于所有通讯都要流经负载平衡器，它通常会成为群集的瓶颈。

IP 隧道通讯进程（IP-IP 封装）

IP 隧道通讯进程允许将发送到某个 IP 地址的包重定向到可能处于其他网络上的另一个地址。LVS 通过 IP 隧道（重定向到其他 IP 地址）向真实服务器

发送请求，然后真实服务器使用自己的路由选择表直接回复到客户端。群集成员可以处于不同的子网中。

直接路由选择

来自最终用户的包将直接转发到真实服务器。IP 包未经修改，因此必须配置真实服务器以接受虚拟服务器 IP 地址的通讯。来自真实服务器的响应会直接发送到客户端。真实服务器和负载均衡器必须处于同一物理网络段中。

10.1.4 调度算法

确定将哪台真实服务器用于客户端请求的新连接，是使用不同算法来实现的。这些算法以模块的形式提供，可进行调整以适应特定需要。有关可用模块的概述，请参阅 `ipvsadm(8)` 手册页。从客户端接收到连接请求时，控制器会根据日程表将一台真实的服务器指派给此客户端。调度程序是 IPVS 内核代码的组成部分，它决定哪台真实的服务器将获取下一个新连接。

10.2 使用 YaST 配置 IP 负载均衡

可使用 YaST `iplb` 模块配置基于内核的 IP 负载均衡。它是 `ldirectord` 的前端。

要访问“IP 负载均衡”对话框，请以 `root` 用户身份启动 YaST 并选择 *High Availability > IP 负载均衡*。或者，以 `root` 用户身份使用 `yast2 iplb` 从命令行启动 YaST 群集模块。

YaST 模块会将其配置写入 `/etc/ha.d/ldirectord.cf`。YaST 模块中的可用选项卡对应于 `/etc/ha.d/ldirectord.cf` 配置文件的结构，此配置文件用于定义全局选项和虚拟服务的选项。

有关配置示例以及负载均衡器和真实服务器之间产生的进程，请参阅例 10.1 “简单的 `ldirectord` 配置”（第 130 页）。

注意：全局参数和虚拟服务器参数

如果在虚拟服务器部分和全局部分都指定了某个参数，那么在虚拟服务器部分中定义的值将覆盖在全局部分中定义的值。

过程 10.1 配置全局参数

以下过程描述了如何配置最重要的全局参数。有关个别参数（以及此处未提及的参数）的更多细节，请单击 [帮助](#) 或参阅 `ldirectord` 手册页。

- 1 通过 **检查间隔** 定义 `ldirectord` 连接到每台真实服务器以检查它们是否仍处于联机状态的间隔。
- 2 通过 **检查超时** 设置真实服务器应该在多长时间内响应上次检查。
- 3 通过 **检查计数** 定义在检查被视为失败前 `ldirectord` 可尝试向真实服务器发送多少次请求。
- 4 通过 **协商超时** 定义协商检查经过多少秒后应视为超时。
- 5 在 **备用中**，输入当所有真实服务器都宕机时，要将 Web 服务重定向到的 Web 服务器的主机名或 IP 地址。
- 6 如果要使用备用路径进行日志记录，请在 **日志文件** 中指定日志的路径。默认情况下，`ldirectord` 会将其日志写入 `/var/log/ldirectord.log`。
- 7 如果希望系统在任何真实服务器的连接状态发生改变时均发送警报，请在 **电子邮件警报** 中输入有效的电子邮件地址。
- 8 通过 **电子邮件警报频率** 定义经过多少秒后，如果任何真实服务器仍无法访问，应重复发出电子邮件警报。
- 9 在 **电子邮件警报状态** 中指定应发送电子邮件警报的服务器状态。如果要定义多个状态，请使用逗号分隔的列表。
- 10 通过 **自动重新装载** 定义 `ldirectord` 是否应连续监视配置文件有无修改。如果设置为 `yes`，则将在发生更改时自动重新装载配置。
- 11 通过 **静止开关** 定义是否应从内核的 LVS 表中删除发生故障的真实服务器。如果设置为 `是`，则不会删除发生故障的服务器。而是将其权重设置为 0，表示不接受新连接。已建立的连接将继续存在，直到超时为止。

图 10.1 YaST IP 负载平衡 - 全局参数

The image shows the 'IPLB - Global Configuration' window in YaST. It has two tabs: 'Global Configure' (selected) and 'Virtual Server Configure'. The 'Global Configure' tab contains several input fields and dropdown menus. At the top, there are four input fields: 'Check Interval' (set to 5), 'Check Timeout' (set to 3), 'Check Count' (empty), and 'Negotiate Timeout' (empty). Below these are two more input fields: 'Fallback' and 'Log File'. Further down are three input fields: 'Email Alert', 'Email Alert Freq', and 'Email Alert Status'. Below these are two more input fields: 'Callback' and 'Execute'. At the bottom, there are four dropdown menus: 'Auto Reload' (set to yes), 'Quiescent' (set to yes), 'Fork' (empty), and 'Supervised' (empty). At the very bottom, there are three buttons: 'Help', 'Cancel', and 'OK'.

过程 10.2 配置虚拟服务


可通过为每种虚拟服务定义若干参数来配置一个或多个虚拟服务。以下过程描述了如何配置虚拟服务最重要的参数。有关个别参数（以及此处未提及的参数）的更多细节，请单击[帮助](#)或参阅 `ldirectord` 手册页。

- 1 在 YaST `iplb` 模块中，切换到*虚拟服务器配置*选项卡。
- 2 添加新虚拟服务器或编辑现有虚拟服务器。一个新对话框将显示可用选项。
- 3 在*虚拟服务器*中，输入负载平衡器和真实服务器可作为 LVS 访问的共享虚拟 IP 地址和端口。还可以指定主机名和服务来代替 IP 地址和端口名称。或者，也可以使用防火墙标记。防火墙标记是一种将任意 VIP:port 服务的集合聚合到一个虚拟服务中的方法。
- 4 要指定*真实服务器*，需要输入服务器的 IP 地址（或主机名）、端口（或服务名称）以及转发方法。转发方法必须是 `gate`、`ipip` 或 `masq` 中的一种，请参见第 10.1.3 节“包的转发”（第 124 页）。

单击添加按钮，为每台真实服务器输入需要的自变量。

- 5 在检查类型中选择用于测试真实服务器是否仍然活动的检查类型。例如，要发送请求并检查响应是否包含预期的字符串，请选择 `Negotiate`。
- 6 如果已将检查类型设置为 `Negotiate`，则还需定义要监视的服务类型。从服务下拉列表中进行选择。
- 7 在请求中输入检查间隔期间每台真实服务器上所请求对象的 URI。
- 8 如果要检查来自真实服务器的响应是否包含特定字符串（如 “I’m alive” 消息），请定义需要匹配的正则表达式。将正则表达式输入到接收中。如果来自真实服务器的响应包含此表达式，则认为真实服务器处于活动状态。
- 9 根据您在步骤 6（第 128 页）中选择的 *服务类型*，还需要指定其他参数，如 *登录*、*密码*、*数据库* 或 *机密*。有关更多信息，请参阅 YaST 帮助文本或 `ldirectord` 手册页。
- 10 选择用于负载均衡的 *调度程序*。有关可用调度程序的信息，请参阅 `ipvsadm(8)` 手册页。
- 11 选择要使用的 *协议*。如果将虚拟服务指定为 IP 地址和端口，则它必须是 `tcp` 或 `udp`。如果将虚拟服务指定为防火墙标记，则协议必须是 `fwm`。
- 12 如果需要，可定义其他参数。单击 *确定* 确认配置。YaST 会将此配置写入 `/etc/ha.d/ldirectord.cf`。

图 10.2 YaST IP 负载均衡 - 虚拟服务

 **IPLB - Virtual Servers Configuration**

Virtual Server

Real Servers

192.168.0.110:80 gate
192.168.0.120:80 gate

Check Type Service Check Command Check Port

Request Receive Http Method Virtual Host

Login Password Database Name Radius Secret

Persistent Netmask Scheduler Protocol

Check Timeout Negotiate Timeout Check Count Email Alert

Email Alert Freq Email Alert Status Fallback Quiescent

例 10.1 简单的 *ldirectord* 配置

图 10.1 “YaST IP 负载平衡 - 全局参数”（第 127 页）和图 10.2 “YaST IP 负载平衡 - 虚拟服务”（第 129 页）中所示的值将生成在 `/etc/ha.d/ldirectord.cf` 中定义的以下配置：

```
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 定义 *ldirectord* 应连续检查配置文件有无修改。
- ❷ *ldirectord* 连接到每台真实服务器以检查它们是否仍处于联机状态的间隔。
- ❸ 真实服务器必须在上次检查后多长时间内作出响应。
- ❹ 定义不要将发生故障的真实服务器从内核的 LVS 表中删除，但将其权重设置为 0。
- ❺ LVS 的虚拟 IP 地址 (VIP)。可通过端口 80 访问 LVS。
- ❻ 用于测试真实服务器是否仍处于活动状态的检查类型。
- ❼ 此服务的所有真实服务器都宕机时，要将 Web 服务重定向到的服务器。
- ❽ 要使用的协议。
- ❾ 定义了两台真实服务器，均可通过端口 80 访问。包的转发方法是 `gate`，表示使用直接路由选择。
- ❿ 需要在真实服务器的响应字符串中匹配的正则表达式。
- ⓫ 检查间隔期间每台真实服务器上所请求对象的 URI。
- ⓬ 用于负载平衡的所选调度程序。
- ⓭ 要监视的服务类型。

此配置将导致以下进程流：ldirectord 每 5 秒连接到每台真实服务器一次 ❷ 以及按 ❸ 和 ❹ 中指定的方式请求 192.168.0.110:80/test.html 或 192.168.0.120:80/test.html。如果上次检查后 3 秒内 ❺ 未收到来自真实服务器的预期的 still alive 字符串 ❻，它会将此真实服务器从可用池中删除。但是，由于 quiescent=yes 设置 ❼，真实服务器不会从 LVS 表中删除，但权重将设置为 0，这样就不会接受此真实服务器的新连接。已建立的连接将继续存在，直到超时为止。

10.3 其他设置

除了使用 YaST 配置 ldirectord 外，还需要确保满足以下条件，才能完成 LVS 设置：

- 正确设置真实服务器以提供所需服务。
- 负载均衡服务器（或服务器）必须能够使用 IP 转发将通讯路由到真实服务器。真实服务器的网络配置取决于选择的包转发方法。
- 为避免负载均衡服务器（或服务器）成为整个系统的单个故障点，需要设置负载均衡器的一个或多个备份。在群集配置中配置 ldirectord 的原始资源，以便在发生硬件故障时，ldirectord 可以故障转移到其他服务器。
- 由于负载均衡器的备份也需要 ldirectord 配置文件才能完成其任务，因此请确保 /etc/ha.d/ldirectord.cf 在要用作负载均衡器备份的所有服务器上都可用。可以按第 3.2.3 节“将配置传送到所有节点”（第 24 页）中所述使用 Csync2 同步配置文件。

10.4 更多信息

要了解更多有关 Linux Virtual Server 的信息，请参阅 <http://www.linuxvirtualserver.org/> 上的项目主页。

有关 ldirectord 的更多信息，请参阅其综合性手册页。

网络设备联接

对于许多系统，需要实施符合典型以太网设备的标准数据安全性或可用性要求的网络连接。在这种情况下，可以将多种以太网设备聚合到单个联接设备。

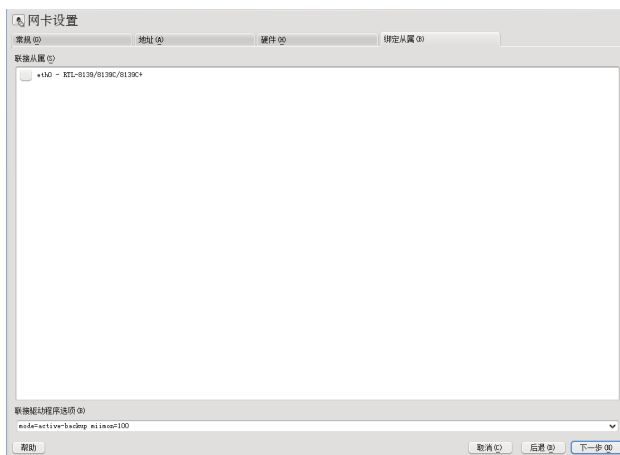
联接设备的配置通过联接模块选项来完成。其行为取决于联接设备的模式。默认情况下是 `mode=active-backup`，即如果活动从属设备发生故障，则其他从属设备将变成为活动设备。

使用 OpenAIS 时，不通过群集软件来管理联接设备。因此，必须在每个可能需要访问联接设备的群集节点上配置联接设备。

11.1 使用 YaST 配置绑定设备

要配置联接设备，请使用以下过程：

- 1 以 `root` 用户身份启动 YaST 并选择 *网络设备 > 网络设置*。
- 2 单击添加配置新网卡，并将设备类型更改为绑定。按下一步继续。



3 选择如何为联接设备指派 IP 地址。有三种方法可供选择：

- 无 IP 地址
- 动态地址（使用 DHCP 或 Zeroconf）
- 静态指派的 IP 地址

使用最适合您环境的方法。如果 OpenAIS 管理虚拟 IP 地址，请选择*静态指派 IP 地址*并将基本 IP 地址指派给接口。

4 切换到*绑定从属*选项卡。

5 要选择需要加入绑定的以太网设备，请激活相关*绑定从属*前面的复选框。

6 编辑*联接驱动程序选项*。可以使用以下模式：

balance-rr
提供负载均衡和容错。

active-backup
提供容错。

balance-xor
提供负载均衡和容错。

广播
提供容错

802.3ad
提供动态链接集合（如果连接的交换机支持动态链接集合）。

balance-tlb
为外发的通讯量提供负载平衡。

balance-alb
为进来的和外发的通讯量提供负载平衡（如果使用的网络设备允许在使用中修改网络设备的硬件地址）。

7 确保将参数 `miimon=100` 添加到**联接驱动程序选项**。如果没有此参数，则不会定期检查数据完整性。

8 单击**下一步**，然后单击**确定退出 YaST** 以创建设备。

11.2 更多信息

*Linux 以太网联接驱动程序操作指南*中详细解释了所有模式及许多其他选项，在安装 `kernel-source` 包后可以在 `/usr/src/linux/Documentation/networking/bonding.txt` 上找到这些内容。

部分 III. 储存和数据复制

Oracle Cluster File System 2

Oracle Cluster File System 2 (OCFS2) 是一个通用日记文件系统，自 Linux 2.6 内核以来就已完全集成。OCFS2 可将应用程序二进制文件、数据文件和数据库储存到共享储存设备。群集中的所有节点对文件系统都有并行的读和写权限。用户空间控制守护程序，通过克隆资源进行管理，提供与 HA 堆栈特别是 OpenAIS/Corosync 和分布式锁管理器 (DLM) 的集成。

12.1 功能和优点

OCFS2 可用于以下储存解决方案，例如：

- 一般应用程序和工作负荷。
- 群集中的 Xen 映像储存。Xen 虚拟机和虚拟服务器可储存在群集服务器安装的 OCFS2 卷上。这为 Xen 虚拟机提供了服务器之间快速方便的可移植性。
- LAMP（Linux、Apache、MySQL 和 PHP | Perl | Python）堆栈。

作为高性能、非对称的并行群集文件系统，OCFS2 支持以下功能：

- 应用程序文件对群集中的所有节点均可用。用户只需在群集中的 OCFS2 上安装它一次。
- 所有节点都可以通过标准文件系统接口直接并行读写至储存区，从而方便地管理运行于群集上的应用程序。

- 文件访问通过 DLM 协调。DLM 控制在多数情况下都运行良好，但如果应用程序的设计与 DLM 争夺对文件访问的协调，则此设计可能会限制可伸缩性。
- 所有后端储存上都可以使用储存备份功能。可以方便地创建共享应用程序文件的图形，它能够帮助提供有效的故障恢复。

OCFS2 还提供以下功能：

- 元数据缓存。
- 元数据日记。
- 跨节点的文件数据一致性。
- 支持最多 4 KB 的多个块大小（每个卷可有不同的块大小），最大卷大小为 16 TB。
- 支持最多 16 个群集节点。
- 对于数据库文件的异步和直接 I/O 支持，提高了数据库性能。

12.2 OCFS2 包和管理实用程序

OCFS2 内核模块 (`ocfs2`) 自动安装在 SUSE® Linux Enterprise Server 11 SP1 上的 High Availability Extension 中。要使用 OCFS2，请确保以下包已安装在群集中的每个节点上：`ocfs2-tools` 和适用于内核的匹配 `ocfs2-kmp-*` 包。

`ocfs2-tools` 包提供以下实用程序，用于管理 OFS2 卷。有关语法信息，请参见其手册页。

表 12.1 OCFS2 实用程序

OCFS2 实用程序	描述
<code>debugfs.ocfs2</code>	为了调试检查 OCFS 文件系统的状态
<code>fsck.ocfs2</code>	检查文件系统的错误并进行选择性的修改。

OCFS2 实用程序	描述
mkfs.ocfs2	在某个设备上创建 OCFS2 文件系统，通常是共享物理或逻辑磁盘上的某个分区。
mounted.ocfs2	检测并列出群集系统上所有的 OCFS2 卷。检测并列出已经安装了 OCFS2 设备的系统上的所有节点或列出所有的 OCFS2 设备。
tunefs.ocfs2	更改 OCFS2 文件系统参数，包括卷标、节点槽号、所有节点槽的日志大小和卷大小。

12.3 配置 OCFS2 服务

必须将以下资源配置为群集中的服务：DLM 和 O2CB，才能创建 OCFS2 卷。OCFS2 使用运行于用户空间中的 Pacemaker 提供的群集成员资格服务。因此，DLM 和 O2CB 需要配置为存在于群集中每个节点上的克隆资源。

过程 12.1 配置 DLM 和 O2CB 资源

以下过程使用 `crm` 外壳配置群集资源。对群集中的一个节点执行以下步骤。或者，也可以使用 `Heartbeat` 配置资源。

1 打开终端窗口，并以 `root` 用户身份或等价用户身份登录。

2 将 DLM（分布式锁管理器）添加为资源：

2a 启动 `crm` 外壳并从零开始创建新配置：

```
crm
cib new stack-glue
```

2b 创建 DLM 服务并使其运行于群集中的所有计算机上：

```
configure
primitive dlm ocf:pacemaker:controld op monitor interval=120s
clone dlm-clone dlm meta globally-unique=false interleave=true
end
```

dlm 克隆资源会控制分布式锁管理器服务，并确保此服务在群集中的所有节点上都启动。

2c 校验所做更改，然后将其提交到 CIB：

```
cib diff
configure verify
```

2d 将配置上载到群集，并退出外壳：

```
cib commit stack-glue
quit
```

3 添加 O2CB 配置：

3a 启动 crm 外壳并从零开始创建新配置：

```
crm
cib new oracle-glue
```

3b 使 O2CB 服务在群集中的每个节点上都启动：

```
configure
primitive o2cb ocf:ocfs2:o2cb op monitor interval=120s
clone o2cb-clone o2cb meta globally-unique=false interleave=true
```

3c 为确保 O2CB 服务仅在也具有已运行的 dlm 服务副本的节点上启动，请添加并列限制：

```
colocation o2cb-with-dlm INFINITY: o2cb-clone dlm-clone
order start-o2cb-after-dlm mandatory: dlm-clone o2cb-clone
```

3d 将配置上载到群集，并退出外壳：

```
cib commit oracle-glue
quit
```

4 配置屏蔽设备：

4a 启动 crm 外壳并从零开始创建新配置：

```
crm
cib new fencing
```

- 4b** 在将 `/dev/sdb2` 作为共享储存区上用于检测信号和屏蔽的专用分区的同时，将 `external/sdb` 配置为屏蔽设备：

```
configure
primitive sbd_stonith stonith:external/sbd \
meta target-role="Started"op monitor \
interval=15 timeout=15 start-delay=15 \
params sbd_device=/dev/sdb2
```

- 4c** 将配置上载到群集，并退出外壳：

```
cib commit fencing
quit
```

12.4 创建 OCFS2 卷

按第 12.3 节“配置 OCFS2 服务”（第 141 页）中所述将 DLM 和 O2CB 配置为群集资源后，配置系统以使用 OCFS2 并创建 OCFS2 卷。

注意：适用于应用程序和数据文件的 **OCFS2 卷**

我们建议您通常将应用程序文件和数据文件储存在不同的 OCFS2 卷上。如果应用程序卷和数据卷具有不同的装入要求，则必须将它们储存在不同的卷上。

开始之前要准备计划用于 OCFS2 卷的块设备。将这些设备留作可用空间。

然后，按过程 12.2,“创建并格式化 OCFS2 卷”（第 145 页）中所述使用 `mkfs.ocfs2` 创建和格式化 OCFS2 卷。此命令最重要的参数列于表 12.2“重要的 OCFS2 参数”（第 143 页）中。有关此命令的更多信息和命令语法，请参阅 `mkfs.ocfs2` 手册页。

表 12.2 重要的 OCFS2 参数

OCFS2 参数	描述和建议
卷标 (-L)	卷的描述性名称能够在不同节点上安装卷时唯一标识它。使用 <code>tuneefs.ocfs2</code> 实用程序根据需要修改该卷标。

OCFS2 参数	描述和建议
群集大小 (-C)	群集大小是分配给文件以保存数据的最小空间单元。有关可用选项和推荐的信息，请参阅 <code>mkfs.ocfs2</code> 手册页。
节点槽数 (-N)	<p>可以同时安装卷的最大节点数。OCFS2 会为每个节点创建单独的系统文件（如日记）。访问卷的节点可以是小尾端结构（如 <code>x86 x86-64</code> 和 <code>ia64</code>）和大尾端结构（如 <code>ppc64</code> 和 <code>s390x</code>）的组合。</p> <p>特定于节点的文件作为本地文件。节点槽号附加到该本地文件。例如：<code>journal:0000</code> 属于指派到槽号 0 的所有节点。</p> <p>根据预期有多少个节点并行装入卷，在创建卷时设置每个卷的最大节点槽数。使用 <code>tunefs.ocfs2</code> 实用程序根据需要增加节点槽数。请注意，此值不能减小。</p>
块大小 (-b)	文件系统可寻址的最小空间单元创建卷时请指定块大小。有关可用选项和推荐的信息，请参阅 <code>mkfs.ocfs2</code> 手册页。
打开/关闭特定功能 (--fs-features)	<p>可以提供功能标志的逗号分隔列表，<code>mkfs.ocfs2</code> 会尝试根据此列表创建具有那些功能的文件系统。要打开某功能，请将其加入列表。要关闭某功能，请在其名称前加 <code>no</code>。</p> <p>有关所有可用标志的概述，请参阅 <code>mkfs.ocfs2</code> 手册页。</p>
预定义功能 (--fs-feature-level)	<p>允许您从一组预定义文件系统功能中进行选择。有关可用选项的信息，请参阅 <code>mkfs.ocfs2</code> 手册页。</p>

如果使用 `mkfs.ocfs2` 创建和格式化卷时未指定任何特定功能，则默认情况下将启用以下功能：`backup-super`、`sparse`、`inline-data`、`unwritten`、`metaecc`、`indexed-dirs` 和 `xattr`。

过程 12.2 创建并格式化 OCFS2 卷

只在群集节点之一上执行以下步骤。

- 1 打开终端窗口，并以 `root` 用户身份登录。
- 2 检查群集是否与 `crm_mon` 命令联机。
- 3 使用 `mkfs.ocfs2` 实用程序创建并格式化卷。有关此命令语法的信息，请参见 `mkfs.ocfs2` 手册页。

例如，要在最多支持 16 个群集节点的 `/dev/sdb1` 上创建新的 OCFS2 文件系统，请使用以下命令：

```
mkfs.ocfs2 -N 16 /dev/sdb1
```

12.5 装入 OCFS2 卷

可以手动装入 OCFS2 卷，也可以按过程 12.4, “使用群集管理器装入 OCFS2 卷”（第 146 页）中所述使用群集管理器将其装入。

过程 12.3 手动装入 OCFS2 卷

- 1 打开终端窗口，并以 `root` 用户身份登录。
- 2 检查群集是否与 `crm_mon` 命令联机。
- 3 使用 `mount` 命令从命令行装入卷。

警告：手动装入的 OCFS2 设备

如果手动装入 OCFS2 文件系统用于测试目的，请务必在开始通过 OpenAIS 使用它之前再次卸载它。

过程 12.4 使用群集管理器装入 OCFS2 卷

要使用 High Availability 软件装入 OCFS2 卷，请在群集中配置 ocfFile System 资源。以下过程使用 crm 外壳配置群集资源。或者，也可以使用 Heartbeat 配置资源。

- 1 启动 crm 外壳并从零开始创建新配置：

```
crm
cib new filesystem
```

- 2 配置 Pacemaker 以在群集中的每个节点上装入 OCFS2 文件系统：

```
configure
primitive fs ocf:heartbeat:Filesystem \
    params device="/dev/sdb1" directory="/mnt/shared" fstype="ocfs2" \
    op monitor interval=120s
clone fs-clone fs meta interleave="true" ordered="true"
```

- 3 为确保 Pacemaker 仅在也具有已运行的 o2cb 资源的克隆的节点上启动 fs 克隆资源，请添加并列限制：

```
colocation fs-with-o2cb INFINITY: fs-clone o2cb-clone
order start-fs-after-o2cb mandatory: o2cb-clone fs-clone
```

- 4 将配置上载到 CIB，并退出外壳：

```
cib commit filesystem
quit
```

12.6 更多信息

有关 OCFS2 的更多信息，请参见以下链接：

<http://oss.oracle.com/projects/ocfs2/>
Oracle 上的 OCFS2 项目主页。

<http://oss.oracle.com/projects/ocfs2/documentation>
项目文档主页上的《OCFS2 用户指南》

分布式复制块设备 (DRBD)

通过 DRBD，您可以为位于 IP 网络上两个不同站点的两个块设备创建镜像。当用于 OpenAIS 时，DRBD 支持分布式高可用性 Linux 群集。本章说明如何安装和设置 DRBD。

13.1 概念概述

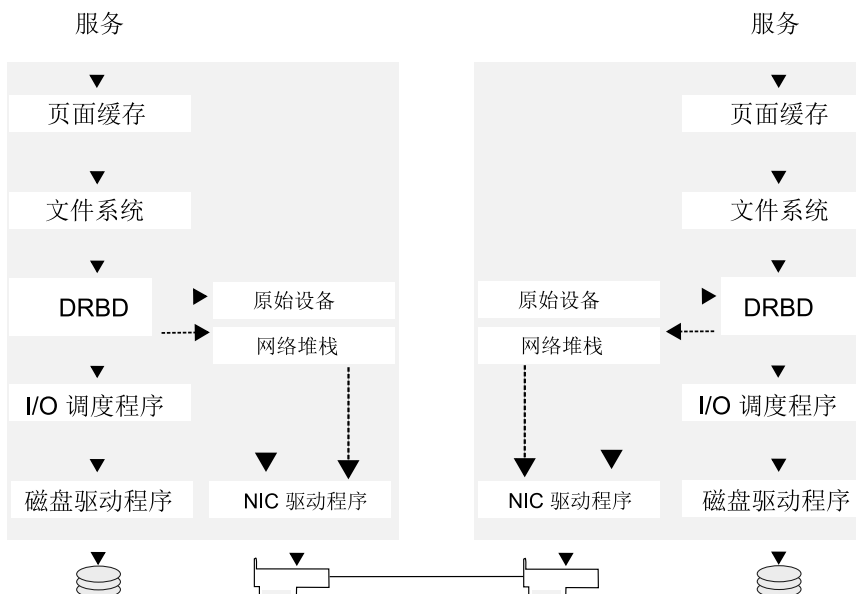
DRBD 以确保数据的两个副本保持相同的方式将主设备上的数据复制到次设备上。将其视为联网的 RAID 1。它实时对数据进行镜像，以便连续复制。应用程序不需要知道实际上它们的数据储存在不同的磁盘上。

重要：未加密数据

镜像之间的数据通讯是不加密的。为实现安全数据交换，您应为连接部署虚拟专用网 (VPN) 解决方案。

DRBD 是 Linux 内核模块，位于下端的 I/O 调度程序与上端的文件系统之间，请参见图 13.1 “DRBD 在 Linux 中的位置”（第 148 页）。要与 DRBD 通讯，用户需使用高级别命令 `drbdadm`。为了提供最大的灵活性，DRBD 附带了低级别工具 `drbdsetup`。

图 13.1 DRBD 在 Linux 中的位置



DRBD 允许使用 Linux 支持的任何块设备，通常包括：

- 硬盘分区或完整硬盘
- 软件 RAID
- 逻辑卷管理器 (LVM)
- 企业卷管理系统 (EVMS)

默认情况下，DRBD使用TCP端口7780及更高端口进行DRBD节点间的通讯。请确定您的防火墙不会阻止此端口上的通讯。

在 DRBD 设备上创建文件系统之前，必须先设置 DRBD 设备。与用户数据相关的所有操作都应通过 `/dev/drbd_R` 设备单独执行而不是在原始设备上执行，因为 DRBD 会将原始设备的最后 128 MB 用于元数据。确保仅在 `/dev/drbd<n>` 设备而不是在原始设备上创建文件系统。

例如，如果原始设备大小为 1024 MB，则 DRBD 设备仅有 896 MB 可用于数据，128 MB 隐藏并保留用于元数据。任何访问 896 MB 和 1024 MB 之间的空间的尝试都会失败，因为它不可用于用户数据。

13.2 安装 DRBD 服务

要安装 DRBD 所需的包，请按第 I 部分“安装和设置”（第 1 页）中所述在联网群集中的两台 SUSE Linux Enterprise Server 计算机上安装 High Availability Extension 外接式附件产品。安装 High Availability Extension 也将安装 DRBD 程序文件。

如果不需要完整的群集堆栈，只想使用 DRBD，表 13.1 “DRBD RPM 包”（第 149 页）包含 DRBD 的所有 RPM 包的列表。在上一版本中，drbd 包分为几个单独的包。

表 13.1 *DRBD RPM 包*

文件名	说明
drbd	简易包，分为其他包
drbd-bash-completion	针对 drbdadm 的可编程 bash completion 支持
drbd-heartbeat	DRBD 的 Heartbeat 资源代理（仅 Heartbeat 需要）
drbd-kmp-default	DRBD 的内核模块（需要）
drbd-kmp-xen	DRBD 的 Xen 内核模块
drbd-udev	DRBD 的 udev 集成脚本，管理指向 /dev/drbd/by-res 和 /dev/drbd/by-disk 中的 DRBD 设备的 symlink
drbd-utils	DRBD 的管理实用程序（需要）
drbd-pacemaker	DRBD 的 Pacemaker 资源代理

文件名	说明
<code>drbd-xen</code>	DRBD 的 Xen 块设备管理脚本
<code>yast2-drbd</code>	YaST DRBD 配置（建议）

要简化 `drbdadm` 的处理，可使用 RPM 包 `drbd-bash-completion` 中的 **Bash completion** 支持。如果要在当前外壳会话中启用它，请插入以下命令：

```
source /etc/bash_completion.d/drbdadm.sh
```

要对 `root` 永久使用它，请创建文件 `/root/.bashrc` 并插入上一行。

13.3 配置 DRBD 服务

注意

以下过程使用服务器名称 `jupiter` 和 `venus` 以及群集资源名称 `r0`。它将 `jupiter` 设置为主节点。确保修改指令以使用您自己的节点和文件名。

开始配置 DRBD 之前，确保 Linux 节点中的块设备已就绪且已分区（如果需要）。以下过程假定您有两个节点 `jupiter` 和 `venus`，它们使用 TCP 端口 7780。确保此端口在防火墙中处于打开状态。

要手动设置 DRBD，请按如下操作：

过程 13.1 手动配置 DRBD

1 以 `root` 用户身份登录。

2 更改 DRBD 的配置文件：

2a 打开文件 `/etc/drbd.conf` 并插入以下行（如不可用）：

```
include "drbd.d/global_common.conf";
include "drbd.d/*.res";
```

从 DRBD 8.3 开始，配置文件分为单独的文件，位于目录 `/etc/drbd.d` 下。

- 2b** 打开文件 `/etc/drbd.d/global_common.conf`。它已包含一些预定义值。转到 `startup` 部分并插入以下行：

```
startup {  
    # wfc-timeout degr-wfc-timeout outdated-wfc-timeout  
    # wait-after-sb;  
    wfc-timeout 1;  
    degr-wfc-timeout 1;  
}
```

这些选项用于在引导时减少超时，有关更多细节，请参见 <http://www.drbd.org/users-guide-emb/re-drbdconf.html>。

- 2c** 创建文件 `/etc/drbd.d/r0.res`，根据具体情况更改行，然后保存：

```
resource r0 { ❶  
    device /dev/drbd_r0 minor 0; ❷  
    disk /dev/sda1; ❸  
    meta-disk internal; ❹  
    on jupiter { ❺  
        address 192.168.1.10:7780; ❻  
    }  
    on venus { ❺ (page 151)  
        address 192.168.1.11:7780; ❻ (page 151)  
    }  
    syncer {  
        rate 7M; ❼  
    }  
}
```

- ❶ 资源名称。建议使用 `r0`、`r1` 之类的资源名称。
- ❷ DRBD 的设备名及其次要编号。建议以 `/dev/drbd` 开头，然后追加资源名称（本例中为 `r0`）。
- ❸ 在节点间复制的设备。请注意，在本例中，两个节点上的设备相同。如果需要不同设备，请将 `disk` 参数移到 `on` 部分中。
- ❹ 元磁盘参数通常包含值 `internal`，但可以指定显式设备以保存元数据。有关更多信息，请参见 <http://www.drbd.org/users-guide-emb/ch-internals.html#s-metadata>。
- ❺ `on` 部分包含节点的主机名
- ❻ 各个节点的 IP 地址和端口号。每个资源都需要单独的端口，通常从 7780 开始。

- ⑦ 同步速率。将其设置为带宽的三分之一。它仅限制重新同步，而不限制镜像。

3 检查配置文件的语法。如果以下命令返回错误，请校验文件：

```
drbdadm dump all
```

4 将 DRBD 配置文件复制到其他节点：

```
scp /etc/drbd.conf venus:/etc/  
scp /etc/drbd.d/* venus:/etc/drbd.d/
```

5 通过在每个节点上输入以下命令，初始化两个系统上的元数据。

```
drbdadm -- --ignore-sanity-checks create-md r0  
rckdrbd start
```

如果磁盘已包含不再需要的文件系统，请使用以下命令清空文件系统结构，然后重复此步骤：

```
dd if=/dev/zero of=/dev/sdb1 count=10000
```

6 通过在每个节点上输入以下命令，监视 DRBD 状态：

```
rckdrbd status
```

应得到类似以下内容的输出：

```
drbd driver loaded OK; device status:  
version: 8.3.7 (api:88/proto:86-91)  
GIT-hash: ea9e28dbff98e331a62bcbcc63a6135808fe2917 build by phil@fat-tyre, 2010-01-13  
17:17:27  
m:res cs ro ds p mounted fstype  
0:r0 Connected Secondary/Secondary Inconsistent/Inconsistent C
```

7 在所需的主节点（本例中为 jupiter）上启动重新同步进程：

```
drbdadm -- --overwrite-data-of-peer primary r0
```

8 再次使用 rckdrbd status 检查状态，将得到：

```
...  
m:res cs ro ds p mounted fstype  
0:r0 Connected Primary/Secondary UpToDate/UpToDate C
```

ds 行中的状态（磁盘状态）在两个节点上都必须为 UpToDate。

9 将 jupiter 设置为主节点：

```
drbdadm primary r0
```

10 在 DRBD 设备上创建文件系统，例如：

```
mkfs.ext3 /dev/drbd_r0
```

11 装入文件系统并使用它：

```
mount /dev/drbd_r0 /mnt/
```

13.4 测试 DRBD 服务

如果安装和配置过程和预期一样，则您就准备好运行 DRBD 功能的基本测试了。此测试还有助于了解该软件的工作原理。

1 测试 jupiter 上的 DRBD 服务。

1a 打开终端控制台，然后以 root 用户身份登录。

1b 在 jupiter 上创建安装点，如 /srv/r0mount：

```
mkdir -p /srv/r0mount
```

1c 装入 drbd 设备：

```
mount -o rw /dev/drbd0 /srv/r0mount
```

1d 从主节点创建文件：

```
touch /srv/r0mount/from_node1
```

2 测试 venus 上的 DRBD 服务。

2a 打开终端控制台，然后以 root 用户身份登录。

2b 卸载 jupiter 上的磁盘：

```
umount /srv/r0mount
```

2c 在 **jupiter** 上输入以下命令，降级 **jupiter** 上的 DRBD 服务：

```
drbdadm secondary r0
```

2d 在 **venus** 上，将 DRBD 服务升级为主服务：

```
drbdadm primary r0
```

2e 在 **venus** 上，检查 **venus** 是否为主节点：

```
rcdrbd status
```

2f 在 **venus** 上，创建安装点，如 **/srv/r0mount**：

```
mkdir /srv/r0mount
```

2g 在 **venus** 上，装入 DRBD 设备：

```
mount -o rw /dev/drbd0 /srv/r0mount
```

2h 校验在 **jupiter** 上创建的文件是否可查看。

```
ls /srv/r0mount
```

/srv/r0mount/from_node1 文件应列出。

3 如果该服务在两个节点上都运行正常，则 DRBD 安装即已完成。

4 再次将 **jupiter** 设置为主节点。

4a 在 **venus** 上输入以下命令，卸下 **venus** 上的磁盘：

```
umount /srv/r0mount
```

4b 在 **venus** 上输入以下命令，降级 **venus** 上的 DRBD 服务：

```
drbdadm secondary r0
```

4c 在 **jupiter** 上，将 DRBD 服务升级为主服务：

```
drbdadm primary r0
```

4d 在 `jupiter` 上，检查 `jupiter` 是否为主节点：

```
rcdrbd status
```

- 5** 要使服务在服务器有问题时自动启动并故障转移，可以使用 `OpenAIS` 将 `DRBD` 设置为高可用性服务。有关为 `SUSE Linux Enterprise 11` 安装和配置 `OpenAIS` 的信息，请参见第 II 部分“配置和管理”（第 31 页）。

13.5 调整 DRBD

可使用几种方式调整 `DRBD`：

1. 对元数据使用外部磁盘。这将加快连接速度。
2. 创建 `udev` 规则以更改 `DRBD` 设备的预读。将以下行保存在文件 `/etc/udev/rules.d/82-dm-ra.rules` 中，并将 `read_ahead_kb` 值更改为工作负载：

```
ACTION=="add", KERNEL=="dm-*", ATTR{bdi/read_ahead_kb}="4100"
```

此行仅在您使用 `LVM` 时有用。

3. 在 `Linux Software RAID` 系统上激活 `bmbv`。在 `DRBD` 配置（通常位于 `/etc/drbd.d/global_common.conf` 中）的通用磁盘部分中使用以下行：

```
disk {  
    use-bmbv;  
}
```

13.6 DRBD 查错

`drbd` 设置涉及很多不同的组件，这些不同的源可能产生问题。以下部分包括多个常用方案和多种建议解决方案。

13.6.1 配置

如果初始 `drbd` 安装未能和预期一样，则配置中可能出错了。

获取关于配置的信息：

- 1 打开终端控制台，然后以 root 用户身份登录。
- 2 通过运行 drbdadm（带 -d 选项），测试配置文件。输入以下命令：

```
drbdadm -d adjust r0
```

在 adjust 选项的干运行中，drbdadm 将 **DRBD** 资源的实际配置与您的 **DRBD** 配置文件进行比较，但它不会执行这些调用。检查输出以确保您了解任何错误的根源。

- 3 如果 /etc/drbd.d/* 和 drbd.conf 文件中存在错误，请先更正，然后再继续。
- 4 如果分区和设置正确，请在不使用 -d 的情况下再次运行 drbdadm。

```
drbdadm adjust r0
```

这会将配置文件应用到 **DRBD** 资源。

13.6.2 主机名

对于 **DRBD**，主机名区分大小写（Node0 是与 node0 不同的主机）。

如果有多个网络设备，且想要使用专用网络设备，可能不会将主机名解析为所用的 IP 地址。在这种情况下，可使用参数 disable-ip-verification。

13.6.3 TCP 端口 7788

如果系统无法连接到对等体，说明本地防火墙可能有问题。默认情况下，**DRBD** 使用 TCP 端口 7788 访问另一个节点。确保在两个节点上该端口均可访问。

13.6.4 DRBD 设备在重引导后损坏

如果 DRBD 不知道哪个实际设备保管了最新数据，它就会变为节点分裂状态。在这种情况下，DRBD 子系统将分别成为次系统，并且互不相连。在这种情况下，会将以下消息写入 `/var/log/messages`：

```
Split-Brain detected, dropping connection!
```

要解决此问题，请在要丢弃其数据的节点上输入以下命令：

```
drbdadm secondary r0  
drbdadm -- --discard-my-data connect r0
```

在具有最新数据的节点上输入以下命令：

```
drbdadm connect r0
```

13.7 更多信息

以下开放源代码资源可用于 DRBD：

- 项目主页：<http://www.drbd.org>。
- Linux Pacemaker 群集堆栈项目的http://clusterlabs.org/wiki/DRBD_HowTo_1.0。
- DRBD 的以下手册页可用于分发：`drbd(8)`、`drbddisk(8)`、`drbdsetup(8)`、`drbdsetup(8)`、`drbdadm(8)` 和 `drbd.conf(5)`。
- 可在 `/usr/share/doc/packages/drbd/drbd.conf` 中查找注释过的 DRBD 示例配置

群集 LVM

当管理群集上的共享储存时，所有节点必须收到有关对储存子系统所做更改的通知。Linux 卷管理器 2 (LVM2) 广泛用于管理本地储存，已扩展为支持对整个群集中的卷组进行透明管理。可使用与本地储存相同的命令来管理群集卷组。

14.1 概念概述

群集 LVM 集成了不同工具：

分布式锁管理器 (DLM)

集成了对 cLVM 的磁盘访问。

逻辑卷管理器 2 (LVM2)

支持将一个文件系统灵活分布到多个磁盘上。LVM 可提供磁盘空间虚拟池。

群集式逻辑卷管理器 (cLVM)

集成了对 LVM2 元数据的访问，以使每个节点都了解相关更改。cLVM 未集成对共享数据本身的访问；要使 cLVM 可执行此操作，必须在 cLVM 管理的储存区上配置 OCFS2 或其他群集感知应用程序。

14.2 cLVM 配置

根据具体情况，可使用具有以下层的 cLVM 创建 RAID 1 设备：

- **LVM** 这是非常灵活的解决方案，可用于增大或减小文件系统大小、添加更多物理储存空间或创建文件系统快照。有关此方法的描述，请参见第 14.2.1 节“方案：iSCSI 在 SAN 上的 cLVM”（第 160 页）。
- **DRBD** 此解决方案仅提供 RAID 0（分段）和 RAID 1（镜像）。有关上一方法的描述，请参见第 14.2.2 节“方案：使用 DRBD 的 cLVM”（第 165 页）。
- **MD 设备（Linux Software RAID 或 mdadm）** 虽然此解决方案可提供所有 RAID 级别，但它尚不支持群集。

请确保已满足以下先决条件：

- 共享储存设备是可用的，如通过光纤通道、FCoE、SCSI、iSCSI SAN 或 DRBD 提供的共享储存设备。
- 如果是 DRBD，那么两个节点都必须是主节点（如以下过程中所述）。
- 检查 LVM2 的锁定类型是否为群集感知。`/etc/lvm/lvm.conf` 中的关键字 `locking_type` 必须包含值 3（应是默认值）。如果需要，将此配置复制到所有节点。

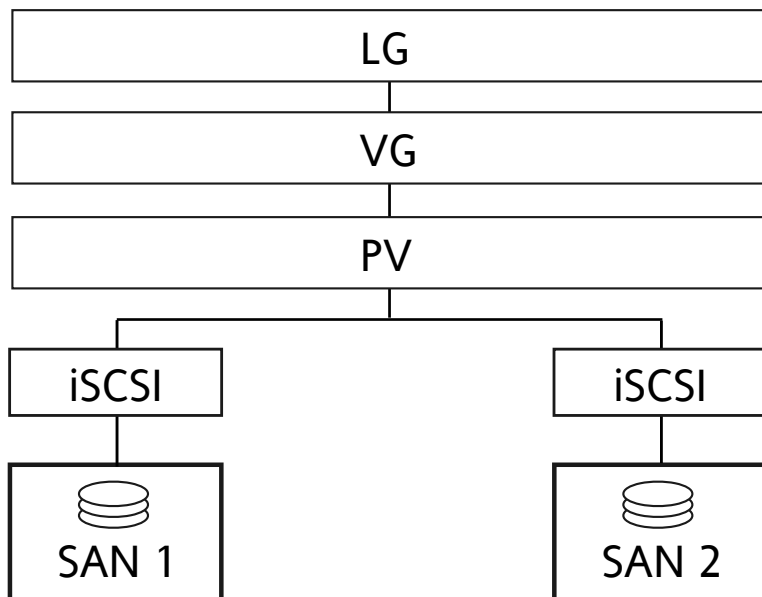
注意：先创建群集资源

先创建群集资源，然后再创建 LVM 卷。否则，稍后将无法删除卷。

14.2.1 方案：iSCSI 在 SAN 上的 cLVM

以下方案使用两个 SAN 盒，将其 iSCSI 目标导出到多个客户端。大致想法如图 14.1 “设置使用 cLVM 的 iSCSI”（第 161 页）所示。

图 14.1 设置使用 *cLVM* 的 iSCSI



警告：数据丢失

以下过程将损坏磁盘上的所有数据！

首先只配置一个 SAN 盒。每个 SAN 盒都必须导出自己的 iSCSI 目标。按如下所示继续：

过程 14.1 配置 iSCSI 目标 (SAN)

- 1 运行 YaST，然后单击 *网络服务 > iSCSI 目标* 以启动 iSCSI 服务器模块。
- 2 如果要在计算机引导时启动 iSCSI 目标，请选择 *引导时*，否则请选择 *手动*。
- 3 如果正在运行防火墙，请启用 *打开防火墙中的端口*。
- 4 切换到全局选项卡。如果需要身份验证，请启用进来的和/或出去的身份验证。在本例中，我们选择 *无身份验证*。

5 添加新的 iSCSI 目标：

5a 切换到 *目标* 选项卡。

5b 单击 *添加*。

5c 输入目标名称。名称格式必须如下所示：

```
iqn.DATE.DOMAIN
```

5d 如果需要描述性更强的名称，可以进行更改，但要确保不同目标之间的标识符是唯一的。

5e 单击 *添加*。

5f 在 *路径* 中输入设备名，并使用 *Scsiid*。

5g 单击 *下一步* 两次。

6 出现警告框时单击是进行确认。

7 打开配置文件 `/etc/iscsi/iscsi.conf`，并将参数 `node.startup` 更改为 `automatic`。

现在按如下方式设置 iSCSI 发起程序：

过程 14.2 配置 iSCSI 发起程序

1 运行 YaST，然后单击 *网络服务 > iSCSI 发起程序*。

2 如果要在计算机引导时启动 iSCSI 发起程序，请选择 *引导时*，否则请将其设置为 *手动*。

3 切换到 *发现* 选项卡并单击 *发现* 按钮。

4 添加 iSCSI 目标的 IP 地址和端口（请参见过程 14.1，“配置 iSCSI 目标 (SAN)”（第 161 页））。通常，可以保留端口并使用其默认值。

5 如果使用身份验证，请插入进来的和出去的用户名和密码，否则请激活 *无身份验证*。

6 选择下一步。找到的连接随即显示在列表中。

7 按完成继续。

8 打开外壳，并以 root 用户身份登录。

9 测试 iSCSI 发起程序是否已成功启动：

```
iscsiadm -m discovery -t st -p 192.168.3.100
192.168.3.100:3260,1 iqn.2010-03.de.jupiter:san1
```

10 建立会话：

```
iscsiadm -m node -l
Logging in to [iface: default, target: iqn.2010-03.de.jupiter:san2,
portal: 192.168.3.100,3260]
Logging in to [iface: default, target: iqn.2010-03.de.venus:san1,
portal: 192.168.3.101,3260]
Login to [iface: default, target: iqn.2010-03.de.jupiter:san2, portal:
192.168.3.100,3260]: successful
Login to [iface: default, target: iqn.2010-03.de.venus:san1, portal:
192.168.3.101,3260]: successful
```

使用 `ls SCSI` 查看设备名：

```
...
[4:0:0:2]    disk      IET        ...      0      /dev/sdd
[5:0:0:1]    disk      IET        ...      0      /dev/sde
```

查找第三列中有 IET 的项。在本例中，设备为 `/dev/sdd` 和 `/dev/sde`。

过程 14.3 创建 DLM 资源

1 启动外壳，并以 root 用户身份登录。

2 运行 `crm configure`。

3 输入以下命令：

```
primitive dlm ocf:pacemaker:controld
primitive clvm ocf:lvm2:clvmd \
    params daemon_timeout="30"
group dlm-clvm dlm clvm
clone dlm-clvm-clone dlm-clvm \
    meta interleave="true" ordered="true"
```

4 使用 `show` 复查更改。

5 如果所有信息都正确，请输入 `commit`，然后使用 `exit` 退出 `crm`。

过程 14.4 创建 LVM 卷组

1 打开已按过程 14.2, “配置 iSCSI 发起程序”（第 162 页）运行 iSCSI 发起程序的一个节点上的 `root` 外壳。

2 使用命令 `pvccreate` 在磁盘 `/dev/sdd` 和 `/dev/sde` 上准备 LVM 的物理卷：

```
pvccreate /dev/sdd
pvccreate /dev/sde
```

3 使用 `pvdisplay` 检查是否所有信息都正确：

```
--- 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-lxB1-mWm2-xn51-agM2-0UuHFC
```

4 在这两个磁盘上创建群集感知卷组：

```
vgcreate --clustered y clustervg /dev/sdd /dev/sde
```

5 使用 `vgdisplay` 检查是否所有信息都正确：

```
--- 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

```

6 根据需要创建逻辑卷：

```
lvcreate --name clusterlv --size 500M clustervg
```

创建卷并启动资源后，应有一个名为 `/dev/dm-0` 的新设备。建议使用 LVM 资源上的群集文件系统，例如 OCFS。有关详细信息，请参见第 12 章 *Oracle Cluster File System 2*（第 139 页）

14.2.2 方案：使用 DRBD 的 cLVM

如果数据中心位于城市、国家/地区或大洲的不同区域，则可使用以下方案。

过程 14.5 创建使用 DRBD 的群集感知卷组

1 创建主/主 DRBD 资源：

1a 首先，按过程 13.1,“手动配置 DRBD”（第 150 页）中所述将 DRBD 设备设置为主/从模式。确保两个节点上的磁盘状态均为 `up-to-date`。使用 `cat /proc/drbd` 或 `rcdrbd status` 检查情况是否如此。

1b 将以下选项添加到配置文件（通常类似 `/etc/drbd.d/r0.res`）：

```
resource r0 {
    startup {
```

```

        become-primary-on both;
    }

    net {
        allow-two-primaries;
    }
    ...
}

```

1c 将更改的配置文件复制到另一个节点，例如：

```
scp /etc/drbd.d/r0.res venus:/etc/drbd.d/
```

1d 在两个节点上运行以下命令：

```

drbdadm disconnect r0
drbdadm connect r0
drbdadm primary r0

```

1e 检查节点的状态：

```

cat /proc/drbd
...
0: cs:Connected ro:Primary/Primary ds:UpToDate/UpToDate C r----

```

2 将 `clvmd` 资源作为克隆品包含在 `Pacemaker` 配置中，并让它依赖于 `DLM` 克隆资源。有关详细指示信息，请参见过程 14.3, “创建 `DLM` 资源”（第 163 页）。继续之前，请确认这些资源已在群集上成功启动。可以使用 `crm_mon` 或 `GUI` 检查正在运行的服务。

3 使用命令 `pvccreate` 准备 `LVM` 的物理卷。例如，在设备 `/dev/drbd_r0` 上，命令应类似于：

```
pvccreate /dev/drbd_r0
```

4 创建群集感知卷组：

```
vgcreate --clustered y myclusterfs /dev/drbd_r0
```

5 根据需要创建逻辑卷。您可能想要更改逻辑卷的大小。例如，使用以下命令创建 4 GB 的逻辑卷：

```
lvcreate --name testlv -L 4G myclusterfs
```

- 6 要确保在群集范围内激活卷组，请按如下方式配置 LVM 资源：

```
primitive vg1 ocf:heartbeat:LVM \  
    params volgrpname="myclusterfs"  
clone vg1-clone vg1 \  
    meta interleave="true" ordered="true"  
colocation colo-vg1 inf: vg1-clone dlm-clvm-clone  
order order-vg1 inf: dlm-clvm-clone vg1-clone
```

- 7 如果希望只在一个节点上激活卷组，可使用以下示例；在这种情况下，cLVM 将防止在多个节点上激活 VG 内的所有逻辑卷，作为非群集应用程序的附加保护措施：

```
primitive vg1 ocf:heartbeat:LVM \  
    params volgrpname="myclusterfs" exclusive="yes"  
colocation colo-vg1 inf: vg1 dlm-clvm-clone  
order order-vg1 inf: dlm-clvm-clone vg1
```

- 8 现在 VG 内的逻辑卷可作为文件系统装入或原始用法提供。确保使用逻辑卷的服务必须具备适当的依赖性，以便在激活 VG 后对它们进行排列和排序。

完成这些配置步骤后，即可像在任何独立工作站中一样进行 LVM2 配置。

14.3 显式配置合格的 LVM2 设备

当若干设备看似共享同一物理卷签名（多路径设备或 DRBD 可能发生这种情况）时，建议显式配置 LVM2 扫描 PV 的设备。

例如，如果命令 `vgcreate` 使用的是物理设备而不是镜像块设备，DRBD 将会感到困惑，并可能导致 DRBD 的裂脑情况。

要停用 LVM2 的单个设备，请执行以下操作：

- 1 编辑文件 `/etc/lvm/lvm.conf` 并搜索以 `filter` 开头的行。
- 2 其中的模式作为正则表达式来处理。前面的“a”表示接受扫描的设备模式，前面的“r”表示拒绝遵守该设备模式的设备。
- 3 要删除名为 `/dev/sdb1` 的设备，请向过滤规则添加以下表达式：

```
"r|^/dev/sdb1$|"
```

完整的过滤行将显示如下：

```
filter = [ "r|^/dev/sdb1$|", "r|/dev/.*/by-path/.*/",  
"r|/dev/.*/by-id/.*/", "a/.*/" ]
```

接受 DRBD 和 MPIO 设备但拒绝所有其他设备的过滤行将显示如下：

```
filter = [ "a|/dev/drbd.*|", "a|/dev/.*/by-id/dm-uuid-mpath-.*/", "r/.*/"  
]
```

4 编写配置文件并将它复制到所有群集节点。

14.4 更多信息

可从 <http://www.clusterlabs.org/wiki/Help:Contents> 处的 pacemaker 邮件列表中获取完整信息。

官方 cLVM 常见问题可在以下网址中找到：<http://sources.redhat.com/cluster/wiki/FAQ/CLVM>。

储存保护

High Availability 群集堆栈的首要任务是保护数据的完整性。这是通过避免未经协调而并发访问数据储存区来实现的：例如，ext3 文件系统只在群集中装入一次，OCFS2 卷只有在与其他群集节点协调后才会装入。在功能良好的群集中，Pacemaker 会检测资源活动是否超出其并发限制，并启动恢复。此外，其策略引擎绝不会超出这些限制。

但是，网络分区或软件故障可能导致选出若干协调程序的情况。如果允许出现这种所谓的“节点分裂”情况，则可能会发生数据损坏。因此，在群集堆栈中增加了若干保护层，以缓解这种情况。

为实现此目标，起作用的主要组件是 IO 屏蔽/STONITH，它可确保在储存激活之前终止所有其他访问。其他机制有 cLVM2 排它激活或 OCFS2 文件锁定支持，以保护系统免受管理或应用程序错误的影响。有了这些机制，再配合进行设置后，就能可靠地避免节点分裂情况所造成的危害。

本章介绍利用储存区本身的 IO 屏蔽机制，后面是对附加保护层（用于确保对储存区的排它访问）的描述。这两套机制可以结合起来使用，以提供更高的保护级别。

15.1 基于储存区的屏蔽

可使用节点分裂检测器 (SBD)、watchdog 支持和 external/sbd STONITH 代理来可靠地避免节点分裂的情况。

15.1.1 概述

在所有节点都可访问共享储存区的环境中，有一个小分区 (1MB) 会格式化为由 SBD 使用。配置完相应的守护程序后，它在其余群集堆栈启动之前将在每个节点上都处于联机状态。它在所有其他群集组件都关闭之后才终止，从而确保了群集资源绝不会在没有 SBD 监督的情况下被激活。

此守护程序会自动将分区上的消息槽之一分配给其自身，并持续监视其中有无发送给它自己的消息。收到消息后，守护程序会立即执行请求，如启动关闭电源或重引导循环以进行屏蔽。

此守护程序会持续监视与储存设备的连接性，并在无法连接分区时自行终止。这就保证了它不会从屏蔽消息断开连接。如果群集数据驻留在不同分区中的同一逻辑单元上，这不是额外的故障点：如果与储存区的连接已丢失，工作负载总是要终止的。

额外的保护是通过 watchdog 支持提供的。现代系统支持 hardware watchdog，后者必须由软件客户端更新，否则硬件会强制执行系统重新启动。这可防止出现 SBD 进程本身的故障，如失去响应或由于 IO 错误而卡住。

15.1.2 设置基于储存区的保护

以下步骤是设置基于储存区的保护所必需的：

- 1 创建 SBD 分区（第 171 页）
- 2 设置软件检查包（第 172 页）
- 3 启动 SBD 守护程序（第 172 页）
- 4 测试 SBD（第 173 页）
- 5 配置屏蔽资源（第 173 页）

以下所有过程都必须以 `root` 用户身份来执行。开始前应确保满足以下要求：

重要：要求

- 环境中必须有所有节点均可到达的共享储存区。
 - 此共享储存段不能使用基于主机的 RAID、cLVM2 或 DRBD。
 - 但是，建议使用基于储存区的 RAID 和多路径，以提高可靠性。
-

创建 SBD 分区

建议在设备启动时创建一个 1MB 的分区。如果 SBD 设备驻留在多路径组上，则需要调整 SBD 所用的超时，因为 MPIO 的沿路径检测可能导致一些等待时间。msgwait 超时后，将假定此消息已传递到节点。对于多路径，这应是 MPIO 检测路径故障并切换到下一个路径所需的时间。可能需要在您的环境中测试此功能。如果它更新检查包计时器的速度不够快，节点就会自行终止。检查包超时时间必须短于 msgwait 超时时间 - 前者是后者的一半是较好的估计值。

在下文中，此 SBD 分区由 `/dev/SBD` 引用。将它替换为实际路径名，例如 `/dev/sdc1`。

重要：覆盖现有数据

确保要用于 SBD 的设备未保存任何数据。sdb 命令不再进一步请求确认就覆盖设备。

1 使用以下命令初始化 SBD 设备：

```
sbd -d /dev/SBD create
```

此操作会将报头写入设备，并创建最多可供 255 个节点以默认时序共享此设备的槽。

2 如果 SBD 设备驻留在多路径组中，请调整 SBD 所用的超时。这可以在初始化 SBD 设备时指定（所有超时的单位都是秒）：

```
/usr/sbin/sbd -d /dev/SBD -4 $msgwait -1 $watchdogtimeout create
```

3 使用以下命令检查已写入设备的内容：

```
sbdd -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
```

正如您看到的，超时数也储存在报头中，以确保所有参与的节点在这方面都一致。

设置软件检查包

强烈建议将 Linux 系统设置为使用检查包。这涉及到在系统引导时装载正确的检查包驱动程序。

- 在 HP 硬件上，这是 `hpwdt` 模块。
- 对于使用 Intel TCO 的系统，可使用 `iTCO_wdt`。`softdog` 是最通用的驱动程序，但建议使用带有实际硬件集成的驱动程序。

有关选项的列表，请参见内核包中的 `drivers/watchdog`。

启动 SBD 守护程序

SBD 守护程序是群集堆栈的关键部分。群集堆栈运行时，甚至部分崩溃时，都必须运行此守护程序，这样才能将其屏蔽。

- 1 要让 OpenAIS init 脚本启动和停止 SDB，请向 `/etc/sysconfig/sbd` 添加以下内容：

```
SBD_DEVICE="/dev/SBD"
# The next line enables the watchdog support:
SBD_OPTS="-W"
```

如果无法访问 SBD 设备，守护程序将不能启动和禁止 OpenAIS 启动。

注意

如果 SBD 设备变得从某个节点无法访问，这会导致此节点进入无限的重引导循环。从技术上来说，这是正确的，但根据您的管理策略，这可能会成为麻烦。您可能希望在这种情况下，引导时不自动启动 OpenAIS。

- 2 在继续下一步之前，请通过执行 `rcopenais restart` 确保所有节点上都启动了 SBD。

测试 SBD

- 1 以下命令会将节点槽及其当前消息从 SBD 设备进行转储：

```
sbd -d /dev/SBD list
```

现在，您应该看到此处列出了使用 SBD 启动过的所有群集节点，并且消息槽应显示 `clear`。

- 2 尝试将测试消息发送到节点之一：

```
sbd -d /dev/SBD message nodea test
```

- 3 此节点将在系统日志中确认收到了该消息：

```
Aug 29 14:10:00 nodea sbd: [13412]: info: Received command test from nodeb
```

这就确认了 SBD 确实在节点上正常运行，并已准备好接收消息。

配置屏蔽资源

- 1 要完成 SBD 设置，必须按如下方式将 SBD 激活为 CIB 中的 STONITH/屏蔽机制：

```
crm configure
crm(live)configure# property stonith-enabled="true"
crm(live)configure# property stonith-timeout="30s"
crm(live)configure# primitive stonith:external/sbd params
sbd_device="/dev/SBD"
crm(live)configure# commit
crm(live)configure# quit
```

由于节点槽是自动分配的，因此无需定义手动主机列表。

- 2 禁用以前可能配置过的任何其他屏蔽设备，因为现在用于此功能的是SBD机制。

现在，启动资源后，群集即已成功配置为共享储存区屏蔽，将在需要屏蔽节点时利用此方法。

15.2 确保储存区的排它激活

此部分将介绍另一种低级别机制：sfex，可将共享储存区的访问以排它的方式锁定于一个节点。请注意，sfex 不会替代 STONITH。由于 sfex 需要共享储存区，因此建议将上述 external/sbd 屏蔽机制用于储存区的另一个分区。

根据设计意图，sfex 不能与需要并发的 workload（如 OCFS2）一起使用，而是用作传统故障转移式 workload 的保护层。这实际上与 SCSI-2 保留相类似，但更具一般性。

15.2.1 概述

在共享储存环境中，储存区的一个小分区专门设置为储存一个或多个锁。

在获取受保护资源之前，节点必须先获取保护锁。此顺序由 Pacemaker 强制实施，sfex 组件可确保即使 Pacemaker 遇到了节点分裂的情况，也不会多次授予锁。

这些锁必须定期刷新，这样某个节点的终止才不会永久性地阻止此锁，其他节点仍可继续操作。

15.2.2 设置

以下内容可帮助您了解如何创建用于 sfex 的共享分区以及如何为 CIB 中的 sfex 锁配置资源。单个 sfex 分区可保存任意数量的锁，默认值为 1，需要为每个锁分配 1 KB 的储存空间。

重要：要求

- sfex 的共享分区应和要保护的数据位于同一逻辑单元上。

- 共享的 **sfex** 分区不能使用基于主机的 RAID 或 DRBD。
- 可以使用 **cLVM2** 逻辑卷。

过程 15.1 创建 *sfex* 分区

- 1 创建用于 **sfex** 的共享分区。注意此分区的名称，并用它替代下面的 `/dev/sfex`。

- 2 使用以下命令创建 **sfex** 元数据：

```
sfex_init -i 1 /dev/sfex
```

- 3 校验元数据已正确创建：

```
sfex_stats -i 1 /dev/sfex ; echo $?
```

此操作应返回 2，因为当前未保存锁。

过程 15.2 为 *sfex* 锁配置资源

- 1 **sfex** 锁通过 CIB 中的资源表示，其配置如下：

```
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 要通过 **sfex** 锁保护资源，请在保护对象和 **sfex** 资源之间创建强制顺序和放置限制。如果受保护资源的 ID 是 `filesystem1`：

```
# order order-sfex-1 inf: sfex_1 filesystem1  
# colocation colo-sfex-1 inf: filesystem1 sfex_1
```

- 3 如果使用组语法，请将 **sfex** 资源添加为组内的第一个资源：

```
# group LAMP sfex_1 filesystem1 apache ipaddr
```


Samba 群集

群集 Samba 服务器提供异构网络的 High Availability 解决方案。本章说明了一些背景信息以及如何设置群集 Samba 服务器。

16.1 概念概述

Samba 使用 Trivial Database (TDB) 已经许多年了。它允许多个应用程序同时写入。为确保所有写操作都成功执行而不会彼此冲突，TDB 使用内部锁定机制。

Cluster Trivial Database (CTDB) 是现有的 TDB 的小扩展。项目本身对 CTDB 的描述是：“Samba 和其他项目用于储存临时数据的 TDB 数据库的群集实现”。

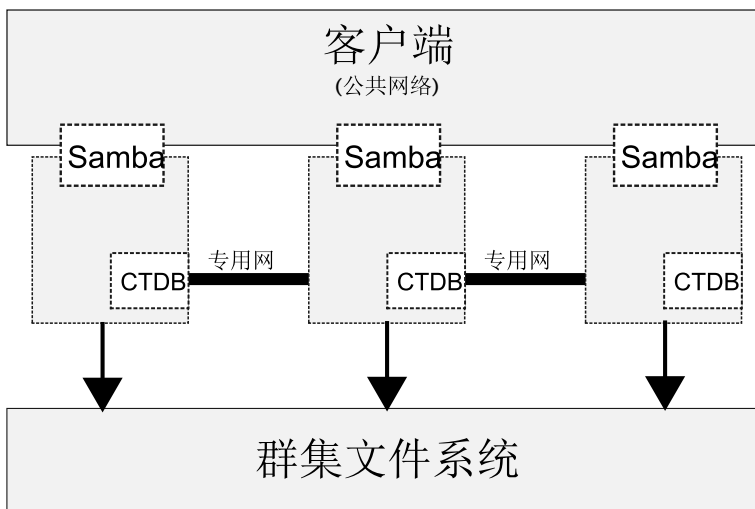
每个群集节点都运行本地 CTDB 守护程序。Samba 与其本地 CTDB 守护程序通讯，而非直接写入其 TDB。守护程序通过网络交换元数据，但实际的读写操作是在快速储存的本地副本上进行的。CTDB 的概念如图 16.1 “CTDB 群集的结构”（第 178 页）中所示。

注意：CTDB 仅用于 Samba

CTDB 资源代理的当前实现将 CTDB 配置为只管理 Samba。任何其他功能，包括 IP 故障转移，都应使用 Pacemaker 进行配置。

此外，CTDB 仅支持完全同类的群集。例如，群集中的所有节点都必须有相同的体系结构，不能混合 i586 与 x86_64。

图 16.1 CTDB 群集的结构



群集 Samba 服务器必须共享某些数据：

- 将 Unix 用户和组 ID 与 Windows 用户和组关联的映射表。
- 用户数据库必须在所有节点间同步。
- Windows 域中的成员服务器的连接信息必须在所有节点上都可用。
- 元数据（如活动 SMB 会话、共享连接和各种锁）必须在所有节点上都可用。

目标是：具有 $N+1$ 个节点的群集 Samba 服务器比只有 N 个节点的快。一个节点不会比非群集 Samba 服务器慢。

16.2 基本配置

注意：已更改的配置文件

CTDB 资源代理会自动更改 `/etc/sysconfig/ctdb` 和 `/etc/samba/smb.conf`。使用 `crm info CTDB` 列出可为 CTDB 资源指定的所有参数。

要设置群集 Samba 服务器，请按如下操作：

1 准备群集：

1a 按本指南的第 II 部分“配置和管理”（第 31 页）中所述配置群集（OpenAIS、acemaker、OCFS2）。

1b 配置共享文件系统（如 OCFS2），并将其装入 /shared 之类的位置。

1c 要打开 POSIX ACL，请启用它：

- 对新的 OCFS2 文件系统，使用：

```
mkfs.ocfs2 --fs-features=xattr ...
```

- 对现有 OCFS2 文件系统，使用：

```
tuneefs.ocfs2 --fs-feature=xattr DEVICE
```

确保在文件系统资源中指定了 `acl` 选项。按如下方式使用 `crm` 外壳：

```
crm(live)configure# primary ocfs2-3 ocf:heartbeat:Filesystem  
options="acl" ...
```

1d 确保服务 `ctdb`、`smb`、`nmb` 和 `winbind` 已禁用：

```
chkconfig ctdb off  
chkconfig smb off  
chkconfig nmb off  
chkconfig winbind off
```

2 在共享文件系统上为 CTDB 锁和 Samba 状态创建目录：

```
mkdir -p /shared/samba/private
```

3 在 /etc/ctdb/nodes 中插入包含群集中每个节点的所有私用 IP 地址的所有节点：

```
192.168.1.10  
192.168.1.11
```

4 将 CTDB 资源添加到群集：

```

crm configure
crm(live)configure# primitive ctdb ocf:heartbeat:CTDB params \
    ctdb_recovery_lock="/shared/samba/ctdb.lock" \
    smb_private_dir="/shared/samba/private" \
    op monitor timeout=20 interval=10
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

```

5 添加群集 IP 地址:

```

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

```

6 检查结果:

```

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: ctdb-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

```

7 从客户端计算机进行测试。在 Linux 客户端上运行以下命令，以检查能否从系统复制文件以及将文件复制到系统:

```
smbclient //192.168.2.222/myshare
```

16.3 调试和测试群集 Samba

要调试群集 Samba 服务器，可使用以下作用于不同级别的工具:

ctdb_diagnostics

运行此工具可诊断群集 Samba 服务器。它可提供大量调试消息，能帮助您找出可能遇到的任何问题。

ctdb_diagnostics 命令可搜索以下文件，这些文件必须在所有节点上都可用：

```
/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
```

如果文件 /etc/ctdb/public_addresses 和 /etc/ctdb/static-routes 存在，它们也会被检查。

ping_pong

检查文件系统是否支持 CTDB 使用 ping_pong。它会对群集文件系统执行一致性和性能之类的特定测试（请参见 http://wiki.samba.org/index.php/Ping_pong），从而给出群集在高负载下将会表现如何的一些预测。

要测试群集文件系统的某些方面，请如下继续操作：

过程 16.1 测试群集文件系统的一致性和性能

- 1 在一个节点上启动命令 ping_pong，将占位符 *N* 替换为节点数 + 1。文件名位于共享储存区上，因此可在所有节点上访问：

```
ping_pong data.txt N
```

应该会得到很高的锁定率，因为只运行一个节点。如果程序不打印锁定率，请替换群集文件系统。

- 2 使用相同的参数在另一个节点上启动 ping_pong 的第二个副本。

应该会看到锁定率急剧下降。如果以下任意情况适用于群集文件系统，请替换它：

- ping_pong 不打印每秒锁定率
- 两个实例中的锁定率并非几乎相等
- 启动第二个实例后锁定率未下降

3 启动ping_pong的第三个副本。添加另一个节点，注意锁定率的变化。

4 逐步停止ping_pong命令。应该观察到锁定率上升，直到回到单一节点的情况。如果未观察到预期行为，请替换群集文件系统。

16.4 更多信息

- [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://ctdb.samba.org>
- http://wiki.samba.org/index.php/Samba_%26_Clustering

部分 IV. 查错和参考

查错

用户经常会遇到奇怪而不易理解的问题（特别是刚开始尝试使用 High Availability 时）。但有几个实用程序可用于近距离地观察 High Availability 的内部进程。本章将推荐各种解决方案。

17.1 安装问题

在安装包或使群集联机的过程中遇到的查错困难。

是否安装了 HA 包？

配置和管理群集所需的包位于 High Availability Extension 提供的高可用性安装模式中。

根据第 3.1 节“安装 High Availability Extension”（第 19 页）中所述，检查是否将 High Availability Extension 作为 SUSE Linux Enterprise Server 11 SP1 的外接式附件安装在每个群集节点上，以及每台计算机上是否安装了高可用性模式。

所有群集节点的初始配置是否相同？

根据第 3.2 节“初始群集设置”（第 20 页）中所述，为了相互通讯，属于同一个群集的所有节点需要使用相同的 `bindnetaddr`、`mcastaddr` 和 `mcastport`。

检查 `/etc/corosync/corosync.conf` 中配置的通讯通道和选项是否对所有群集节点都相同。

如果使用加密通讯，请检查 `/etc/corosync/authkey` 文件是否在所有群集节点上都可用。

除 `nodeid` 以外的所有 `corosync.conf` 设置都必须相同；所有节点上的 `authkey` 文件都必须相同。

防火墙是否允许通过 `mcastport` 进行通讯？

如果用于群集节点之间通讯的 `mcastport` 由防火墙阻止，这些节点将无法相互可见。根据第 3.1 节“安装 High Availability Extension”（第 19 页）中所述，在使用 YaST 配置初始设置时，通常会自动调整防火墙设置。

要确保 `mcastport` 不被防火墙阻止，请检查每个节点上的 `/etc/sysconfig/SuSEfirewall12` 中的设置。或者，在每个群集节点上启动 YaST 防火墙模块。在单击 *允许的服务* > 高级后，将 `mcastport` 添加到允许的 *UDP* 端口列表中并确认更改。

是否在每个群集节点上启动了 OpenAIS？

使用 `/etc/init.d/openais status` 检查每个群集节点上的 OpenAIS 状态。如果 OpenAIS 未在运行，执行 `/etc/init.d/openais start` 命令来启动它。

17.2 “调试” HA 群集

下面显示了资源操作的历史记录（选项 `-o`）和处于不活动状态的资源（`-r`）：

```
crm_mon -o -r
```

状态改变时会刷新显示（要取消，请按 `Ctrl + C`）。示例显示如下：

例 17.1 已停止的资源

Refresh in 10s...

```
=====
Last updated: Mon Jan 19 08:56:14 2009
Current DC: d42 (d42)
3 Nodes configured.
3 Resources configured.
=====

Online: [ d230 d42 ]
OFFLINE: [ clusternode-1 ]

Full list of resources:

Clone Set: o2cb-clone
    Stopped: [ o2cb:0 o2cb:1o2cb:2 ]
Clone Set: dlm-clone
    Stopped [ dlm:0 dlm:1 dlm:2 ]
mySecondIP      (ocf::heartbeat:IPaddr):      Stopped

Operations:
* Node d230:
  aa: migration-threshold=1000000
    + (5) probe: rc=0 (ok)
    + (37) stop: rc=0 (ok)
    + (38) start: rc=0 (ok)
    + (39) monitor: interval=15000ms rc=0 (ok)
* Node d42:
  aa: migration-threshold=1000000
    + (3) probe: rc=0 (ok)
    + (12) stop: rc=0 (ok)
```

首先使您的节点联机（参见第 17.3 节（第 187 页））。然后，检查您的资源和操作。

<http://clusterlabs.org/wiki/Documentation> 下的 *Configuration Explained*（配置说明）PDF 涵盖了 *How Does the Cluster Interpret the OCF Return Codes?*（群集如何解释 OCF 返回代码？）部分中所述的三种不同的恢复类型。

17.3 常见问题解答

我的群集状态是什么？

要检查群集的当前状态，请使用程序 `crm_mon` 或 `crm status` 之一。这将显示当前的 DC 以及当前节点已知的所有节点和资源。

我的群集的一些节点无法互相可见。

这可能有几个原因：

- 先查看配置文件 `/etc/corosync/corosync.conf`，检查群集中每个节点的多路广播地址是否相同（使用关键字 `mcastaddr` 在 `interface` 部分中查找）。
- 检查您的防火墙设置。
- 检查您的交换机是否支持多路广播地址
- 检查节点间的连接是否已断开。这通常是错误配置防火墙的结果。这也可能是节点分裂情况（其中群集已分区）的原因。

我想列出当前已知的资源。

使用命令 `crm_resource -L` 可以了解您的当前资源。

我配置了一个资源，但是它总是失败。

要检查 OCF 脚本，请使用 `ocf-tester`，例如：

```
ocf-tester -n ipl -o ip=YOUR_IP_ADDRESS \  
/usr/lib/ocf/resource.d/heartbeat/IPaddr
```

对更多参数，请多次使用 `-o`。通过运行 `crm ra info AGENT` 可获取必需参数和可选参数的列表，例如：

```
crm ra info ocf:heartbeat:IPaddr
```

运行 `ocf-tester` 之前，请确保资源不受群集管理。

我刚刚收到一条失败消息。有可能收到更多信息吗？

您随时可以向命令添加 `--verbose` 参数。如果您多次执行该操作，该调试输出会变得非常详细。请参见 `/var/log/messages` 了解有用提示。

如何清理我的资源？

使用以下命令：

```
crm resource list  
crm resource cleanup rscid [node]
```

如果遗漏此节点，则资源将在所有节点上清除。更多信息可以在第 6.4.2 节“清理资源”（第 102 页）中找到。

我无法装入 ocfs2 设备。

检查 `/var/log/messages` 中是否有以下行：

```
Jan 12 09:58:55 clusternode2 lrmd: [3487]: info: RA output:
(o2cb:1:start:stderr) 2009/01/12_09:58:55
    ERROR: Could not load ocfs2_stackglue
Jan 12 16:04:22 clusternode2 modprobe: FATAL: Module ocfs2_stackglue not
found.
```

在这种情况下，将缺少内核模块 `ocfs2_stackglue.ko`。请根据安装的内核安装包 `ocfs2-kmp-default`、`ocfs2-kmp-pae` 或 `ocfs2-kmp-xen`。

17.4 更多信息

有关 Linux 和 Heartbeat 上的高可用性的更多信息（包括配置群集资源以及管理和自定义 Heartbeat 群集），请参见 <http://clusterlabs.org/wiki/Documentation>。

群集管理工具

High Availability Extension 附带了一套全面的工具，帮助您从命令行管理群集。本章主要介绍管理 CIB 中的群集配置和群集资源所需的工具。用于管理资源代理的其他命令行工具或用于调试设置（和查错）的工具在第 17 章 [查错](#)（第 185 页）中有所介绍。

以下列表提供了一些与群集管理相关的任务，并简要介绍了完成这些任务所使用的工具：

监视群集的状态

`crm_mon` 命令可用于监视您的群集状态和配置。其输出包括节点数、`uname`、`uuid`、状态、群集中配置的资源及其各自的当前状态。`crm_mon` 的输出可显示在控制台上或打印到 HTML 文件。当具有不包含状态部分的群集配置文件时，`crm_mon` 会按文件中所指定的方式创建节点和资源概览。有关对此工具的使用和命令语法的详细介绍，请参见 [crm_mon\(8\)](#)（第 213 页）。

管理 CIB

`cibadmin` 命令是用于操作 **Heartbeat CIB** 的低级管理命令。它可用于转储、更新和修改全部或部分 CIB，删除整个 CIB 或执行其他 CIB 管理操作。有关对此工具的使用和命令语法的详细介绍，请参见 [cibadmin\(8\)](#)（第 193 页）。

管理配置更改

`crm_diff` 命令可帮助您创建和应用 XML 增补程序。它对于观察群集配置的两个版本之间的更改或保存这些更改供日后使用 [cibadmin\(8\)](#)（第 193 页）来应用它们非常有用。有关对此工具的使用和命令语法的详细介绍，请参见 [crm_diff\(8\)](#)（第 205 页）。

操作 CIB 属性

您可以使用 `crm_attribute` 命令来查询和操作 CIB 中使用的节点属性和群集配置选项。有关对此工具的使用和命令语法的详细介绍，请参见 `crm_attribute(8)`（第 202 页）。

验证群集配置

`crm_verify` 命令可检查配置数据库 (CIB) 的一致性和其他问题。它可检查包含配置的文件或连接到运行中的群集。它可报告两类问题。虽然警告解决方法已经传达到管理员，但是必须修复错误 `Heartbeat` 才能正常工作。`crm_verify` 可帮助创建新的或已修改的配置。您可以本地复制运行的群集中的 CIB，编辑它，使用 `crm_verify` 验证它，然后使用 `cibadmin` 使新配置生效。有关对此工具的使用和命令语法的详细介绍，请参见 `crm_verify(8)`（第 241 页）。

管理资源配置

`crm_resource` 命令可在群集上执行各种资源相关的操作。它可以修改已配置资源的定义，启动和停止资源，删除资源或在节点间迁移资源。有关对此工具的使用和命令语法的详细介绍，请参见 `crm_resource(8)`（第 217 页）。

管理资源故障计数

`crm_failcount` 命令可查询指定节点上每个资源的故障计数。此工具还可用于重置故障计数，同时允许资源在它多次失败的节点上再次运行。有关对此工具的使用和命令语法的详细介绍，请参阅 `crm_failcount(8)`（第 208 页）。

管理节点的备用状态

`crm_standby` 命令可操作节点的备用属性。备用模式中的所有节点都不再具备主管资源的资格，并且必须移动那里的所有资源。备用模式对于执行维护任务（如内核更新）很有用。从节点删除备用属性，使之再次成为群集中完全处于活动状态的成员。有关对此工具的使用和命令语法的详细介绍，请参见 `crm_standby(8)`（第 238 页）。

cibadmin (8)

cibadmin — Provides direct access to the cluster configuration

Synopsis

Allows the configuration, or sections of it, to be queried, modified, replaced and/or deleted.

```
cibadmin (--query|-Q) -[Vrwlsmfbp] [-i xml-object-id|-o
    xml-object-type] [-t t-flag-whatever] [-h hostname]

cibadmin (--create|-C) -[Vrwlsmfbp] [-X xml-string]
    [-x xml- filename] [-t t-flag-whatever] [-h hostname]

cibadmin (--replace|-R) -[Vrwlsmfbp] [-i xml-object-id|
    -o xml-object-type] [-X xml-string] [-x xml-filename] [-t
    t-flag- whatever] [-h hostname]

cibadmin (--update|-U) -[Vrwlsmfbp] [-i xml-object-id|
    -o xml-object-type] [-X xml-string] [-x xml-filename] [-t
    t-flag- whatever] [-h hostname]

cibadmin (--modify|-M) -[Vrwlsmfbp] [-i xml-object-id|
    -o xml-object-type] [-X xml-string] [-x xml-filename] [-t
    t-flag- whatever] [-h hostname]

cibadmin (--delete|-D) -[Vrwlsmfbp] [-i xml-object-id|
    -o xml-object-type] [-t t-flag-whatever] [-h hostname]

cibadmin (--delete_alt|-d) -[Vrwlsmfbp] -o
    xml-object-type [-X xml-string|-x xml-filename]
    [-t t-flag-whatever] [-h hostname]

cibadmin --erase (-E)

cibadmin --bump (-B)

cibadmin --ismaster (-m)

cibadmin --master (-w)

cibadmin --slave (-r)

cibadmin --sync (-S)

cibadmin --help (-?)
```

Description

The `cibadmin` command is the low-level administrative command for manipulating the Heartbeat CIB. Use it 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.

`cibadmin` operates on the XML trees of the CIB, largely without knowledge of the purpose of the updates or queries performed. This means that shortcuts that seem natural to users who understand the meaning of the elements in the XML tree are impossible to use with `cibadmin`. It requires a complete lack of ambiguity and can only deal with valid XML subtrees (tags and elements) for both input and output.

注意

`cibadmin` should always be used in preference to editing the `cib.xml` file by hand—especially if the cluster is active. The cluster goes to great lengths to detect and discourage this practice so that your data is not lost or corrupted.

Options

`--obj_type object-type, -o object-type`

Specify the type of object on which to operate. Valid values are `nodes`, `resources`, `constraints`, `crm_status`, and `status`.

`--verbose, -V`

Turn on debug mode. Additional `-V` options increase the detail and frequency of the output.

`--help, -?`

Obtain a help message from `cibadmin`.

`--xpath PATHSPEC, -A PATHSPEC`

Supply a valid XPath to use instead of an `obj_type`.

Commands

`--bump, -B`

Increase the `epoch` version counter in the CIB. Normally this value is increased automatically by the cluster when a new leader is elected. Manually increasing it can be useful if you want to make an older configuration obsolete (such as one stored on inactive cluster nodes).

`--create, -C`

Create a new CIB from the XML content of the argument.

`--delete, -D`

Delete the first object matching the supplied criteria, for example, `<op id="rsc1_op1" name="monitor"/>`. The tag name and all attributes must match in order for the element to be deleted

`--erase, -E`

Erase the contents of the entire CIB.

`--ismaster, -m`

Print a message indicating whether or not the local instance of the CIB software is the master instance or not. Exits with return code 0 if it is the master instance or 35 if not.

`--modify, -M`

Find the object somewhere in the CIB's XML tree and update it.

`--query, -Q`

Query a portion of the CIB.

`--replace, -R`

Recursively replace an XML object in the CIB.

`--sync, -S`

Force a resync of all nodes with the CIB on the specified host (if `-h` is used) or with the DC (if no `-h` option is used).

XML Data

`--xml-text string, -X string`

Specify an XML tag or fragment on which `crmadmin` should operate. It must be a complete tag or XML fragment.

`--xml-file filename, -x filename`

Specify the XML from a file on which `cibadmin` should operate. It must be a complete tag or an XML fragment.

`--xml_pipe, -p`

Specify that the XML on which `cibadmin` should operate comes from standard input. It must be a complete tag or an XML fragment.

Advanced Options

`--host hostname, -h hostname`

Send command to specified host. Applies to `query` and `sync` commands only.

`--local, -l`

Let a command take effect locally (rarely used, advanced option).

`--no-bcast, -b`

Command will not be broadcast even if it altered the CIB.

重要

Use this option with care to avoid ending up with a divergent cluster.

`--sync-call, -s`

Wait for call to complete before returning.

Examples

To get a copy of the entire active CIB (including status section, etc.) delivered to stdout, issue this command:

```
cibadmin -Q
```


To add an IPaddr2 resource to the *resources* section, first create a file `foo` with the following contents:

```
<primitive id="R_10.10.10.101" class="ocf" type="IPaddr2"
  provider="heartbeat">
  <instance_attributes id="RA_R_10.10.10.101">
    <attributes>
      <nvpair id="R_ip_P_ip" name="ip" value="10.10.10.101"/>
      <nvpair id="R_ip_P_nic" name="nic" value="eth0"/>
    </attributes>
  </instance_attributes>
</primitive>
```

Then issue the following command:

```
cibadmin --obj_type resources -U -x foo
```

To change the IP address of the IPaddr2 resource previously added, issue the command below:

```
cibadmin -M -X '<nvpair id="R_ip_P_ip" name="ip" value="10.10.10.102"/>'
```

注意

This does not change the resource name to match the new IP address. To do that, delete then re-add the resource with a new ID tag.

To stop (disable) the IP address resource added previously, and without removing it, create a file called `bar` with the following content in it:

```
<primitive id="R_10.10.10.101">
  <instance_attributes id="RA_R_10.10.10.101">
    <attributes>
      <nvpair id="stop_R_10.10.10.101" name="target-role" value="Stopped"/>
    </attributes>
  </instance_attributes>
</primitive>
```

Then issue the following command:

```
cibadmin --obj_type resources -U -x bar
```

To restart the IP address resource stopped by the previous step, issue:

```
cibadmin -D -X '<nvpair id="stop_R_10.10.10.101">'
```

To completely remove the IP address resource from the CIB, issue this command:

```
cibadmin -D -X '<primitive id="R_10.10.10.101"/>'
```

To replace the CIB with a new manually-edited version of the CIB, use the following command:

```
cibadmin -R -x $HOME/cib.xml
```

Files

`/var/lib/heartbeat/crm/cib.xml`—the CIB (minus status section) on disk.

See Also

`crm_resource(8)` (第 217 页), `crmadmin(8)` (第 199 页), `lrmadmin(8)`, `heartbeat(8)`

Caveats

Avoid working on the automatically maintained copy of the CIB on the local disk. Whenever anything in the cluster changes, the CIB is updated. Therefore using an outdated backup copy of the CIB to propagate your configuration changes might result in an inconsistent cluster.

crmadmin (8)

crmadmin — controls the Cluster Resource Manager

Synopsis

```
crmadmin [-V|-q] [-i|-d|-K|-S|-E] node
```

```
crmadmin [-V|-q] -N -B
```

```
crmadmin [-V|-q] -D
```

```
crmadmin -v
```

```
crmadmin -?
```

Description

`crmadmin` was originally designed to control most of the actions of the CRM daemon. However, the largest part of its functionality has been made obsolete by other tools, such as `crm_attribute` and `crm_resource`. Its remaining functionality is mostly related to testing and the status of the `crmd` process.

警告

Some `crmadmin` options are geared towards testing and cause trouble if used incorrectly. In particular, do not use the `--kill` or `--election` options unless you know exactly what you are doing.

Options

`--help, -?`

Print the help text.

`--version, -v`

Print version details for HA, CRM, and CIB feature set.

`--verbose, -V`

Turn on command debug information.

注意

Increase the level of verbosity by providing additional instances.

`--quiet, -q`

Do not provide any debug information at all and reduce the output to a minimum.

`--bash-export, -B`

Create bash export entries of the form `export uname=uuid`. This applies only to the `crmadmin -N node` command.

注意

The `-B` functionality is rarely useful and may be removed in future versions.

Commands

`--debug_inc node, -i node`

Incrementally increase the CRM daemon's debug level on the specified node. This can also be achieved by sending the USR1 signal to the `crmd` process.

`--debug_dec node, -d node`

Incrementally decrease the CRM daemon's debug level on the specified node. This can also be achieved by sending the USR2 signal to the `crmd` process.

`--kill node, -K node`

Shut down the CRM daemon on the specified node.

警告

Use this with extreme caution. This action should normally only be issued by Heartbeat and may have unintended side effects.

`--status node, -S node`

Query the status of the CRM daemon on the specified node.

The output includes a general health indicator and the internal FSM state of the `crmd` process. This can be helpful when determining what the cluster is doing.

`--election node, -E node`

Initiate an election from the specified node.

警告

Use this with extreme caution. This action is normally initiated internally and may have unintended side effects.

`--dc_lookup, -D`

Query the uname of the current DC.

The location of the DC is only of significance to the `crmd` internally and is rarely useful to administrators except when deciding on which node to examine the logs.

`--nodes, -N`

Query the uname of all member nodes. The results of this query may include nodes in `offline` mode.

注意

The `-i`, `-d`, `-K`, and `-E` options are rarely used and may be removed in future versions.

See Also

`crm_attribute(8)` (第 202 页), `crm_resource(8)` (第 217 页)

crm_attribute (8)

crm_attribute — Allows node attributes and cluster options to be queried, modified and deleted

Synopsis

```
crm_attribute [options]
```

Description

The `crm_attribute` command queries and manipulates node attributes and cluster configuration options that are used in the CIB.

Options

`--help, -?`
Print a help message.

`--verbose, -V`
Turn on debug information.

注意

Increase the level of verbosity by providing additional instances.

`--quiet, -Q`
When doing an attribute query using `-G`, print just the value to stdout. Use this option with `-G`.

`--get-value, -G`
Retrieve, rather than set, the preference.

`--delete-attr, -D`
Delete, rather than set, the attribute.

`--attr-id string, -i string`
For advanced users only. Identifies the id attribute.

`--attr-value string, -v string`
Value to set. This is ignored when used with `-G`.

`--node node_name, -N node_name`
The uname of the node to change

`--set-name string, -s string`
Specify the set of attributes in which to read or write the attribute.

`--attr-name string, -n string`
Specify the attribute to set or query.

`--type string, -t type`
Determine to which section of the CIB the attribute should be set or to which section of the CIB the attribute that is queried belongs. Possible values are `nodes`, `status`, or `crm_config`.

Examples

Query the value of the `location` attribute in the `nodes` section for the host `myhost` in the CIB:

```
crm_attribute -G -t nodes -U myhost -n location
```

Query the value of the `cluster-delay` attribute in the `crm_config` section in the CIB:

```
crm_attribute -G -t crm_config -n cluster-delay
```

Query the value of the `cluster-delay` attribute in the `crm_config` section in the CIB. Print just the value:

```
crm_attribute -G -Q -t crm_config -n cluster-delay
```

Delete the `location` attribute for the host `myhost` from the `nodes` section of the CIB:

```
crm_attribute -D -t nodes -U myhost -n location
```

Add a new attribute called `location` with the value of `office` to the `set` subsection of the `nodes` section in the CIB (settings applied to the host *myhost*):

```
crm_attribute -t nodes -U myhost -s set -n location -v office
```

Change the value of the `location` attribute in the `nodes` section for the *myhost* host:

```
crm_attribute -t nodes -U myhost -n location -v backoffice
```

Files

`/var/lib/heartbeat/crm/cib.xml`—the CIB (minus status section) on disk. Editing this file directly is strongly discouraged.

See Also

`cibadmin(8)` (第 193 页)

crm_diff (8)

`crm_diff` — identify changes to the cluster configuration and apply patches to the configuration files

Synopsis

```
crm_diff [-?|-V] [-o filename] [-O string] [-p filename] [-n filename] [-N string]
```

Description

The `crm_diff` command assists 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`.

Options

`--help, -?`

Print a help message.

`--original filename, -o filename`

Specify the original file against which to diff or apply patches.

`--new filename, -n filename`

Specify the name of the new file.

`--original-string string, -O string`

Specify the original string against which to diff or apply patches.

`--new-string string, -N string`

Specify the new string.

`--patch filename, -p filename`

Apply a patch to the original XML. Always use with `-o`.

`--cib, -c`

Compare or patch the inputs as a CIB. Always specify the base version with `-o` and provide either the patch file or the second version with `-p` or `-n`, respectively.

`--stdin, -s`

Read the inputs from stdin.

Examples

Use `crm_diff` to determine the differences between various CIB configuration files and to create patches. By means of patches, easily reuse configuration parts without having to use the `cibadmin` command on every single one of them.

- 1 Obtain the two different configuration files by running `cibadmin` on the two cluster setups to compare:

```
cibadmin -Q > cib1.xml
cibadmin -Q > cib2.xml
```

- 2 Determine whether to diff the entire files against each other or compare just a subset of the configurations.

- 3 To print the difference between the files to stdout, use the following command:

```
crm_diff -o cib1.xml -n cib2.xml
```

- 4 To print the difference between the files to a file and create a patch, use the following command:

```
crm_diff -o cib1.xml -n cib2.xml > patch.xml
```

- 5 Apply the patch to the original file:

```
crm_diff -o cib1.xml -p patch.xml
```

Files

`/var/lib/heartbeat/crm/cib.xml`—the CIB (minus status section) on disk. Editing this file directly is strongly discouraged.

See Also

cibadmin(8) (第 193 页)

crm_failcount (8)

crm_failcount — Manage the counter recording each resource's failures

Synopsis

```
crm_failcount [-?|-V] -D -u|-U node -r resource
crm_failcount [-?|-V] -G -u|-U node -r resource
crm_failcount [-?|-V] -v string -u|-U node -r resource
```

Description

Heartbeat implements a sophisticated method to compute and force failover of a resource to another node in case that resource tends to fail on the current node. A resource carries a `resource-stickiness` attribute to determine how much it prefers to run on a certain node. It also carries a `migration-threshold` that determines the threshold at which the resource should failover to another node.

The `failcount` attribute is added to the resource and increased on resource monitoring failure. The value of `failcount` multiplied by the value of `migration-threshold` determines the *failover score* of this resource. If this number exceeds the preference set for this resource, the resource is moved to another node and not run again on the original node until the failure count is reset.

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 run again on nodes where it had previously failed too many times.

Options

```
--help, -?
    Print a help message.

--verbose, -V
    Turn on debug information.
```

注意

Increase the level of verbosity by providing additional instances.

`--quiet, -Q`

When doing an attribute query using `-G`, print just the value to stdout. Use this option with `-G`.

`--get-value, -G`

Retrieve rather than set the preference.

`--delete-attr, -D`

Specify the attribute to delete.

`--attr-id string, -i string`

For advanced users only. Identifies the id attribute.

`--attr-value string, -v string`

Specify the value to use. This option is ignored when used with `-G`.

`--node node_uname, -U node_uname`

Specify the uname of the node to change.

`--resource-id resource name, -r resource name`

Specify the name of the resource on which to operate.

Examples

Reset the failcount for the resource `myrsc` on the node `node1`:

```
crm_failcount -D -U node1 -r my_rsc
```

Query the current failcount for the resource `myrsc` on the node `node1`:

```
crm_failcount -G -U node1 -r my_rsc
```

Files

`/var/lib/heartbeat/crm/cib.xml`—the CIB (minus status section) on disk. Editing this file directly is strongly discouraged.

See Also

`crm_attribute(8)` (第 202 页), `cibadmin(8)` (第 193 页), and the Linux High Availability FAQ Web site [http://www.linux-ha.org/v2/faq/forced_failover]

crm_master (8)

`crm_master` — Manage a master/slave resource's preference for being promoted on a given node

Synopsis

```
crm_master [-V|-Q] -D [-l lifetime]  
crm_master [-V|-Q] -G [-l lifetime]  
crm_master [-V|-Q] -v string [-l string]
```

Description

`crm_master` is called from inside the resource agent scripts to determine which resource instance should be promoted to master mode. It should never be used from the command line and is just a helper utility for the resource agents. RAs use `crm_master` to promote a particular instance to master mode or to remove this preference from it. By assigning a lifetime, determine whether this setting should survive a reboot of the node (set lifetime to `forever`) or whether it should not survive a reboot (set lifetime to `reboot`).

A resource agent needs to determine on which resource `crm_master` should operate. These queries must be handled inside the resource agent script. The actual calls of `crm_master` follow a syntax similar to those of the `crm_attribute` command.

Options

`--help, -?`
Print a help message.

`--verbose, -V`
Turn on debug information.

注意

Increase the level of verbosity by providing additional instances.

`--quiet, -Q`

When doing an attribute query using `-G`, print just the value to stdout. Use this option with `-G`.

`--get-value, -G`

Retrieve rather than set the preference to be promoted.

`--delete-attr, -D`

Delete rather than set the attribute.

`--attr-id string, -i string`

For advanced users only. Identifies the id attribute.

`--attr-value string, -v string`

Value to set. This is ignored when used with `-G`.

`--lifetime string, -l string`

Specify how long the preference lasts. Possible values are `reboot` or `forever`.

Environment Variables

`OCF_RESOURCE_INSTANCE`—the name of the resource instance

Files

`/var/lib/heartbeat/crm/cib.xml`—the CIB (minus status section) on disk.

See Also

`cibadmin(8)` (第 193 页), `crm_attribute(8)` (第 202 页)

crm_mon (8)

crm_mon — monitor the cluster's status

Synopsis

```
crm_mon [-V] -d -pfilename -h filename
crm_mon [-V] [-l|-n|-r] -h filename
crm_mon [-V] [-n|-r] -X filename
crm_mon [-V] [-n|-r] -c|-l
crm_mon [-V] -i interval
crm_mon -?
```

Description

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.

Options

`--help, -?`

Provide help.

`--verbose, -V`

Increase the debug output.

`--interval seconds, -i seconds`

Determine the update frequency. If `-i` is not specified, the default of 15 seconds is assumed.

`--group-by-node, -n`
Group resources by node.

`--inactive, -r`
Display inactive resources.

`--simple-status, -s`
Display the cluster status once as a simple one line output (suitable for nagios).

`--one-shot, -l`
Display the cluster status once on the console then exit (does not use ncurses).

`--as-html filename, -h filename`
Write the cluster's status to the specified file.

`--web-cgi, -w`
Web mode with output suitable for CGI.

`--daemonize, -d`
Run in the background as a daemon.

`--pid-file filename, -p filename`
Specify the daemon's pid file.

Examples

Display your cluster's status and get an updated listing every 15 seconds:

```
crm_mon
```

Display your cluster's status and get an updated listing after an interval specified by `-i`. If `-i` is not given, the default refresh interval of 15 seconds is assumed:

```
crm_mon -i interval[s]
```

Display your cluster's status on the console:

```
crm_mon -c
```

Display your cluster's status on the console just once then exit:

```
crm_mon -l
```

Display your cluster's status and group resources by node:

```
crm_mon -n
```

Display your cluster's status, group resources by node, and include inactive resources in the list:

```
crm_mon -n -r
```

Write your cluster's status to an HTML file:

```
crm_mon -h filename
```

Run `crm_mon` as a daemon in the background, specify the daemon's pid file for easier control of the daemon process, and create HTML output. This option allows you to constantly create HTML output that can be easily processed by other monitoring applications:

```
crm_mon -d -p filename -h filename
```

Display the cluster configuration laid out in an existing cluster configuration file (*filename*), group the resources by node, and include inactive resources. This command can be used for dry runs of a cluster configuration before rolling it out to a live cluster.

```
crm_mon -r -n -X filename
```

Files

`/var/lib/heartbeat/crm/cib.xml`—the CIB (minus status section) on disk. Editing this file directly is strongly discouraged.

crm_node (8)

crm_node — Lists the members of a cluster

Synopsis

```
crm_node [-V] [-p|-e|-q]
```

Description

Lists the members of a cluster.

Options

- V
be verbose
- partition, -p
print the members of this partition
- epoch, -e
print the epoch this node joined the partition
- quorum, -q
print a 1 if our partition has quorum

crm_resource (8)

crm_resource — Perform tasks related to cluster resources

Synopsis

```
crm_resource [-?|-V|-S] -L|-Q|-W|-D|-C|-P|-p [options]
```

Description

The `crm_resource` command performs various resource-related actions on the cluster. It can modify the definition of configured resources, start and stop resources, and delete and migrate resources between nodes.

`--help, -?`

Print the help message.

`--verbose, -V`

Turn on debug information.

注意

Increase the level of verbosity by providing additional instances.

`--quiet, -Q`

Print only the value on stdout (for use with `-W`).

Commands

`--list, -L`

List all resources.

`--query-xml, -x`

Query a resource.

Requires: `-r`

`--locate, -W`
Locate a resource.

Requires: `-r`

`--migrate, -M`
Migrate a resource from its current location. Use `-N` to specify a destination.

If `-N` is not specified, the resource is forced to move by creating a rule for the current location and a score of `-INFINITY`.

注意

This prevents the resource from running on this node until the constraint is removed with `-U`.

Requires: `-r`, Optional: `-N, -f`

`--un-migrate, -U`
Remove all constraints created by `-M`

Requires: `-r`

`--delete, -D`
Delete a resource from the CIB.

Requires: `-r, -t`

`--cleanup, -C`
Delete a resource from the LRM.

Requires: `-r`. Optional: `-H`

`--reprobe, -P`
Recheck for resources started outside the CRM.

Optional: `-H`

`--refresh, -R`
Refresh the CIB from the LRM.

Optional: -H

`--set-parameter string, -p string`

Set the named parameter for a resource.

Requires: -r, -v. Optional: -i, -s, and --meta

`--get-parameter string, -g string`

Get the named parameter for a resource.

Requires: -r. Optional: -i, -s, and --meta

`--delete-parameter string, -d string`

Delete the named parameter for a resource.

Requires: -r. Optional: -i, and --meta

`--list-operations string, -O string`

List the active resource operations. Optionally filtered by resource, node, or both.

Optional: -N, -r

`--list-all-operations string, -o string`

List all resource operations. Optionally filtered by resource, node, or both. Optional:

-N, -r

Options

`--resource string, -r string`

Specify the resource ID.

`--resource-type string, -t string`

Specify the resource type (primitive, clone, group, etc.).

`--property-value string, -v string`

Specify the property value.

`--node string, -N string`

Specify the hostname.

`--meta`

Modify a resource's configuration option rather than one which is passed to the resource agent script. For use with `-p`, `-g` and `-d`.

`--lifetime string, -u string`

Lifespan of migration constraints.

`--force, -f`

Force the resource to move by creating a rule for the current location and a score of `-INFINITY`

This should be used if the resource's stickiness and constraint scores total more than `INFINITY` (currently 100,000).

注意

This prevents the resource from running on this node until the constraint is removed with `-U`.

`-s string`

(Advanced Use Only) Specify the ID of the `instance_attributes` object to change.

`-i string`

(Advanced Use Only) Specify the ID of the `nvpair` object to change or delete.

Examples

Listing all resources:

```
crm_resource -L
```

Checking where a resource is running (and if it is):

```
crm_resource -W -r my_first_ip
```

If the `my_first_ip` resource is running, the output of this command reveals the node on which it is running. If it is not running, the output shows this.

Start or stop a resource:

```
crm_resource -r my_first_ip -p target_role -v started
crm_resource -r my_first_ip -p target_role -v stopped
```

Query the definition of a resource:

```
crm_resource -Q -r my_first_ip
```

Migrate a resource away from its current location:

```
crm_resource -M -r my_first_ip
```

Migrate a resource to a specific location:

```
crm_resource -M -r my_first_ip -H c001n02
```

Allow a resource to return to its normal location:

```
crm_resource -U -r my_first_ip
```

注意

The values of `resource_stickiness` and `default_resource_stickiness` may mean that it does not move back. In such cases, you should use `-M` to move it back before running this command.

Delete a resource from the CRM:

```
crm_resource -D -r my_first_ip -t primitive
```

Delete a resource group from the CRM:

```
crm_resource -D -r my_first_group -t group
```

Disable resource management for a resource in the CRM:

```
crm_resource -p is-managed -r my_first_ip -t primitive -v off
```

Enable resource management for a resource in the CRM:

```
crm_resource -p is-managed -r my_first_ip -t primitive -v on
```

Reset a failed resource after having been manually cleaned up:

```
crm_resource -C -H c001n02 -r my_first_ip
```

Recheck all nodes for resources started outside the CRM:

```
crm_resource -P
```

Recheck one node for resources started outside the CRM:

```
crm_resource -P -H c001n02
```

Files

`/var/lib/heartbeat/crm/cib.xml`—the CIB (minus status section) on disk. Editing this file directly is strongly discouraged.

See Also

`cibadmin(8)` (第 193 页), `crmadmin(8)` (第 199 页), `lrmdadmin(8)`, `heartbeat(8)`

crm_shadow (8)

crm_shadow — Perform Configuration Changes in a Sandbox Before Updating The Live Cluster

Synopsis

```
crm_shadow [-V] [-p|-e|-q]
```

Description

Sets up an environment in which configuration tools (`cibadmin`, `crm_resource`, etc) work offline instead of against a live cluster, allowing changes to be previewed and tested for side-effects.

Options

- `--verbose, -V`
turn on debug info. additional instance increase verbosity
- `--which, -w`
indicate the active shadow copy
- `--display, -p`
display the contents of the shadow copy
- `--diff, -d`
display the changes in the shadow copy
- `--create-empty, -eNAME`
create the named shadow copy with an empty cluster configuration
- `--create, -cNAME`
create the named shadow copy of the active cluster configuration

`--reset, -rNAME`
 recreate the named shadow copy from the active cluster configuration

`--commit, -cNAME`
 upload the contents of the named shadow copy to the cluster

`--delete, -dNAME`
 delete the contents of the named shadow copy

`--edit, -eNAME`
 Edit the contents of the named shadow copy with your favorite editor

`--batch, -b`
 do not spawn a new shell

`--force, -f`
 do not spawn a new shell

`--switch, -s`
 switch to the named shadow copy

Internal Commands

To work with a shadow configuration, you need to create one first:

```
crm_shadow --create-empty YOUR_NAME
```

It gives you an internal shell like the one from the `crm` tool. Use `help` to get an overview of all internal commands, or `help subcommand` for a specific command.

表 18.1 *Overview of Internal Commands*

Command	Syntax/Description
<code>alias</code>	<pre>alias [-p] [name[=value] ...]</pre> <p><code>alias</code> with no arguments or with the <code>-p</code> option prints the list of aliases in the form <code>alias NAME=VALUE</code> on standard output. Otherwise, an alias is defined for each <code>NAME</code> whose <code>VALUE</code> is given. A trailing space in <code>VALUE</code> causes the next word to be checked for alias</p>

Command	Syntax/Description
	substitution when the alias is expanded. Alias returns true unless a NAME is given for which no alias has been defined.
bg	bg [JOB_SPEC ...] Place each JOB_SPEC in the background, as if it had been started with &. If JOB_SPEC is not present, the shell's notion of the current job is used.
bind	bind [-lpvsPVS] [-m keymap] [-f filename] [-q name] [-u name] [-r keyseq] [-x keyseq:shell-command] [keyseq:readline-function or readline-command] Bind a key sequence to a Readline function or a macro, or set a Readline variable. The non-option argument syntax is equivalent to that found in ~/.inputrc, but must be passed as a single argument: bind "\C-x\C-r": re-read-init-file.
break	break [N] Exit from within a for, while or until loop. If N is specified, break N levels.
builtin	builtin [shell-builtin [arg ...]] Run a shell builtin. This is useful when you wish to rename a shell builtin to be a function, but need the functionality of the builtin within the function itself.
caller	caller [EXPR] Returns the context of the current subroutine call. Without EXPR, returns \$line \$filename. With EXPR, returns \$line \$subroutine \$filename; this extra information can be used to provide a stack trace.
case	case WORD in [PATTERN [PATTERN] [COMMANDS;;] ... esac

Command	Syntax/Description
	Selectively execute <i>COMMANDS</i> based upon <i>WORD</i> matching <i>PATTERN</i> . The ' ' is used to separate multiple patterns.
cd	cd [-L -P] [dir] Change the current directory to DIR.
command	command [-pVv] command [arg ...] Runs <i>COMMAND</i> with <i>ARGS</i> ignoring shell functions. If you have a shell function called 'ls', and you wish to call the command 'ls', you can say "command ls". If the -p option is given, a default value is used for PATH that is guaranteed to find all of the standard utilities. If the -V or -v option is given, a string is printed describing <i>COMMAND</i> . The -V option produces a more verbose description.
compgen	compgen [-abcdefgjkusv] [-o option] [-A action] [-G globpat] [-W wordlist] [-P prefix] [-S suffix] [-X filterpat] [-F function] [-C command] [WORD] Display the possible completions depending on the options. Intended to be used from within a shell function generating possible completions. If the optional <i>WORD</i> argument is supplied, matches against <i>WORD</i> are generated.
complete	complete [-abcdefgjkusv] [-pr] [-o option] [-A action] [-G globpat] [-W wordlist] [-P prefix] [-S suffix] [-X filterpat] [-F function] [-C command] [name ...] For each <i>NAME</i> , specify how arguments are to be completed. If the -p option is supplied, or if no options are supplied, existing completion specifications are printed in a way that allows them to be reused as input. The -r option removes a completion specification for each <i>NAME</i> , or, if no <i>NAMES</i> are supplied, all completion specifications.
continue	continue [N]

Command	Syntax/Description
	Resume the next iteration of the enclosing FOR, WHILE or UNTIL loop. If <i>N</i> is specified, resume at the <i>N</i> -th enclosing loop.
<code>declare</code>	<code>declare [-afirtx] [-p] [name[=value] ...]</code> Declare variables and/or give them attributes. If no <i>NAMES</i> are given, then display the values of variables instead. The <code>-p</code> option will display the attributes and values of each <i>NAME</i> .
<code>dirs</code>	<code>dirs [-clpv] [+N] [-N]</code> Display the list of currently remembered directories. Directories find their way onto the list with the <code>pushd</code> command; you can get back up through the list with the <code>popd</code> command.
<code>disown</code>	<code>disown [-h] [-ar] [JOBSPEC ...]</code> By default, removes each <i>JOBSPEC</i> argument from the table of active jobs. If the <code>-h</code> option is given, the job is not removed from the table, but is marked so that SIGHUP is not sent to the job if the shell receives a SIGHUP. The <code>-a</code> option, when <i>JOBSPEC</i> is not supplied, means to remove all jobs from the job table; the <code>-r</code> option means to remove only running jobs.
<code>echo</code>	<code>echo [-neE] [arg ...]</code> Output the ARGs. If <code>-n</code> is specified, the trailing newline is suppressed. If the <code>-e</code> option is given, interpretation of the following backslash-escaped characters is turned on: \a (alert, bell) \b (backspace) \c (suppress trailing newline) \E (escape character) \f (form feed) \n (new line)

Command	Syntax/Description
	<p> \ <i>r</i> (carriage return) \ <i>t</i> (horizontal tab) \ <i>v</i> (vertical tab) \ <i>\</i> (backslash) \ <i>0nnn</i> (the character whose ASCII code is <i>NNN</i> (octal). <i>NNN</i> can be 0 to 3 octal digits) </p> <p>You can turn off the interpretation of the above characters with the <code>-E</code> option.</p>
<code>enable</code>	<p> <code>enable [-pnds] [-a] [-f filename] [name...]</code> </p> <p>Enable and disable builtin shell commands. This allows you to use a disk command which has the same name as a shell builtin without specifying a full pathname. If <code>-n</code> is used, the <i>NAMES</i> become disabled; otherwise <i>NAMES</i> are enabled. For example, to use the <code>test</code> found in <code>\$PATH</code> instead of the shell builtin version, type <code>enable -n test</code>. On systems supporting dynamic loading, the <code>-f</code> option may be used to load new builtins from the shared object <i>FILENAME</i>. The <code>-d</code> option will delete a builtin previously loaded with <code>-f</code>. If no non-option names are given, or the <code>-p</code> option is supplied, a list of builtins is printed. The <code>-a</code> option means to print every builtin with an indication of whether or not it is enabled. The <code>-s</code> option restricts the output to the POSIX.2 'special' builtins. The <code>-n</code> option displays a list of all disabled builtins.</p>
<code>eval</code>	<p> <code>eval [ARG ...]</code> </p> <p>Read <i>ARGS</i> as input to the shell and execute the resulting command(s).</p>
<code>exec</code>	<p> <code>exec [-cl] [-a name] file [redirection ...]</code> </p> <p>Exec <i>FILE</i>, replacing this shell with the specified program. If <i>FILE</i> is not specified, the redirections take effect in this shell. If the first argument is <code>-l</code>, then place a dash in the zeroth arg passed to <i>FILE</i>, as <code>login</code> does. If the <code>-c</code> option is supplied, <i>FILE</i> is executed with a null environment. The <code>-a</code> option means to make <code>set argv[0]</code> of the</p>

Command	Syntax/Description
	executed process to <i>NAME</i> . If the file cannot be executed and the shell is not interactive, then the shell exits, unless the shell option <code>execfail</code> is set.
<code>exit</code>	<code>exit [N]</code> Exit the shell with a status of <i>N</i> . If <i>N</i> is omitted, the exit status is that of the last command executed.
<code>export</code>	<code>export [-nf] [NAME[=value] ...]</code> <code>export -p</code> <i>NAMES</i> are marked for automatic export to the environment of subsequently executed commands. If the <code>-f</code> option is given, the <i>NAMES</i> refer to functions. If no <i>NAMES</i> are given, or if <code>-p</code> is given, a list of all names that are exported in this shell is printed. An argument of <code>-n</code> says to remove the export property from subsequent <i>NAMES</i> . An argument of <code>--</code> disables further option processing.
<code>false</code>	<code>false</code> Return an unsuccessful result.
<code>fc</code>	<code>fc [-e ename] [-nlr] [FIRST] [LAST]</code> <code>fc -s [pat=rep] [cmd]</code> <code>fc</code> is used to list or edit and re-execute commands from the history list. <i>FIRST</i> and <i>LAST</i> can be numbers specifying the range, or <i>FIRST</i> can be a string, which means the most recent command beginning with that string.
<code>fg</code>	<code>fg [JOB_SPEC]</code> Place <i>JOB_SPEC</i> in the foreground, and make it the current job. If <i>JOB_SPEC</i> is not present, the shell's notion of the current job is used.
<code>for</code>	<code>for NAME [in WORDS ... ;] do COMMANDS; done</code>

Command	Syntax/Description
	<p>The <code>for</code> loop executes a sequence of commands for each member in a list of items. If <code>in WORDS ... ;</code> is not present, then <code>in "\$@"</code> is assumed. For each element in <i>WORDS</i>, <i>NAME</i> is set to that element, and the <i>COMMANDS</i> are executed.</p>
<code>function</code>	<pre>function NAME { COMMANDS ; } function NAME () { COMMANDS ; }</pre> <p>Create a simple command invoked by <i>NAME</i> which runs <i>COMMANDS</i>. Arguments on the command line along with <i>NAME</i> are passed to the function as <code>\$0 .. \$n</code>.</p>
<code>getopts</code>	<pre>getopts OPTSTRING NAME [arg]</pre> <p>Getopts is used by shell procedures to parse positional parameters.</p>
<code>hash</code>	<pre>hash [-lr] [-p PATHNAME] [-dt] [NAME...]</pre> <p>For each <i>NAME</i>, the full pathname of the command is determined and remembered. If the <code>-p</code> option is supplied, <i>PATHNAME</i> is used as the full pathname of <i>NAME</i>, and no path search is performed. The <code>-r</code> option causes the shell to forget all remembered locations. The <code>-d</code> option causes the shell to forget the remembered location of each <i>NAME</i>. If the <code>-t</code> option is supplied the full pathname to which each <i>NAME</i> corresponds is printed. If multiple <i>NAME</i> arguments are supplied with <code>-t</code>, the <i>NAME</i> is printed before the hashed full pathname. The <code>-l</code> option causes output to be displayed in a format that may be reused as input. If no arguments are given, information about remembered commands is displayed.</p>
<code>history</code>	<pre>history [-c] [-d OFFSET] [n] history -ps arg [arg...] history -awrm [filename]</pre> <p>Display the history list with line numbers. Lines listed with with a <code>*</code> have been modified. Argument of <i>N</i> says to list only the last <i>N</i> lines. The <code>-c</code> option causes the history list to be cleared by deleting all of</p>

Command	Syntax/Description
	<p>the entries. The <code>-d</code> option deletes the history entry at offset <i>OFFSET</i>. The <code>-w</code> option writes out the current history to the history file; <code>-r</code> means to read the file and append the contents to the history list instead. <code>-a</code> means to append history lines from this session to the history file. Argument <code>-n</code> means to read all history lines not already read from the history file and append them to the history list.</p>
<code>jobs</code>	<pre>jobs [-lnprs] [JOBSPEC ...] job -x COMMAND [ARGS]</pre> <p>Lists the active jobs. The <code>-l</code> option lists process id's in addition to the normal information; the <code>-p</code> option lists process id's only. If <code>-n</code> is given, only processes that have changed status since the last notification are printed. <i>JOBSPEC</i> restricts output to that job. The <code>-r</code> and <code>-s</code> options restrict output to running and stopped jobs only, respectively. Without options, the status of all active jobs is printed. If <code>-x</code> is given, <i>COMMAND</i> is run after all job specifications that appear in <i>ARGS</i> have been replaced with the process ID of that job's process group leader.</p>
<code>kill</code>	<pre>kill [-s sigspec -n signum -sigspec] pid JOBSPEC ... kill -l [sigspec]</pre> <p>Send the processes named by PID (or <i>JOBSPEC</i>) the signal <i>SIGSPEC</i>. If <i>SIGSPEC</i> is not present, then <i>SIGTERM</i> is assumed. An argument of <code>-l</code> lists the signal names; if arguments follow <code>-l</code> they are assumed to be signal numbers for which names should be listed. Kill is a shell builtin for two reasons: it allows job IDs to be used instead of process IDs, and, if you have reached the limit on processes that you can create, you don't have to start a process to kill another one.</p>
<code>let</code>	<pre>let ARG [ARG ...]</pre> <p>Each <i>ARG</i> is a mathematical expression to be evaluated. Evaluation is done in fixed-width integers with no check for overflow, though division by 0 is trapped and flagged as an error. The following list of operators is grouped into levels of equal-precedence operators. The levels are listed in order of decreasing precedence.</p>

Command	Syntax/Description
<code>local</code>	<pre>local NAME[=VALUE] ...</pre> <p>Create a local variable called <i>NAME</i>, and give it <i>VALUE</i>. <code>local</code> can only be used within a function; it makes the variable <i>NAME</i> have a visible scope restricted to that function and its children.</p>
<code>logout</code>	<pre>logout</pre> <p>Logout of a login shell.</p>
<code>popd</code>	<pre>popd [+N -N] [-n]</pre> <p>Removes entries from the directory stack. With no arguments, removes the top directory from the stack, and <code>cd</code>'s to the new top directory.</p>
<code>printf</code>	<pre>printf [-v var] format [ARGUMENTS]</pre> <p><code>printf</code> formats and prints <i>ARGUMENTS</i> under control of the <i>FORMAT</i>. <i>FORMAT</i> is a character string which contains three types of objects: plain characters, which are simply copied to standard output, character escape sequences which are converted and copied to the standard output, and format specifications, each of which causes printing of the next successive argument. In addition to the standard <code>printf(1)</code> formats, <code>%b</code> means to expand backslash escape sequences in the corresponding argument, and <code>%q</code> means to quote the argument in a way that can be reused as shell input. If the <code>-v</code> option is supplied, the output is placed into the value of the shell variable <i>VAR</i> rather than being sent to the standard output.</p>
<code>pushd</code>	<pre>pushd [dir +N -N] [-n]</pre> <p>Adds a directory to the top of the directory stack, or rotates the stack, making the new top of the stack the current working directory. With no arguments, exchanges the top two directories.</p>
<code>pwd</code>	<pre>pwd [-LP]</pre>

Command	Syntax/Description
	<p>Print the current working directory. With the <code>-P</code> option, <code>pwd</code> prints the physical directory, without any symbolic links; the <code>-L</code> option makes <code>pwd</code> follow symbolic links.</p>
<code>read</code>	<p><code>read [-ers] [-u fd] [-t timeout] [-p prompt] [-a array] [-n nchars] [-d delim] [NAME ...]</code></p> <p>The given <i>NAMES</i> are marked readonly and the values of these <i>NAMES</i> may not be changed by subsequent assignment. If the <code>-f</code> option is given, then functions corresponding to the <i>NAMES</i> are so marked. If no arguments are given, or if <code>-p</code> is given, a list of all readonly names is printed. The <code>-a</code> option means to treat each <i>NAME</i> as an array variable. An argument of <code>--</code> disables further option processing.</p>
<code>readonly</code>	<p><code>readonly [-af] [NAME[=VALUE] ...]</code> <code>readonly -p</code></p> <p>The given <i>NAMES</i> are marked readonly and the values of these <i>NAMES</i> may not be changed by subsequent assignment. If the <code>-f</code> option is given, then functions corresponding to the <i>NAMES</i> are so marked. If no arguments are given, or if <code>-p</code> is given, a list of all readonly names is printed. The <code>-a</code> option means to treat each <i>NAME</i> as an array variable. An argument of <code>--</code> disables further option processing.</p>
<code>return</code>	<p><code>return [N]</code></p> <p>Causes a function to exit with the return value specified by <i>N</i>. If <i>N</i> is omitted, the return status is that of the last command.</p>
<code>select</code>	<p><code>select NAME [in WORDS ... ;] do COMMANDS; done</code></p> <p>The <i>WORDS</i> are expanded, generating a list of words. The set of expanded words is printed on the standard error, each preceded by a number. If <code>in WORDS</code> is not present, <code>in "\$@"</code> is assumed. The PS3 prompt is then displayed and a line read from the standard input. If the line consists of the number corresponding to one of the displayed words, then <i>NAME</i> is set to that word. If the line is empty, <i>WORDS</i> and</p>

Command	Syntax/Description
	the prompt are redisplayed. If EOF is read, the command completes. Any other value read causes <i>NAME</i> to be set to null. The line read is saved in the variable <i>REPLY</i> . <i>COMMANDS</i> are executed after each selection until a break command is executed.
set	<pre>set [--abefhkmnptuvxBCHP] [-o OPTION] [ARG...]</pre> <p>Sets internal shell options.</p>
shift	<pre>shift [n]</pre> <p>The positional parameters from $\\$N+1$. . . are renamed to $\\$1$. . . If <i>N</i> is not given, it is assumed to be 1.</p>
shopt	<pre>shopt [-pqsu] [-o long-option] OPTNAME [OPTNAME...]</pre> <p>Toggle the values of variables controlling optional behavior. The <i>-s</i> flag means to enable (set) each <i>OPTNAME</i>; the <i>-u</i> flag unsets each <i>OPTNAME</i>. The <i>-q</i> flag suppresses output; the exit status indicates whether each <i>OPTNAME</i> is set or unset. The <i>-o</i> option restricts the <i>OPTNAME</i>s to those defined for use with <code>set -o</code>. With no options, or with the <i>-p</i> option, a list of all settable options is displayed, with an indication of whether or not each is set.</p>
source	<pre>source FILENAME [ARGS]</pre> <p>Read and execute commands from <i>FILENAME</i> and return. The pathnames in $\\$PATH$ are used to find the directory containing <i>FILENAME</i>. If any <i>ARGS</i> are supplied, they become the positional parameters when <i>FILENAME</i> is executed.</p>
suspend	<pre>suspend [-f]</pre> <p>Suspend the execution of this shell until it receives a SIGCONT signal. The <i>-f</i> if specified says not to complain about this being a login shell if it is; just suspend anyway.</p>

Command	Syntax/Description
<code>test</code>	<pre>test [expr]</pre> <p>Exits with a status of 0 (true) or 1 (false) depending on the evaluation of <i>EXPR</i>. Expressions may be unary or binary. Unary expressions are often used to examine the status of a file. There are string operators as well, and numeric comparison operators.</p>
<code>time</code>	<pre>time [-p] PIPELINE</pre> <p>Execute <i>PIPELINE</i> and print a summary of the real time, user CPU time, and system CPU time spent executing <i>PIPELINE</i> when it terminates. The return status is the return status of <i>PIPELINE</i>. The <code>-p</code> option prints the timing summary in a slightly different format. This uses the value of the <code>TIMEFORMAT</code> variable as the output format.</p>
<code>times</code>	<pre>times</pre> <p>Print the accumulated user and system times for processes run from the shell.</p>
<code>trap</code>	<pre>trap [-lp] [ARG SIGNAL_SPEC ...]</pre> <p>The command <i>ARG</i> is to be read and executed when the shell receives signal(s) <i>SIGNAL_SPEC</i>. If <i>ARG</i> is absent (and a single <i>SIGNAL_SPEC</i> is supplied) or <code>-</code>, each specified signal is reset to its original value. If <i>ARG</i> is the null string each <i>SIGNAL_SPEC</i> is ignored by the shell and by the commands it invokes. If a <i>SIGNAL_SPEC</i> is <code>EXIT</code> (0) the command <i>ARG</i> is executed on exit from the shell. If a <i>SIGNAL_SPEC</i> is <code>DEBUG</code>, <i>ARG</i> is executed after every simple command. If the <code>-p</code> option is supplied then the trap commands associated with each <i>SIGNAL_SPEC</i> are displayed. If no arguments are supplied or if only <code>-p</code> is given, trap prints the list of commands associated with each signal. Each <i>SIGNAL_SPEC</i> is either a signal name in <code>signal.h</code> or a signal number. Signal names are case insensitive and the <code>SIG</code> prefix is optional. <code>trap -l</code> prints a list of signal names and their corresponding numbers. Note that a signal can be sent to the shell with <code>kill -signal \$\$</code>.</p>

Command	Syntax/Description
true	<pre>true</pre> <p>Return a successful result.</p>
type	<pre>type [-afptP] NAME [NAME ...]</pre> <p>Obsolete, see declare.</p>
typeset	<pre>typeset [-afFirtx] [-p] name[=value]</pre> <p>Obsolete, see declare.</p>
ulimit	<pre>ulimit [-SHacdfilmpqstuvx] [limit]</pre> <p>Ulimit provides control over the resources available to processes started by the shell, on systems that allow such control.</p>
umask	<pre>umask [-p] [-S] [MODE]</pre> <p>The user file-creation mask is set to <i>MODE</i>. If <i>MODE</i> is omitted, or if <i>-S</i> is supplied, the current value of the mask is printed. The <i>-S</i> option makes the output symbolic; otherwise an octal number is output. If <i>-p</i> is supplied, and <i>MODE</i> is omitted, the output is in a form that may be used as input. If <i>MODE</i> begins with a digit, it is interpreted as an octal number, otherwise it is a symbolic mode string like that accepted by <code>chmod(1)</code>.</p>
unalias	<pre>unalias [-a] NAME [NAME ...]</pre> <p>Remove <i>NAMES</i> from the list of defined aliases. If the <i>-a</i> option is given, then remove all alias definitions.</p>
unset	<pre>unset [-f] [-v] [NAME ...]</pre> <p>For each <i>NAME</i>, remove the corresponding variable or function. Given the <i>-v</i>, unset will only act on variables. Given the <i>-f</i> flag, unset will only act on functions. With neither flag, unset first tries to unset a variable. If that fails, it then tries to unset a function. Some variables cannot be unset; also see <code>readonly</code>.</p>

Command	Syntax/Description
<code>until</code>	<pre>until COMMANDS; do COMMANDS; done</pre> <p>Expand and execute <i>COMMANDS</i> as long as the final command in the <code>until</code> <i>COMMANDS</i> has an exit status which is not zero.</p>
<code>wait</code>	<pre>wait [N]</pre> <p>Wait for the specified process and report its termination status. If <i>N</i> is not given, all currently active child processes are waited for, and the return code is zero. <i>N</i> may be a process ID or a job specification; if a job spec is given, all processes in the job's pipeline are waited for.</p>
<code>while</code>	<pre>while COMMANDS; do COMMANDS; done</pre> <p>Expand and execute <i>COMMANDS</i> as long as the final command in the <code>while</code> <i>COMMANDS</i> has an exit status of zero.</p>

crm_standby (8)

`crm_standby` — manipulate a node's standby attribute to determine whether resources can be run on this node

Synopsis

```
crm_standby [-?|-V] -D -u|-U node -r resource
crm_standby [-?|-V] -G -u|-U node -r resource
crm_standby [-?|-V] -v string -u|-U node -r resource [-l string]
```

Description

The `crm_standby` command manipulates 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 when it needs to become a fully active member of the cluster again.

By assigning a lifetime to the `standby` attribute, determine whether the standby setting should survive a reboot of the node (set lifetime to `forever`) or should be reset with reboot (set lifetime to `reboot`). Alternatively, remove the `standby` attribute and bring the node back from standby manually.

Options

`--help, -?`

Print a help message.

`--verbose, -V`

Turn on debug information.

注意

Increase the level of verbosity by providing additional instances.

`--quiet, -Q`

When doing an attribute query using `-G`, print just the value to stdout. Use this option with `-G`.

`--get-value, -G`

Retrieve rather than set the preference.

`--delete-attr, -D`

Specify the attribute to delete.

`--attr-value string, -v string`

Specify the value to use. This option is ignored when used with `-G`.

`--attr-id string, -i string`

For advanced users only. Identifies the id attribute..

`--node node_undef, -u node_undef`

Specify the uname of the node to change.

`--lifetime string, -l string`

Determine how long this preference lasts. Possible values are `reboot` or `forever`.

注意

If a `forever` value exists, it is always used by the CRM instead of any `reboot` value.

Examples

Have a local node go to standby:

```
crm_standby -v true
```

Have a node (`node1`) go to standby:

```
crm_standby -v true -U node1
```

Query the standby status of a node:

```
crm_standby -G -U node1
```

Remove the standby property from a node:

```
crm_standby -D -U node1
```

Have a node go to standby for an indefinite period of time:

```
crm_standby -v true -l forever -U node1
```

Have a node go to standby until the next reboot of this node:

```
crm_standby -v true -l reboot -U node1
```

Files

`/var/lib/heartbeat/crm/cib.xml`—the CIB (minus status section) on disk.
Editing this file directly is strongly discouraged.

See Also

`cibadmin(8)` (第 193 页), `crm_attribute(8)` (第 202 页)

crm_verify (8)

crm_verify — check the CIB for consistency

Synopsis

```
crm_verify [-V] -x file
crm_verify [-V] -X string
crm_verify [-V] -L|-p
crm_verify [-?]
```

Description

crm_verify checks the configuration database (CIB) for consistency and other problems. It can be used to check a file containing the configuration or can it can connect to a running cluster. It reports two classes of problems, errors and warnings. Errors must be fixed before High Availability can work properly. However, it is left up to the administrator to decide if the warnings should also be fixed.

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.

Options

--help, -h
Print a help message.

--verbose, -V
Turn on debug information.

注意

Increase the level of verbosity by providing additional instances.

`--live-check, -L`

Connect to the running cluster and check the CIB.

`--crm_xml string, -X string`

Check the configuration in the supplied string. Pass complete CIBs only.

`--xml-file file, -x file`

Check the configuration in the named file.

`--xml-pipe, -p`

Use the configuration piped in via stdin. Pass complete CIBs only.

Examples

Check the consistency of the configuration in the running cluster and produce verbose output:

```
crm_verify -VL
```

Check the consistency of the configuration in a given file and produce verbose output:

```
crm_verify -Vx file1
```

Pipe a configuration into `crm_verify` and produce verbose output:

```
cat file1.xml | crm_verify -Vp
```

Files

`/var/lib/heartbeat/crm/cib.xml`—the CIB (minus status section) on disk. Editing this file directly is strongly discouraged.

See Also

`cibadmin(8)` (第 193 页)

HA OCF Agents

All OCF agents require several parameters to be set when they are started. The following overview shows how to manually operate these agents. The data that is available in this appendix is directly taken from the `meta-data` invocation of the respective RA. Find all these agents in `/usr/lib/ocf/resource.d/heartbeat/`.

When configuring an RA, omit the `OCF_RESKEY_` prefix to the parameter name. Parameters that are in square brackets may be omitted in the configuration.

ocf:anything (7)

ocf:anything — Manages an arbitrary service

Synopsis

```
OCF_RESKEY_binfile=string [OCF_RESKEY_cmdline_options=string]
[OCF_RESKEY_pidfile=string] [OCF_RESKEY_logfile=string]
[OCF_RESKEY_errlogfile=string] [OCF_RESKEY_user=string]
[OCF_RESKEY_monitor_hook=string] [OCF_RESKEY_stop_timeout=string]
anything [start | stop | monitor | meta-data | validate-all]
```

Description

This is a generic OCF RA to manage almost anything.

Supported Parameters

OCF_RESKEY_binfile=Full path name of the binary to be executed
The full name of the binary to be executed. This is expected to keep running with the same pid and not just do something and exit.

OCF_RESKEY_cmdline_options=Command line options
Command line options to pass to the binary

OCF_RESKEY_pidfile=File to write STDOUT to
File to read/write the PID from/to.

OCF_RESKEY_logfile=File to write STDOUT to
File to write STDOUT to

OCF_RESKEY_errlogfile=File to write STDERR to
File to write STDERR to

OCF_RESKEY_user=User to run the command as
User to run the command as

OCF_RESKEY_monitor_hook=Command to run in monitor operation
Command to run in monitor operation

OCF_RESKEY_stop_timeout=Seconds to wait after having sent SIGTERM before
sending SIGKILL in stop operation

In the stop operation: Seconds to wait for kill -TERM to succeed before sending
kill -SIGKILL. Defaults to 2/3 of the stop operation timeout.

ocf:AoEtarget (7)

ocf:AoEtarget — Manages ATA-over-Ethernet (AoE) target exports

Synopsis

```
[OCF_RESKEY_device=string] [OCF_RESKEY_nic=string]  
[OCF_RESKEY_shelf=integer] [OCF_RESKEY_slot=integer]  
OCF_RESKEY_pid=string [OCF_RESKEY_binary=string] AoEtarget [start |  
stop | monitor | reload | meta-data | validate-all]
```

Description

This resource agent manages an ATA-over-Ethernet (AoE) target using vblade. It exports any block device, or file, as an AoE target using the specified Ethernet device, shelf, and slot number.

Supported Parameters

OCF_RESKEY_device=Device to export

The local block device (or file) to export as an AoE target.

OCF_RESKEY_nic=Ethernet interface

The local Ethernet interface to use for exporting this AoE target.

OCF_RESKEY_shelf=AoE shelf number

The AoE shelf number to use when exporting this target.

OCF_RESKEY_slot=AoE slot number

The AoE slot number to use when exporting this target.

OCF_RESKEY_pid=Daemon pid file

The file to record the daemon pid to.

OCF_RESKEY_binary=vblade binary

Location of the vblade binary.

ocf:apache (7)

ocf:apache — Manages an Apache web server instance

Synopsis

```
OCF_RESKEY_configfile=string [OCF_RESKEY_httpd=string]
[OCF_RESKEY_port=integer] [OCF_RESKEY_statusurl=string]
[OCF_RESKEY_testregex=string] [OCF_RESKEY_client=string]
[OCF_RESKEY_testurl=string] [OCF_RESKEY_testregex10=string]
[OCF_RESKEY_testconf=string] [OCF_RESKEY_testname=string]
[OCF_RESKEY_options=string] [OCF_RESKEY_envfiles=string] apache
[start | stop | status | monitor | meta-data | validate-all]
```

Description

This is the resource agent for the Apache web server. This resource agent operates both version 1.x and version 2.x Apache servers. The start operation ends with a loop in which monitor is repeatedly called to make sure that the server started and that it is operational. Hence, if the monitor operation does not succeed within the start operation timeout, the apache resource will end with an error status. The monitor operation by default loads the server status page which depends on the mod_status module and the corresponding configuration file (usually /etc/apache2/mod_status.conf). Make sure that the server status page works and that the access is allowed **only** from localhost (address 127.0.0.1). See the statusurl and testregex attributes for more details. See also <http://httpd.apache.org/>

Supported Parameters

OCF_RESKEY_configfile=configuration file path

The full pathname of the Apache configuration file. This file is parsed to provide defaults for various other resource agent parameters.

OCF_RESKEY_httpd=httpd binary path

The full pathname of the httpd binary (optional).

`OCF_RESKEY_port=httpd port`

A port number that we can probe for status information using the `statusurl`. This will default to the port number found in the configuration file, or 80, if none can be found in the configuration file.

`OCF_RESKEY_statusurl=url name`

The URL to monitor (the apache server status page by default). If left unspecified, it will be inferred from the apache configuration file. If you set this, make sure that it succeeds **only** from the localhost (127.0.0.1). Otherwise, it may happen that the cluster complains about the resource being active on multiple nodes.

`OCF_RESKEY_testregex=monitor regular expression`

Regular expression to match in the output of `statusurl`. Case insensitive.

`OCF_RESKEY_client=http client`

Client to use to query to Apache. If not specified, the RA will try to find one on the system. Currently, `wget` and `curl` are supported. For example, you can set this parameter to "curl" if you prefer that to `wget`.

`OCF_RESKEY_testurl=test url`

URL to test. If it does not start with "http", then it's considered to be relative to the Listen address.

`OCF_RESKEY_testregex10=extended monitor regular expression`

Regular expression to match in the output of `testurl`. Case insensitive.

`OCF_RESKEY_testconf file=test configuration file`

A file which contains test configuration. Could be useful if you have to check more than one web application or in case sensitive info should be passed as arguments (passwords). Furthermore, using a config file is the only way to specify certain parameters. Please see `README.webapps` for examples and file description.

`OCF_RESKEY_testname=test name`

Name of the test within the test configuration file.

`OCF_RESKEY_options=command line options`

Extra options to apply when starting apache. See `man httpd(8)`.

OCF_RESKEY_envfiles=environment settings files

Files (one or more) which contain extra environment variables. If you want to prevent script from reading the default file, set this parameter to empty string.

ocf:AudibleAlarm (7)

ocf:AudibleAlarm — Emits audible beeps at a configurable interval

Synopsis

```
[OCF_RESKEY_nodelist=string] AudibleAlarm [start | stop | restart | status |  
monitor | meta-data | validate-all]
```

Description

Resource script for AudibleAlarm. It sets an audible alarm running by beeping at a set interval.

Supported Parameters

OCF_RESKEY_nodelist=Node list

The node list that should never sound the alarm.

ocf:ClusterMon (7)

ocf:ClusterMon — Runs `crm_mon` in the background, recording the cluster status to an HTML file

Synopsis

```
[OCF_RESKEY_user=string] [OCF_RESKEY_update=integer]  
[OCF_RESKEY_extra_options=string] OCF_RESKEY_pidfile=string  
OCF_RESKEY_htmlfile=string ClusterMon [start | stop | monitor | meta-data |  
validate-all]
```

Description

This is a ClusterMon Resource Agent. It outputs current cluster status to the html.

Supported Parameters

OCF_RESKEY_user=The user we want to run `crm_mon` as
The user we want to run `crm_mon` as

OCF_RESKEY_update=Update interval
How frequently should we update the cluster status

OCF_RESKEY_extra_options=Extra options
Additional options to pass to `crm_mon`. Eg. `-n -r`

OCF_RESKEY_pidfile=PID file
PID file location to ensure only one instance is running

OCF_RESKEY_htmlfile=HTML output
Location to write HTML output to.

ocf:CTDB (7)

ocf:CTDB — CTDB Resource Agent

Synopsis

```
OCF_RESKEY_ctdb_recovery_lock=string
OCF_RESKEY_smb_private_dir=string
[OCF_RESKEY_ctdb_config_dir=string]
[OCF_RESKEY_ctdb_binary=string] [OCF_RESKEY_ctdbd_binary=string]
[OCF_RESKEY_ctdb_socket=string] [OCF_RESKEY_ctdb_dbdir=string]
[OCF_RESKEY_ctdb_logfile=string]
[OCF_RESKEY_ctdb_debuglevel=integer] [OCF_RESKEY_smb_conf=string]
CTDB [start | stop | monitor | meta-data | validate-all]
```

Description

This resource agent manages CTDB, allowing one to use Clustered Samba in a Linux-HA/Pacemaker cluster. You need a shared filesystem (e.g. OCFS2) on which CTDB lock and Samba state will be stored. Configure shares in `smb.conf` on all nodes, and create `/etc/ctdb/nodes` containing a list of private IP addresses of each node in the cluster. Configure this RA as a clone, and it will take care of the rest. For more information see [http://linux-ha.org/wiki/CTDB_\(resource_agent\)](http://linux-ha.org/wiki/CTDB_(resource_agent))

Supported Parameters

OCF_RESKEY_ctdb_recovery_lock=CTDB shared lock file

The location of a shared lock file, common across all nodes. This must be on shared storage, e.g.: `/shared-fs/samba/ctdb.lock`

OCF_RESKEY_smb_private_dir=Samba private dir

The directory for `smbd` to use for storing such files as `smbpasswd` and `secrets.tdb`. This must be on shared storage, e.g.: `/shared-fs/samba/private`

`OCF_RESKEY_ctdb_config_dir`=CTDB config file directory

The directory containing various CTDB configuration files. The "nodes" and "notify.sh" scripts are expected to be in this directory, as is the "events.d" subdirectory.

`OCF_RESKEY_ctdb_binary`=CTDB binary path

Full path to the CTDB binary.

`OCF_RESKEY_ctdbd_binary`=CTDB Daemon binary path

Full path to the CTDB cluster daemon binary.

`OCF_RESKEY_ctdb_socket`=CTDB socket location

Full path to the domain socket that ctdbd will create, used for local clients to attach and communicate with the ctdb daemon.

`OCF_RESKEY_ctdb_dbdir`=CTDB database directory

The directory to put the local CTDB database files in. Persistent database files will be put in ctdb_dbdir/persistent.

`OCF_RESKEY_ctdb_logfile`=CTDB log file location

Full path to log file. To log to syslog instead, use the value "syslog".

`OCF_RESKEY_ctdb_debuglevel`=CTDB debug level

What debug level to run at (0-10). Higher means more verbose.

`OCF_RESKEY_smb_conf`=Path to smb.conf

Path to default samba config file.

ocf:db2 (7)

ocf:db2 — Manages an IBM DB2 Universal Database instance

Synopsis

```
[OCF_RESKEY_instance=string] [OCF_RESKEY_admin=string] db2 [start | stop  
| status | monitor | validate-all | meta-data | methods]
```

Description

Resource script for db2. It manages a DB2 Universal Database instance as an HA resource.

Supported Parameters

OCF_RESKEY_instance=instance
The instance of database.

OCF_RESKEY_admin=admin
The admin user of the instance.

ocf:Delay (7)

ocf:Delay — Waits for a defined timespan

Synopsis

```
[OCF_RESKEY_startdelay=integer] [OCF_RESKEY_stopdelay=integer]  
[OCF_RESKEY_mondelay=integer] Delay [start | stop | status | monitor | meta-data  
| validate-all]
```

Description

This script is a test resource for introducing delay.

Supported Parameters

OCF_RESKEY_startdelay=Start delay
How long in seconds to delay on start operation.

OCF_RESKEY_stopdelay=Stop delay
How long in seconds to delay on stop operation. Defaults to "startdelay" if unspecified.

OCF_RESKEY_mondelay=Monitor delay
How long in seconds to delay on monitor operation. Defaults to "startdelay" if unspecified.

ocf:drbd (7)

ocf:drbd — Manages a DRBD resource (deprecated)

Synopsis

```
OCF_RESKEY_drbd_resource=string [OCF_RESKEY_drbdconf=string]
[OCF_RESKEY_clone_overrides_hostname=boolean]
[OCF_RESKEY_ignore_deprecation=boolean] drbd [start | promote | demote
| notify | stop | monitor | monitor | meta-data | validate-all]
```

Description

Deprecation warning: This agent is deprecated and may be removed from a future release. See the ocf:linbit:drbd resource agent for a supported alternative. -- This resource agent manages a Distributed Replicated Block Device (DRBD) object as a master/slave resource. DRBD is a mechanism for replicating storage; please see the documentation for setup details.

Supported Parameters

OCF_RESKEY_drbd_resource=drbd resource name
The name of the drbd resource from the drbd.conf file.

OCF_RESKEY_drbdconf=Path to drbd.conf
Full path to the drbd.conf file.

OCF_RESKEY_clone_overrides_hostname=Override drbd hostname
Whether or not to override the hostname with the clone number. This can be used to create floating peer configurations; drbd will be told to use node_<cloneno> as the hostname instead of the real uname, which can then be used in drbd.conf.

OCF_RESKEY_ignore_deprecation=Suppress deprecation warning
If set to true, suppresses the deprecation warning for this agent.

ocf:Dummy (7)

ocf:Dummy — Example stateless resource agent

Synopsis

`OCF_RESKEY_state=string Dummy [start | stop | monitor | reload | migrate_to | migrate_from | meta-data | validate-all]`

Description

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.

Supported Parameters

`OCF_RESKEY_state=State file`
Location to store the resource state in.

ocf:eDir88 (7)

ocf:eDir88 — Manages a Novell eDirectory directory server

Synopsis

```
OCF_RESKEY_eDir_config_file=string  
[OCF_RESKEY_eDir_monitor_ldap=boolean]  
[OCF_RESKEY_eDir_monitor_idm=boolean]  
[OCF_RESKEY_eDir_jvm_initial_heap=integer]  
[OCF_RESKEY_eDir_jvm_max_heap=integer]  
[OCF_RESKEY_eDir_jvm_options=string] eDir88 [start | stop | monitor | meta-  
data | validate-all]
```

Description

Resource script for managing an eDirectory instance. Manages a single instance of eDirectory as an HA resource. The "multiple instances" feature of eDirectory has been added in version 8.8. This script will not work for any version of eDirectory prior to 8.8. This RA can be used to load multiple eDirectory instances on the same host. It is very strongly recommended to put eDir configuration files (as per the eDir_config_file parameter) on local storage on each node. This is necessary for this RA to be able to handle situations where the shared storage has become unavailable. If the eDir configuration file is not available, this RA will fail, and heartbeat will be unable to manage the resource. Side effects include STONITH actions, unmanageable resources, etc... Setting a high action timeout value is very strongly recommended. eDir with IDM can take in excess of 10 minutes to start. If heartbeat times out before eDir has had a chance to start properly, mayhem WILL ENSUE. The LDAP module seems to be one of the very last to start. So this script will take even longer to start on installations with IDM and LDAP if the monitoring of IDM and/or LDAP is enabled, as the start command will wait for IDM and LDAP to be available.

Supported Parameters

OCF_RESKEY_eDir_config_file=eDir config file

Path to configuration file for eDirectory instance.

OCF_RESKEY_eDir_monitor_ldap=eDir monitor ldap

Should we monitor if LDAP is running for the eDirectory instance?

OCF_RESKEY_eDir_monitor_idm=eDir monitor IDM

Should we monitor if IDM is running for the eDirectory instance?

OCF_RESKEY_eDir_jvm_initial_heap=DHOST_INITIAL_HEAP value

Value for the DHOST_INITIAL_HEAP java environment variable. If unset, java defaults will be used.

OCF_RESKEY_eDir_jvm_max_heap=DHOST_MAX_HEAP value

Value for the DHOST_MAX_HEAP java environment variable. If unset, java defaults will be used.

OCF_RESKEY_eDir_jvm_options=DHOST_OPTIONS value

Value for the DHOST_OPTIONS java environment variable. If unset, original values will be used.

ocf:Evmsd (7)

ocf:Evmsd — Controls clustered EVMS volume management (deprecated)

Synopsis

[OCF_RESKEY_ignore_deprecation=boolean] Evmsd [start | stop | monitor | meta-data]

Description

Deprecation warning: EVMS is no longer actively maintained and should not be used. This agent is deprecated and may be removed from a future release. -- This is a Evmsd Resource Agent.

Supported Parameters

OCF_RESKEY_ignore_deprecation=Suppress deprecation warning
If set to true, suppresses the deprecation warning for this agent.

ocf:EvmsSCC (7)

ocf:EvmsSCC — Manages EVMS Shared Cluster Containers (SCCs) (deprecated)

Synopsis

```
[OCF_RESKEY_ignore_deprecation=boolean] EvmsSCC [start | stop | notify  
| status | monitor | meta-data]
```

Description

Deprecation warning: EVMS is no longer actively maintained and should not be used. This agent is deprecated and may be removed from a future release. -- Resource script for EVMS shared cluster container. It runs `evms_activate` on one node in the cluster.

Supported Parameters

`OCF_RESKEY_ignore_deprecation=Suppress deprecation warning`
If set to true, suppresses the deprecation warning for this agent.

ocf:Filesystem (7)

ocf:Filesystem — Manages filesystem mounts

Synopsis

```
[OCF_RESKEY_device=string] [OCF_RESKEY_directory=string]  
[OCF_RESKEY_fstype=string] [OCF_RESKEY_options=string]  
[OCF_RESKEY_statusfile_prefix=string] Filesystem [start | stop | notify  
| monitor | validate-all | meta-data]
```

Description

Resource script for Filesystem. It manages a Filesystem on a shared storage medium. The standard monitor operation of depth 0 (also known as probe) checks if the filesystem is mounted. If you want deeper tests, set `OCF_CHECK_LEVEL` to one of the following values: 10: read first 16 blocks of the device (raw read) This doesn't exercise the filesystem at all, but the device on which the filesystem lives. This is noop for non-block devices such as NFS, SMBFS, or bind mounts. 20: test if a status file can be written and read The status file must be writable by root. This is not always the case with an NFS mount, as NFS exports usually have the "root_squash" option set. In such a setup, you must either use read-only monitoring (depth=10), export with "no_root_squash" on your NFS server, or grant world write permissions on the directory where the status file is to be placed.

Supported Parameters

`OCF_RESKEY_device=block device`

The name of block device for the filesystem, or -U, -L options for mount, or NFS mount specification.

`OCF_RESKEY_directory=mount point`

The mount point for the filesystem.

OCF_RESKEY_fstype=filesystem type

The optional type of filesystem to be mounted.

OCF_RESKEY_options=options

Any extra options to be given as -o options to mount. For bind mounts, add "bind" here and set fstype to "none". We will do the right thing for options such as "bind,ro".

OCF_RESKEY_statusfile_prefix=status file prefix

The prefix to be used for a status file for resource monitoring with depth 20. If you don't specify this parameter, all status files will be created in a separate directory.

ocf:ICP (7)

ocf:ICP — Manages an ICP Vortex clustered host drive

Synopsis

```
[OCF_RESKEY_driveid=string] [OCF_RESKEY_device=string] ICP [start | stop  
| status | monitor | validate-all | meta-data]
```

Description

Resource script for ICP. It Manages an ICP Vortex clustered host drive as an HA resource.

Supported Parameters

OCF_RESKEY_driveid=ICP cluster drive ID
The ICP cluster drive ID.

OCF_RESKEY_device=device
The device name.

ocf:ids (7)

ocf:ids — Manages an Informix Dynamic Server (IDS) instance

Synopsis

```
[OCF_RESKEY_informixdir=string] [OCF_RESKEY_informixserver=string]  
[OCF_RESKEY_onconfig=string] [OCF_RESKEY_dbname=string]  
[OCF_RESKEY_sqltestquery=string] ids [start | stop | status | monitor | validate-  
all | meta-data | methods | usage]
```

Description

OCF resource agent to manage an IBM Informix Dynamic Server (IDS) instance as an High-Availability resource.

Supported Parameters

OCF_RESKEY_informixdir= INFORMIXDIR environment variable

The value the environment variable INFORMIXDIR has after a typical installation of IDS. Or in other words: the path (without trailing '/') where IDS was installed to. If this parameter is unspecified the script will try to get the value from the shell environment.

OCF_RESKEY_informixserver= INFORMIXSERVER environment variable

The value the environment variable INFORMIXSERVER has after a typical installation of IDS. Or in other words: the name of the IDS server instance to manage. If this parameter is unspecified the script will try to get the value from the shell environment.

OCF_RESKEY_onconfig= ONCONFIG environment variable

The value the environment variable ONCONFIG has after a typical installation of IDS. Or in other words: the name of the configuration file for the IDS instance specified in INFORMIXSERVER. The specified configuration file will be searched

at '/etc/'. If this parameter is unspecified the script will try to get the value from the shell environment.

OCF_RESKEY_dbname= database to use for monitoring, defaults to 'sysmaster'
This parameter defines which database to use in order to monitor the IDS instance. If this parameter is unspecified the script will use the 'sysmaster' database as a default.

OCF_RESKEY_sqltestquery= SQL test query to use for monitoring, defaults to 'SELECT COUNT(*) FROM systables;'
SQL test query to run on the database specified by the parameter 'dbname' in order to monitor the IDS instance and determine if it's functional or not. If this parameter is unspecified the script will use 'SELECT COUNT(*) FROM systables;' as a default.

ocf:IPAddr2 (7)

ocf:IPAddr2 — Manages virtual IPv4 addresses (Linux specific version)

Synopsis

```
OCF_RESKEY_ip=string [OCF_RESKEY_nic=string]
[OCF_RESKEY_cidr_netmask=string] [OCF_RESKEY_broadcast=string]
[OCF_RESKEY_iflabel=string] [OCF_RESKEY_lvs_support=boolean]
[OCF_RESKEY_mac=string] [OCF_RESKEY_clusterip_hash=string]
[OCF_RESKEY_unique_clone_address=boolean]
[OCF_RESKEY_arp_interval=integer] [OCF_RESKEY_arp_count=integer]
[OCF_RESKEY_arp_bg=string] [OCF_RESKEY_arp_mac=string] IPAddr2 [start
| stop | status | monitor | meta-data | validate-all]
```

Description

This Linux-specific resource manages IP alias IP addresses. It can add an IP alias, or remove one. In addition, it can implement Cluster Alias IP functionality if invoked as a clone resource.

Supported Parameters

OCF_RESKEY_ip=IPv4 address

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

OCF_RESKEY_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.

OCF_RESKEY_cidr_netmask=CIDR netmask

The netmask for the interface in CIDR format (e.g., 24 and not 255.255.255.0) If unspecified, the script will also try to determine this from the routing table.

OCF_RESKEY_broadcast=Broadcast address

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

OCF_RESKEY_iflabel=Interface label

You can specify an additional label for your IP address here. This label is appended to your interface name. If a label is specified in nic name, this parameter has no effect.

OCF_RESKEY_lvs_support=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.

OCF_RESKEY_mac=Cluster IP MAC address

Set the interface MAC address explicitly. Currently only used in case of the Cluster IP Alias. Leave empty to chose automatically.

OCF_RESKEY_clusterip_hash=Cluster IP hashing function

Specify the hashing algorithm used for the Cluster IP functionality.

OCF_RESKEY_unique_clone_address=Create a unique address for cloned instances

If true, add the clone ID to the supplied value of ip to create a unique address to manage

OCF_RESKEY_arp_interval=ARP packet interval in ms

Specify the interval between unsolicited ARP packets in milliseconds.

OCF_RESKEY_arp_count=ARP packet count

Number of unsolicited ARP packets to send.

OCF_RESKEY_arp_bg=ARP from background

Whether or not to send the arp packets in the background.

OCF_RESKEY_arp_mac=ARP MAC

MAC address to send the ARP packets too. You really shouldn't be touching this.

ocf:IPaddr (7)

ocf:IPaddr — Manages virtual IPv4 addresses (portable version)

Synopsis

```
OCF_RESKEY_ip=string [OCF_RESKEY_nic=string]
[OCF_RESKEY_cidr_netmask=string] [OCF_RESKEY_broadcast=string]
[OCF_RESKEY_iflabel=string] [OCF_RESKEY_lvs_support=boolean]
[OCF_RESKEY_local_stop_script=string]
[OCF_RESKEY_local_start_script=string]
[OCF_RESKEY_ARP_INTERVAL_MS=integer]
[OCF_RESKEY_ARP_REPEAT=integer]
[OCF_RESKEY_ARP_BACKGROUND=boolean]
[OCF_RESKEY_ARP_NETMASK=string] IPaddr [start | stop | monitor | validate-all
| meta-data]
```

Description

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

Supported Parameters

OCF_RESKEY_ip=IPv4 address

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

OCF_RESKEY_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.

OCF_RESKEY_cidr_netmask=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.

OCF_RESKEY_broadcast=Broadcast address

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

OCF_RESKEY_iflabel=Interface label

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

OCF_RESKEY_lvs_support=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.

OCF_RESKEY_local_stop_script=Script called when the IP is released

Script called when the IP is released

OCF_RESKEY_local_start_script=Script called when the IP is added

Script called when the IP is added

OCF_RESKEY_ARP_INTERVAL_MS=milliseconds between gratuitous ARPs

milliseconds between ARPs

OCF_RESKEY_ARP_REPEAT=repeat count

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

OCF_RESKEY_ARP_BACKGROUND=run in background

run in background (no longer any reason to do this)

OCF_RESKEY_ARP_NETMASK=netmask for ARP

netmask for ARP - in nonstandard hexadecimal format.

ocf:IPsrcaddr (7)

ocf:IPsrcaddr — Manages the preferred source address for outgoing IP packets

Synopsis

```
[OCF_RESKEY_ipaddress=string] IPsrcaddr [start | stop | stop | monitor |  
validate-all | meta-data]
```

Description

Resource script for IPsrcaddr. It manages the preferred source address modification.

Supported Parameters

OCF_RESKEY_ipaddress=IP address
The IP address.

ocf:IPv6addr (7)

ocf:IPv6addr — Manages IPv6 aliases

Synopsis

```
[OCF_RESKEY_ipv6addr=string] [OCF_RESKEY_cidr_netmask=string]  
[OCF_RESKEY_nic=string] IPv6addr [start | stop | status | monitor | validate-all |  
meta-data]
```

Description

This script manages IPv6 alias IPv6 addresses, It can add an IP6 alias, or remove one.

Supported Parameters

OCF_RESKEY_ipv6addr=IPv6 address
The IPv6 address this RA will manage

OCF_RESKEY_cidr_netmask=Netmask
The netmask for the interface in CIDR format. (ie, 24). The value of this parameter overwrites the value of `_prefix_` of `ipv6addr` parameter.

OCF_RESKEY_nic=Network interface
The base network interface on which the IPv6 address will be brought online.

ocf:iSCSILogicalUnit (7)

ocf:iSCSILogicalUnit — Manages iSCSI Logical Units (LUs)

Synopsis

```
[OCF_RESKEY_implementation=string] [OCF_RESKEY_target_iqn=string]
[OCF_RESKEY_lun=integer] [OCF_RESKEY_path=string]
OCF_RESKEY_scsi_id=string OCF_RESKEY_scsi_sn=string
[OCF_RESKEY_vendor_id=string] [OCF_RESKEY_product_id=string]
[OCF_RESKEY_additional_parameters=string] iSCSILogicalUnit [start
| stop | monitor | meta-data | validate-all]
```

Description

Manages iSCSI Logical Unit. An iSCSI Logical unit is a subdivision of an SCSI Target, exported via a daemon that speaks the iSCSI protocol.

Supported Parameters

OCF_RESKEY_implementation=iSCSI target daemon implementation

The iSCSI target daemon implementation. Must be one of "iet", "tgt", or "lio". If unspecified, an implementation is selected based on the availability of management utilities, with "iet" being tried first, then "tgt", then "lio".

OCF_RESKEY_target_iqn=iSCSI target IQN

The iSCSI Qualified Name (IQN) that this Logical Unit belongs to.

OCF_RESKEY_lun=Logical Unit number (LUN)

The Logical Unit number (LUN) exposed to initiators.

OCF_RESKEY_path=Block device (or file) path

The path to the block device exposed. Some implementations allow this to be a regular file, too.

OCF_RESKEY_scsi_id=SCSI ID

The SCSI ID to be configured for this Logical Unit. The default is the resource name, truncated to 24 bytes.

OCF_RESKEY_scsi_sn=SCSI serial number

The SCSI serial number to be configured for this Logical Unit. The default is a hash of the resource name, truncated to 8 bytes.

OCF_RESKEY_vendor_id=SCSI vendor ID

The SCSI vendor ID to be configured for this Logical Unit.

OCF_RESKEY_product_id=SCSI product ID

The SCSI product ID to be configured for this Logical Unit.

OCF_RESKEY_additional_parameters=List of iSCSI LU parameters

Additional LU parameters. A space-separated list of "name=value" pairs which will be passed through to the iSCSI daemon's management interface. The supported parameters are implementation dependent. Neither the name nor the value may contain whitespace.

ocf:iSCSITarget (7)

ocf:iSCSITarget — iSCSI target export agent

Synopsis

```
[OCF_RESKEY_implementation=string] OCF_RESKEY_iqn=string  
OCF_RESKEY_tid=integer [OCF_RESKEY_portals=string]  
[OCF_RESKEY_allowed_initiators=string]  
OCF_RESKEY_incoming_username=string  
[OCF_RESKEY_incoming_password=string]  
[OCF_RESKEY_additional_parameters=string] iSCSITarget [start | stop  
| monitor | meta-data | validate-all]
```

Description

Manages iSCSI targets. An iSCSI target is a collection of SCSI Logical Units (LUs) exported via a daemon that speaks the iSCSI protocol.

Supported Parameters

`OCF_RESKEY_implementation=`Manages an iSCSI target export

The iSCSI target daemon implementation. Must be one of "iet", "tgt", or "lio". If unspecified, an implementation is selected based on the availability of management utilities, with "iet" being tried first, then "tgt", then "lio".

`OCF_RESKEY_iqn=`iSCSI target IQN

The target iSCSI Qualified Name (IQN). Should follow the conventional "iqn.yyyy-mm.<reversed domain name>[:identifier]" syntax.

`OCF_RESKEY_tid=`iSCSI target ID

The iSCSI target ID. Required for tgt.

OCF_RESKEY_portals=iSCSI portal addresses

iSCSI network portal addresses. Not supported by all implementations. If unset, the default is to create one portal that listens on .

OCF_RESKEY_allowed_initiators=List of iSCSI initiators allowed to connect to this target

Allowed initiators. A space-separated list of initiators allowed to connect to this target. Initiators may be listed in any syntax the target implementation allows. If this parameter is empty or not set, access to this target will be allowed from any initiator.

OCF_RESKEY_incoming_username=Incoming account username

A username used for incoming initiator authentication. If unspecified, allowed initiators will be able to log in without authentication.

OCF_RESKEY_incoming_password=Incoming account password

A password used for incoming initiator authentication.

OCF_RESKEY_additional_parameters=List of iSCSI target parameters

Additional target parameters. A space-separated list of "name=value" pairs which will be passed through to the iSCSI daemon's management interface. The supported parameters are implementation dependent. Neither the name nor the value may contain whitespace.

ocf:iscsi (7)

ocf:iscsi — Manages a local iSCSI initiator and its connections to iSCSI targets

Synopsis

```
[OCF_RESKEY_portal=string] OCF_RESKEY_target=string  
[OCF_RESKEY_discovery_type=string] [OCF_RESKEY_iscsiadm=string]  
[OCF_RESKEY_udev=string] iscsi [start | stop | status | monitor | validate-all |  
methods | meta-data]
```

Description

OCF Resource Agent for iSCSI. Add (start) or remove (stop) iSCSI targets.

Supported Parameters

OCF_RESKEY_portal=portal

The iSCSI portal address in the form: {ip_address|hostname}[:"port"]

OCF_RESKEY_target=target

The iSCSI target.

OCF_RESKEY_discovery_type=discovery_type

Discovery type. Currently, with open-iscsi, only the sendtargets type is supported.

OCF_RESKEY_iscsiadm=iscsiadm

iscsiadm program path.

OCF_RESKEY_udev=udev

If the next resource depends on the udev creating a device then we wait until it is finished. On a normally loaded host this should be done quickly, but you may be unlucky. If you are not using udev set this to "no", otherwise we will spin in a loop until a timeout occurs.

ocf:ldirectord (7)

ocf:ldirectord — Wrapper OCF Resource Agent for ldirectord

Synopsis

```
OCF_RESKEY_configfile=string [OCF_RESKEY_ldirectord=string]  
ldirectord [start | stop | monitor | meta-data | validate-all]
```

Description

It's a simple OCF RA wrapper for ldirectord and uses the ldirectord interface to create the OCF compliant interface. You win monitoring of ldirectord. Be warned: Asking ldirectord status is an expensive action.

Supported Parameters

OCF_RESKEY_configfile=configuration file path
The full pathname of the ldirectord configuration file.

OCF_RESKEY_ldirectord=ldirectord binary path
The full pathname of the ldirectord.

ocf:LinuxSCSI (7)

ocf:LinuxSCSI — Enables and disables SCSI devices through the kernel SCSI hot-plug subsystem (deprecated)

Synopsis

```
[OCF_RESKEY_scsi=string] [OCF_RESKEY_ignore_deprecation=boolean]  
LinuxSCSI [start | stop | methods | status | monitor | meta-data | validate-all]
```

Description

Deprecation warning: This agent makes use of Linux SCSI hot-plug functionality which has been superseded by SCSI reservations. It is deprecated and may be removed from a future release. See the `scsi2reservation` and `sfex` agents for alternatives. -- This is a resource agent for LinuxSCSI. It manages the availability of a SCSI device from the point of view of the linux kernel. It make Linux believe the device has gone away, and it can make it come back again.

Supported Parameters

`OCF_RESKEY_scsi=SCSI instance`
The SCSI instance to be managed.

`OCF_RESKEY_ignore_deprecation=Suppress deprecation warning`
If set to true, suppresses the deprecation warning for this agent.

ocf:LVM (7)

ocf:LVM — Controls the availability of an LVM Volume Group

Synopsis

```
[OCF_RESKEY_volgrpname=string] [OCF_RESKEY_exclusive=string] LVM  
[start | stop | status | monitor | methods | meta-data | validate-all]
```

Description

Resource script for LVM. It manages an Linux Volume Manager volume (LVM) as an HA resource.

Supported Parameters

OCF_RESKEY_volgrpname=Volume group name
The name of volume group.

OCF_RESKEY_exclusive=Exclusive activation
If set, the volume group will be activated exclusively.

ocf:MailTo (7)

ocf:MailTo — Notifies recipients by email in the event of resource takeover

Synopsis

```
[OCF_RESKEY_email=string] [OCF_RESKEY_subject=string] MailTo [start |  
stop | status | monitor | meta-data | validate-all]
```

Description

This is a resource agent for MailTo. It sends email to a sysadmin whenever a takeover occurs.

Supported Parameters

OCF_RESKEY_email=Email address
The email address of sysadmin.

OCF_RESKEY_subject=Subject
The subject of the email.

ocf:ManageRAID (7)

ocf:ManageRAID — Manages RAID devices

Synopsis

[OCF_RESKEY_raidname=string] ManageRAID [start | stop | status | monitor |
validate-all | meta-data]

Description

Manages starting, stopping and monitoring of RAID devices which are preconfigured in /etc/conf.d/HB-ManageRAID.

Supported Parameters

OCF_RESKEY_raidname=RAID name

Name (case sensitive) of RAID to manage. (preconfigured in /etc/conf.d/HB-
ManageRAID)

ocf:ManageVE (7)

ocf:ManageVE — Manages an OpenVZ Virtual Environment (VE)

Synopsis

```
[OCF_RESKEY_veid=integer] ManageVE [start | stop | status | monitor | validate-all  
| meta-data]
```

Description

This OCF complaint resource agent manages OpenVZ VEs and thus requires a proper OpenVZ installation including a recent vzctl util.

Supported Parameters

OCF_RESKEY_veid=OpenVZ ID of VE

OpenVZ ID of virtual environment (see output of vzlist -a for all assigned IDs)

ocf:mysql-proxy (7)

ocf:mysql-proxy — Manages a MySQL Proxy daemon

Synopsis

```
[OCF_RESKEY_binary=string] OCF_RESKEY_defaults_file=string  
[OCF_RESKEY_proxy_backend_addresses=string]  
[OCF_RESKEY_proxy_read_only_backend_addresses=string]  
[OCF_RESKEY_proxy_address=string] [OCF_RESKEY_log_level=string]  
[OCF_RESKEY_heartbeat=string] [OCF_RESKEY_admin_address=string]  
[OCF_RESKEY_admin_username=string]  
[OCF_RESKEY_admin_password=string]  
[OCF_RESKEY_admin_lua_script=string]  
[OCF_RESKEY_parameters=string] OCF_RESKEY_pidfile=string  
mysql-proxy [start | stop | reload | monitor | validate-all | meta-data]
```

Description

This script manages MySQL Proxy as an OCF resource in a high-availability setup.
Tested with MySQL Proxy 0.7.0 on Debian 5.0.

Supported Parameters

OCF_RESKEY_binary=Full path to MySQL Proxy binary
Full path to the MySQL Proxy binary. For example, "/usr/sbin/mysql-proxy".

OCF_RESKEY_defaults_file=Full path to configuration file
Full path to a MySQL Proxy configuration file. For example, "/etc/mysql-proxy.conf".

OCF_RESKEY_proxy_backend_addresses=MySQL Proxy backend-servers
Address:port of the remote backend-servers (default: 127.0.0.1:3306).

OCF_RESKEY_proxy_read_only_backend_addresses=MySQL Proxy read only backend-servers

Address:port of the remote (read only) slave-server (default:).

OCF_RESKEY_proxy_address=MySQL Proxy listening address

Listening address:port of the proxy-server (default: :4040). You can also specify a socket like "/tmp/mysql-proxy.sock".

OCF_RESKEY_log_level=MySQL Proxy log level.

Log all messages of level (error|warning|info|message|debug) or higher. An empty value disables logging.

OCF_RESKEY_keepalive=Use keepalive option

Try to restart the proxy if it crashed (default:). Valid values: true or false. An empty value equals "false".

OCF_RESKEY_admin_address=MySQL Proxy admin-server address

Listening address:port of the admin-server (default: 127.0.0.1:4041).

OCF_RESKEY_admin_username=MySQL Proxy admin-server username

Username to allow to log in (default:).

OCF_RESKEY_admin_password=MySQL Proxy admin-server password

Password to allow to log in (default:).

OCF_RESKEY_admin_lua_script=MySQL Proxy admin-server lua script

Script to execute by the admin plugin.

OCF_RESKEY_parameters=MySQL Proxy additional parameters

The MySQL Proxy daemon may be called with additional parameters. Specify any of them here.

OCF_RESKEY_pidfile=PID file

PID file

ocf:mysql (7)

ocf:mysql — Manages a MySQL database instance

Synopsis

```
[OCF_RESKEY_binary=string] [OCF_RESKEY_config=string]
[OCF_RESKEY_datadir=string] [OCF_RESKEY_user=string]
[OCF_RESKEY_group=string] [OCF_RESKEY_log=string]
[OCF_RESKEY_pid=string] [OCF_RESKEY_socket=string]
[OCF_RESKEY_test_table=string] [OCF_RESKEY_test_user=string]
[OCF_RESKEY_test_passwd=string]
[OCF_RESKEY_enable_creation=integer]
[OCF_RESKEY_additional_parameters=string]
[OCF_RESKEY_replication_user=string]
[OCF_RESKEY_replication_passwd=string] mysql [start | stop | status | monitor
| monitor | monitor | notify | promote | demote | validate-all | meta-data]
```

Description

Resource script for MySQL. It manages a MySQL Database instance as an HA resource.

Supported Parameters

OCF_RESKEY_binary=MySQL binary
Location of the MySQL binary

OCF_RESKEY_config=MySQL config
Configuration file

OCF_RESKEY_datadir=MySQL datadir
Directory containing databases

OCF_RESKEY_user=MySQL user
User running MySQL daemon

OCF_RESKEY_group=MySQL group
Group running MySQL daemon (for logfile and directory permissions)

OCF_RESKEY_log=MySQL log file
The logfile to be used for mysqld.

OCF_RESKEY_pid=MySQL pid file
The pidfile to be used for mysqld.

OCF_RESKEY_socket=MySQL socket
The socket to be used for mysqld.

OCF_RESKEY_test_table=MySQL test table
Table to be tested in monitor statement (in database.table notation)

OCF_RESKEY_test_user=MySQL test user
MySQL test user

OCF_RESKEY_test_passwd=MySQL test user password
MySQL test user password

OCF_RESKEY_enable_creation=Create the database if it does not exist
If the MySQL database does not exist, it will be created

OCF_RESKEY_additional_parameters=Additional parameters to pass to mysqld
Additional parameters which are passed to the mysqld on startup. (e.g. --skip-external-locking or --skip-grant-tables)

OCF_RESKEY_replication_user=MySQL replication user
MySQL replication user. Used for replication client and slave.

OCF_RESKEY_replication_passwd=MySQL replication user password
MySQL replication password. Used for replication client and slave.

ocf:nfsserver (7)

ocf:nfsserver — Manages an NFS server

Synopsis

```
[OCF_RESKEY_nfs_init_script=string]  
[OCF_RESKEY_nfs_notify_cmd=string]  
[OCF_RESKEY_nfs_shared_infodir=string] [OCF_RESKEY_nfs_ip=string]  
nfsserver [start | stop | monitor | meta-data | validate-all]
```

Description

Nfsserver helps to manage the Linux nfs server as a failover-able resource in Linux-HA. It depends on Linux specific NFS implementation details, so is considered not portable to other platforms yet.

Supported Parameters

OCF_RESKEY_nfs_init_script= Init script for nfsserver

The default init script shipped with the Linux distro. The nfsserver resource agent offloads the start/stop/monitor work to the init script because the procedure to start/stop/monitor nfsserver varies on different Linux distro.

OCF_RESKEY_nfs_notify_cmd= The tool to send out notification.

The tool to send out NSM reboot notification. Failover of nfsserver can be considered as rebooting to different machines. The nfsserver resource agent use this command to notify all clients about the happening of failover.

OCF_RESKEY_nfs_shared_infodir= Directory to store nfs server related information.

The nfsserver resource agent will save nfs related information in this specific directory. And this directory must be able to fail-over before nfsserver itself.

OCF_RESKEY_nfs_ip= IP address.

The floating IP address used to access the nfs service

ocf:oracle (7)

ocf:oracle — Manages an Oracle Database instance

Synopsis

```
OCF_RESKEY_sid=string [OCF_RESKEY_home=string]
[OCF_RESKEY_user=string] [OCF_RESKEY_ipcrm=string]
[OCF_RESKEY_clear_backupmode=boolean]
[OCF_RESKEY_shutdown_method=string] oracle [start | stop | status | monitor
| validate-all | methods | meta-data]
```

Description

Resource script for oracle. Manages an Oracle Database instance as an HA resource.

Supported Parameters

OCF_RESKEY_sid=sid
The Oracle SID (aka ORACLE_SID).

OCF_RESKEY_home=home
The Oracle home directory (aka ORACLE_HOME). If not specified, then the SID along with its home should be listed in /etc/oratab.

OCF_RESKEY_user=user
The Oracle owner (aka ORACLE_OWNER). If not specified, then it is set to the owner of file \$ORACLE_HOME/dbs/*\${ORACLE_SID}.ora. If this does not work for you, just set it explicitly.

OCF_RESKEY_ipcrm=ipcrm
Sometimes IPC objects (shared memory segments and semaphores) belonging to an Oracle instance might be left behind which prevents the instance from starting. It is not easy to figure out which shared segments belong to which instance, in particular when more instances are running as same user. What we use here is the

"oradebug" feature and its "ipc" trace utility. It is not optimal to parse the debugging information, but I am not aware of any other way to find out about the IPC information. In case the format or wording of the trace report changes, parsing might fail. There are some precautions, however, to prevent stepping on other peoples toes. There is also a dumpinstipc option which will make us print the IPC objects which belong to the instance. Use it to see if we parse the trace file correctly. Three settings are possible: - none: don't mess with IPC and hope for the best (beware: you'll probably be out of luck, sooner or later) - instance: try to figure out the IPC stuff which belongs to the instance and remove only those (default; should be safe) - orauser: remove all IPC belonging to the user which runs the instance (don't use this if you run more than one instance as same user or if other apps running as this user use IPC) The default setting "instance" should be safe to use, but in that case we cannot guarantee that the instance will start. In case IPC objects were already left around, because, for instance, someone mercilessly killing Oracle processes, there is no way any more to find out which IPC objects should be removed. In that case, human intervention is necessary, and probably all instances running as same user will have to be stopped. The third setting, "orauser", guarantees IPC objects removal, but it does that based only on IPC objects ownership, so you should use that only if every instance runs as separate user. Please report any problems. Suggestions/fixes welcome.

```
OCF_RESKEY_clear_backupmode=clear_backupmode
```

The clear of the backup mode of ORACLE.

```
OCF_RESKEY_shutdown_method=shutdown_method
```

How to stop Oracle is a matter of taste it seems. The default method ("checkpoint/abort") is: alter system checkpoint; shutdown abort; This should be the fastest safe way bring the instance down. If you find "shutdown abort" distasteful, set this attribute to "immediate" in which case we will shutdown immediate; If you still think that there's even better way to shutdown an Oracle instance we are willing to listen.

ocf:oralsnr (7)

ocf:oralsnr — Manages an Oracle TNS listener

Synopsis

```
OCF_RESKEY_sid=string [OCF_RESKEY_home=string]  
[OCF_RESKEY_user=string] OCF_RESKEY_listener=string oralsnr [start |  
stop | status | monitor | validate-all | meta-data | methods]
```

Description

Resource script for Oracle Listener. It manages an Oracle Listener instance as an HA resource.

Supported Parameters

OCF_RESKEY_sid=sid

The Oracle SID (aka ORACLE_SID). Necessary for the monitor op, i.e. to do tnsping SID.

OCF_RESKEY_home=home

The Oracle home directory (aka ORACLE_HOME). If not specified, then the SID should be listed in /etc/oratab.

OCF_RESKEY_user=user

Run the listener as this user.

OCF_RESKEY_listener=listener

Listener instance to be started (as defined in listener.ora). Defaults to LISTENER.

ocf:pgsql (7)

ocf:pgsql — Manages a PostgreSQL database instance

Synopsis

```
[OCF_RESKEY_pgctl=string] [OCF_RESKEY_start_opt=string]
[OCF_RESKEY_ctl_opt=string] [OCF_RESKEY_psql=string]
[OCF_RESKEY_pgdata=string] [OCF_RESKEY_pgdba=string]
[OCF_RESKEY_pghost=string] [OCF_RESKEY_pgport=string]
[OCF_RESKEY_pgdb=string] [OCF_RESKEY_logfile=string]
[OCF_RESKEY_stop_escalate=string] psql [start | stop | status | monitor |
meta-data | validate-all | methods]
```

Description

Resource script for PostgreSQL. It manages a PostgreSQL as an HA resource.

Supported Parameters

OCF_RESKEY_pgctl=pgctl
Path to pg_ctl command.

OCF_RESKEY_start_opt=start_opt
Start options (-o start_opt in pgi_ctl). "-i -p 5432" for example.

OCF_RESKEY_ctl_opt=ctl_opt
Additional pg_ctl options (-w, -W etc..). Default is ""

OCF_RESKEY_psql=psql
Path to psql command.

OCF_RESKEY_pgdata=pgdata
Path PostgreSQL data directory.

OCF_RESKEY_pgdba=pgdba
User that owns PostgreSQL.

OCF_RESKEY_pghost=pghost
Hostname/IP Address where PostgreSQL is listening

OCF_RESKEY_pgport=pgport
Port where PostgreSQL is listening

OCF_RESKEY_pgdb=pgdb
Database that will be used for monitoring.

OCF_RESKEY_logfile=logfile
Path to PostgreSQL server log output file.

OCF_RESKEY_stop_escalate=stop escalation
Number of retries (using -m fast) before resorting to -m immediate

ocf:pingd (7)

ocf:pingd — Monitors connectivity to specific hosts or IP addresses ("ping nodes")
(deprecated)

Synopsis

```
[OCF_RESKEY_pidfile=string] [OCF_RESKEY_user=string]  
[OCF_RESKEY_dampen=integer] [OCF_RESKEY_set=integer]  
[OCF_RESKEY_name=integer] [OCF_RESKEY_section=integer]  
[OCF_RESKEY_multiplier=integer] [OCF_RESKEY_host_list=integer]  
[OCF_RESKEY_ignore_deprecation=boolean] pingd [start | stop | monitor |  
meta-data | validate-all]
```

Description

Deprecation warning: This agent is deprecated and may be removed from a future release. See the ocf:pacemaker:pingd resource agent for a supported alternative. -- This is a pingd Resource Agent. It records (in the CIB) the current number of ping nodes a node can connect to.

Supported Parameters

OCF_RESKEY_pidfile=PID file
PID file

OCF_RESKEY_user=The user we want to run pingd as
The user we want to run pingd as

OCF_RESKEY_dampen=Dampening interval
The time to wait (dampening) further changes occur

OCF_RESKEY_set=Set name
The name of the instance_attributes set to place the value in. Rarely needs to be specified.

OCF_RESKEY_name=Attribute name

The name of the attributes to set. This is the name to be used in the constraints.

OCF_RESKEY_section=Section name

The section place the value in. Rarely needs to be specified.

OCF_RESKEY_multiplier=Value multiplier

The number by which to multiply the number of connected ping nodes by

OCF_RESKEY_host_list=Host list

The list of ping nodes to count. Defaults to all configured ping nodes. Rarely needs to be specified.

OCF_RESKEY_ignore_deprecation=Suppress deprecation warning

If set to true, suppresses the deprecation warning for this agent.

ocf:portblock (7)

ocf:portblock — Block and unblocks access to TCP and UDP ports

Synopsis

```
[OCF_RESKEY_protocol=string] [OCF_RESKEY_portno=integer]
[OCF_RESKEY_action=string] [OCF_RESKEY_ip=string]
[OCF_RESKEY_tickle_dir=string] [OCF_RESKEY_sync_script=string]
portblock [start | stop | status | monitor | meta-data | validate-all]
```

Description

Resource script for portblock. It is used to temporarily block ports using iptables. In addition, it may allow for faster TCP reconnects for clients on failover. Use that if there are long lived TCP connections to an HA service. This feature is enabled by setting the tickle_dir parameter and only in concert with action set to unblock. Note that the tickle ACK function is new as of version 3.0.2 and hasn't yet seen widespread use.

Supported Parameters

OCF_RESKEY_protocol=protocol
The protocol used to be blocked/unblocked.

OCF_RESKEY_portno=portno
The port number used to be blocked/unblocked.

OCF_RESKEY_action=action
The action (block/unblock) to be done on the protocol::portno.

OCF_RESKEY_ip=ip
The IP address used to be blocked/unblocked.

OCF_RESKEY_tickle_dir=Tickle directory

The shared or local directory (must be absolute path) which stores the established TCP connections.

OCF_RESKEY_sync_script=Connection state file synchronization script

If the tickle_dir is a local directory, then the TCP connection state file has to be replicated to other nodes in the cluster. It can be csync2 (default), some wrapper of rsync, or whatever. It takes the file name as a single argument. For csync2, set it to "csync2 -xv".

ocf:proftpd (7)

ocf:proftpd — OCF Resource Agent compliant FTP script.

Synopsis

```
[OCF_RESKEY_binary=string] [OCF_RESKEY_confdir=string]
[OCF_RESKEY_pidfile=string] [OCF_RESKEY_curl_binary=string]
[OCF_RESKEY_curl_url=string] [OCF_RESKEY_test_user=string]
[OCF_RESKEY_test_pass=string] proftpd [start | stop | monitor | monitor |
validate-all | meta-data]
```

Description

This script manages Proftpd in an Active-Passive setup

Supported Parameters

OCF_RESKEY_binary=The Proftpd binary
The Proftpd binary

OCF_RESKEY_confdir=Configuration file name with full path
The Proftpd configuration file name with full path. For example, "/etc/proftpd.conf"

OCF_RESKEY_pidfile=PID file
The Proftpd PID file. The location of the PID file is configured in the Proftpd configuration file.

OCF_RESKEY_curl_binary=The absolut path to the curl binary
The absolut path to the curl binary for monitoring with OCF_CHECK_LEVEL greater zero.

OCF_RESKEY_curl_url=The URL which is checked by curl
The URL which is checked by curl with OCF_CHECK_LEVEL greater zero.

OCF_RESKEY_test_user=The name of the ftp user

The name of the ftp user for monitoring with OCF_CHECK_LEVEL greater zero.

OCF_RESKEY_test_pass=The password of the ftp user

The password of the ftp user for monitoring with OCF_CHECK_LEVEL greater zero.

ocf:Pure-FTPd (7)

ocf:Pure-FTPd — Manages a Pure-FTPd FTP server instance

Synopsis

```
OCF_RESKEY_script=string OCF_RESKEY_conf=string  
OCF_RESKEY_daemon_type=string [OCF_RESKEY_pidfile=string]  
Pure-FTPd [start | stop | monitor | validate-all | meta-data]
```

Description

This script manages Pure-FTPd in an Active-Passive setup

Supported Parameters

OCF_RESKEY_script=Script name with full path
The full path to the Pure-FTPd startup script. For example, "/sbin/pure-config.pl"

OCF_RESKEY_conf=Configuration file name with full path
The Pure-FTPd configuration file name with full path. For example, "/etc/pure-ftp/pure-ftp.conf"

OCF_RESKEY_daemon_type=Configuration file name with full path
The Pure-FTPd daemon to be called by pure-ftp-wrapper. Valid options are "" for pure-ftp, "mysql" for pure-ftp-mysql, "postgresql" for pure-ftp-postgresql and "ldap" for pure-ftp-ldap

OCF_RESKEY_pidfile=PID file
PID file

ocf:Raid1 (7)

ocf:Raid1 — Manages a software RAID1 device on shared storage

Synopsis

```
[OCF_RESKEY_raidconf=string] [OCF_RESKEY_raiddev=string]  
[OCF_RESKEY_homehost=string] Raid1 [start | stop | status | monitor | validate-  
all | meta-data]
```

Description

Resource script for RAID1. It manages a software Raid1 device on a shared storage medium.

Supported Parameters

OCF_RESKEY_raidconf=RAID config file
The RAID configuration file. e.g. /etc/raidtab or /etc/mdadm.conf.

OCF_RESKEY_raiddev=block device
The block device to use.

OCF_RESKEY_homehost=Homehost for mdadm
The value for the homehost directive; this is an mdadm feature to protect RAIDs against being activated by accident. It is recommended to create RAIDs managed by the cluster with "homehost" set to a special value, so they are not accidentally auto-assembled by nodes not supposed to own them.

ocf:Route (7)

ocf:Route — Manages network routes

Synopsis

```
OCF_RESKEY_destination=string OCF_RESKEY_device=string  
OCF_RESKEY_gateway=string OCF_RESKEY_source=string  
[OCF_RESKEY_table=string] Route [start | stop | monitor | reload | meta-data |  
validate-all]
```

Description

Enables and disables network routes. Supports host and net routes, routes via a gateway address, and routes using specific source addresses. This resource agent is useful if a node's routing table needs to be manipulated based on node role assignment. Consider the following example use case: - One cluster node serves as an IPsec tunnel endpoint. - All other nodes use the IPsec tunnel to reach hosts in a specific remote network. Then, here is how you would implement this scheme making use of the Route resource agent: - Configure an ipsec LSB resource. - Configure a cloned Route OCF resource. - Create an order constraint to ensure that ipsec is started before Route. - Create a colocation constraint between the ipsec and Route resources, to make sure no instance of your cloned Route resource is started on the tunnel endpoint itself.

Supported Parameters

OCF_RESKEY_destination=Destination network

The destination network (or host) to be configured for the route. Specify the netmask suffix in CIDR notation (e.g. "/24"). If no suffix is given, a host route will be created. Specify "0.0.0.0/0" or "default" if you want this resource to set the system default route.

OCF_RESKEY_device=Outgoing network device

The outgoing network device to use for this route.

OCF_RESKEY_gateway=Gateway IP address
The gateway IP address to use for this route.

OCF_RESKEY_source=Source IP address
The source IP address to be configured for the route.

OCF_RESKEY_table=Routing table
The routing table to be configured for the route.

ocf:rsyncd (7)

ocf:rsyncd — Manages an rsync daemon

Synopsis

```
[OCF_RESKEY_binpath=string] [OCF_RESKEY_conf file=string]  
[OCF_RESKEY_bwlimit=string] rsyncd [start | stop | monitor | validate-all | meta-  
data]
```

Description

This script manages rsync daemon

Supported Parameters

OCF_RESKEY_binpath=Full path to the rsync binary
The rsync binary path. For example, "/usr/bin/rsync"

OCF_RESKEY_conf file=Configuration file name with full path
The rsync daemon configuration file name with full path. For example,
"/etc/rsyncd.conf"

OCF_RESKEY_bwlimit=limit I/O bandwidth, KBytes per second
This option allows you to specify a maximum transfer rate in kilobytes per second.
This option is most effective when using rsync with large files (several megabytes
and up). Due to the nature of rsync transfers, blocks of data are sent, then if rsync
determines the transfer was too fast, it will wait before sending the next data block.
The result is an average transfer rate equaling the specified limit. A value of zero
specifies no limit.

ocf:SAPDatabase (7)

ocf:SAPDatabase — Manages any SAP database (based on Oracle, MaxDB, or DB2)

Synopsis

```
OCF_RESKEY_SID=string OCF_RESKEY_DIR_EXECUTABLE=string
OCF_RESKEY_DBTYPE=string OCF_RESKEY_NETSERVICENAME=string
OCF_RESKEY_DBJ2EE_ONLY=boolean OCF_RESKEY_JAVA_HOME=string
OCF_RESKEY_STRICT_MONITORING=boolean
OCF_RESKEY_AUTOMATIC_RECOVER=boolean
OCF_RESKEY_DIR_BOOTSTRAP=string OCF_RESKEY_DIR_SECSTORE=string
OCF_RESKEY_DB_JARS=string OCF_RESKEY_PRE_START_USEREXIT=string
OCF_RESKEY_POST_START_USEREXIT=string
OCF_RESKEY_PRE_STOP_USEREXIT=string
OCF_RESKEY_POST_STOP_USEREXIT=string SAPDatabase [start | stop | status
| monitor | validate-all | meta-data | methods]
```

Description

Resource script for SAP databases. It manages a SAP database of any type as an HA resource.

Supported Parameters

OCF_RESKEY_SID=SAP system ID

The unique SAP system identifier. e.g. P01

OCF_RESKEY_DIR_EXECUTABLE=path of sapstartsrv and sapcontrol

The full qualified path where to find sapstartsrv and sapcontrol.

OCF_RESKEY_DBTYPE=database vendor

The name of the database vendor you use. Set either: ORA,DB6,ADA

OCF_RESKEY_NETSERVICENAME=listener name

The Oracle TNS listener name.

OCF_RESKEY_DBJ2EE_ONLY=only JAVA stack installed

If you do not have a ABAP stack installed in the SAP database, set this to TRUE

OCF_RESKEY_JAVA_HOME=Path to Java SDK

This is only needed if the DBJ2EE_ONLY parameter is set to true. Enter the path to the Java SDK which is used by the SAP WebAS Java

OCF_RESKEY_STRICT_MONITORING=Activates application level monitoring

This controls how the resource agent monitors the database. If set to true, it will use SAP tools to test the connect to the database. Do not use with Oracle, because it will result in unwanted failovers in case of an archiver stuck

OCF_RESKEY_AUTOMATIC_RECOVER=Enable or disable automatic startup recovery

The SAPDatabase resource agent tries to recover a failed start attempt automatically one time. This is done by running a forced abort of the RDBMS and/or executing recovery commands.

OCF_RESKEY_DIR_BOOTSTRAP=path to j2ee bootstrap directory

The full qualified path where to find the J2EE instance bootstrap directory. e.g.
/usr/sap/P01/J00/j2ee/cluster/bootstrap

OCF_RESKEY_DIR_SECSTORE=path to j2ee secure store directory

The full qualified path where to find the J2EE security store directory. e.g.
/usr/sap/P01/SYS/global/security/lib/tools

OCF_RESKEY_DB_JARS=file name of the jdbc driver

The full qualified filename of the jdbc driver for the database connection test. It will be automatically read from the bootstrap.properties file in Java engine 6.40 and 7.00. For Java engine 7.10 the parameter is mandatory.

OCF_RESKEY_PRE_START_USEREXIT=path to a pre-start script

The full qualified path where to find a script or program which should be executed before this resource gets started.

OCF_RESKEY_POST_START_USEREXIT=path to a post-start script

The full qualified path where to find a script or program which should be executed after this resource got started.

OCF_RESKEY_PRE_STOP_USEREXIT=path to a pre-start script

The full qualified path where to find a script or program which should be executed before this resource gets stopped.

OCF_RESKEY_POST_STOP_USEREXIT=path to a post-start script

The full qualified path where to find a script or program which should be executed after this resource got stopped.

ocf:SAPInstance (7)

ocf:SAPInstance — Manages a SAP instance

Synopsis

```
OCF_RESKEY_InstanceName=string OCF_RESKEY_DIR_EXECUTABLE=string
OCF_RESKEY_DIR_PROFILE=string OCF_RESKEY_START_PROFILE=string
OCF_RESKEY_START_WAITTIME=string
OCF_RESKEY_AUTOMATIC_RECOVER=boolean
OCF_RESKEY_MONITOR_SERVICES=string
OCF_RESKEY_ERS_InstanceName=string
OCF_RESKEY_ERS_START_PROFILE=string
OCF_RESKEY_PRE_START_USEREXIT=string
OCF_RESKEY_POST_START_USEREXIT=string
OCF_RESKEY_PRE_STOP_USEREXIT=string
OCF_RESKEY_POST_STOP_USEREXIT=string SAPInstance [start | stop | status
| monitor | promote | demote | validate-all | meta-data | methods]
```

Description

Resource script for SAP. It manages a SAP Instance as an HA resource.

Supported Parameters

OCF_RESKEY_InstanceName=**instance name: SID_INSTANCE_VIR-HOSTNAME**
The full qualified SAP instance name. e.g. P01_DVEBMGS00_sapp01ci

OCF_RESKEY_DIR_EXECUTABLE=**path of sapstartsrv and sapcontrol**
The full qualified path where to find sapstartsrv and sapcontrol.

OCF_RESKEY_DIR_PROFILE=**path of start profile**
The full qualified path where to find the SAP START profile.

OCF_RESKEY_START_PROFILE=start profile name

The name of the SAP START profile.

OCF_RESKEY_START_WAITTIME=Check the successful start after that time (do not wait for J2EE-Addin)

After that time in seconds a monitor operation is executed by the resource agent.

Does the monitor return SUCCESS, the start is handled as SUCCESS. This is useful to resolve timing problems with e.g. the J2EE-Addin instance.

OCF_RESKEY_AUTOMATIC_RECOVER=Enable or disable automatic startup recovery

The SAPInstance resource agent tries to recover a failed start attempt automatically one time. This is done by killing running instance processes and executing cleanipc.

OCF_RESKEY_MONITOR_SERVICES=

OCF_RESKEY_ERS_InstanceName=

OCF_RESKEY_ERS_START_PROFILE=

OCF_RESKEY_PRE_START_USEREXIT=path to a pre-start script

The full qualified path where to find a script or program which should be executed before this resource gets started.

OCF_RESKEY_POST_START_USEREXIT=path to a post-start script

The full qualified path where to find a script or program which should be executed after this resource got started.

OCF_RESKEY_PRE_STOP_USEREXIT=path to a pre-stop script

The full qualified path where to find a script or program which should be executed before this resource gets stopped.

OCF_RESKEY_POST_STOP_USEREXIT=path to a post-stop script

The full qualified path where to find a script or program which should be executed after this resource got stopped.

ocf:scsi2reservation (7)

ocf:scsi2reservation — scsi-2 reservation

Synopsis

```
[OCF_RESKEY_scsi_reserve=string] [OCF_RESKEY_sharedisk=string]  
[OCF_RESKEY_start_loop=string] scsi2reservation [start | stop | monitor  
| meta-data | validate-all]
```

Description

The scsi-2-reserve resource agent is a place holder for SCSI-2 reservation. A healthy instance of scsi-2-reserve resource, indicates the own of the specified SCSI device. This resource agent depends on the scsi_reserve from scsires package, which is Linux specific.

Supported Parameters

OCF_RESKEY_scsi_reserve=Manages exclusive access to shared storage media through SCSI-2 reservations

The `scsi_reserve` is a command from scsires package. It helps to issue SCSI-2 reservation on SCSI devices.

OCF_RESKEY_sharedisk= Shared disk.

The shared disk that can be reserved.

OCF_RESKEY_start_loop= Times to re-try before giving up.

We are going to try several times before giving up. `Start_loop` indicates how many times we are going to re-try.

ocf:SendArp (7)

ocf:SendArp — Broadcasts unsolicited ARP announcements

Synopsis

```
[OCF_RESKEY_ip=string] [OCF_RESKEY_nic=string] SendArp [start | stop |  
monitor | meta-data | validate-all]
```

Description

This script send out gratuitous Arp for an IP address

Supported Parameters

OCF_RESKEY_ip=IP address

The IP address for sending arp package.

OCF_RESKEY_nic=NIC

The nic for sending arp package.

ocf:ServeRAID (7)

ocf:ServeRAID — Enables and disables shared ServeRAID merge groups

Synopsis

```
[OCF_RESKEY_serveraid=integer] [OCF_RESKEY_mergegroup=integer]  
ServeRAID [start | stop | status | monitor | validate-all | meta-data | methods]
```

Description

Resource script for ServeRAID. It enables/disables shared ServeRAID merge groups.

Supported Parameters

OCF_RESKEY_serveraid=serveraid
The adapter number of the ServeRAID adapter.

OCF_RESKEY_mergegroup=mergegroup
The logical drive under consideration.

ocf:sfex (7)

ocf:sfex — Manages exclusive access to shared storage using Shared Disk File EXclusiveness (SF-EX)

Synopsis

```
[OCF_RESKEY_device=string] [OCF_RESKEY_index=integer]
[OCF_RESKEY_collision_timeout=integer]
[OCF_RESKEY_monitor_interval=integer]
[OCF_RESKEY_lock_timeout=integer] sfex [start | stop | monitor | meta-data]
```

Description

Resource script for SF-EX. It manages a shared storage medium exclusively .

Supported Parameters

OCF_RESKEY_device=block device

Block device path that stores exclusive control data.

OCF_RESKEY_index=index

Location in block device where exclusive control data is stored. 1 or more is specified. Default is 1.

OCF_RESKEY_collision_timeout=waiting time for lock acquisition

Waiting time when a collision of lock acquisition is detected. Default is 1 second.

OCF_RESKEY_monitor_interval=monitor interval

Monitor interval(sec). Default is 10 seconds

OCF_RESKEY_lock_timeout=Valid term of lock

Valid term of lock(sec). Default is 20 seconds.

ocf:SphinxSearchDaemon (7)

ocf:SphinxSearchDaemon — Manages the Sphinx search daemon.

Synopsis

```
OCF_RESKEY_config=string [OCF_RESKEY_searchd=string]
[OCF_RESKEY_search=string] [OCF_RESKEY_testQuery=string]
SphinxSearchDaemon [start | stop | monitor | meta-data | validate-all]
```

Description

This is a searchd Resource Agent. It manages the Sphinx Search Daemon.

Supported Parameters

OCF_RESKEY_config=Configuration file
searchd configuration file

OCF_RESKEY_searchd=searchd binary
searchd binary

OCF_RESKEY_search=search binary
Search binary for functional testing in the monitor action.

OCF_RESKEY_testQuery=test query
Test query for functional testing in the monitor action. The query does not need to match any documents in the index. The purpose is merely to test whether the search daemon is able to query its indices and respond properly.

ocf:Squid (7)

ocf:Squid — Manages a Squid proxy server instance

Synopsis

```
[OCF_RESKEY_squid_exe=string] OCF_RESKEY_squid_conf=string  
OCF_RESKEY_squid_pidfile=string OCF_RESKEY_squid_port=integer  
[OCF_RESKEY_squid_stop_timeout=integer]  
[OCF_RESKEY_debug_mode=string] [OCF_RESKEY_debug_log=string] Squid  
[start | stop | status | monitor | meta-data | validate-all]
```

Description

The resource agent of Squid. This manages a Squid instance as an HA resource.

Supported Parameters

OCF_RESKEY_squid_exe=Executable file

This is a required parameter. This parameter specifies squid's executable file.

OCF_RESKEY_squid_conf=Configuration file

This is a required parameter. This parameter specifies a configuration file for a squid instance managed by this RA.

OCF_RESKEY_squid_pidfile=Pidfile

This is a required parameter. This parameter specifies a process id file for a squid instance managed by this RA.

OCF_RESKEY_squid_port=Port number

This is a required parameter. This parameter specifies a port number for a squid instance managed by this RA. If plural ports are used, you must specify the only one of them.

OCF_RESKEY_squid_stop_timeout=Number of seconds to await to confirm a normal stop method

This is an omittable parameter. On a stop action, a normal stop method is firstly used. and then the confirmation of its completion is awaited for the specified seconds by this parameter. The default value is 10.

OCF_RESKEY_debug_mode=Debug mode

This is an optional parameter. This RA runs in debug mode when this parameter includes 'x' or 'v'. If 'x' is included, both of STDOUT and STDERR redirect to the logfile specified by "debug_log", and then the builtin shell option 'x' is turned on. It is similar about 'v'.

OCF_RESKEY_debug_log=A destination of the debug log

This is an optional and omittable parameter. This parameter specifies a destination file for debug logs and works only if this RA run in debug mode. Refer to "debug_mode" about debug mode. If no value is given but it's required, it's made by the following rules: "/var/log/" as a directory part, the basename of the configuration file given by "syslog_ng_conf" as a basename part, ".log" as a suffix.

ocf:Stateful (7)

ocf:Stateful — Example stateful resource agent

Synopsis

`OCF_RESKEY_state=string Stateful [start | stop | monitor | meta-data | validate-all]`

Description

This is an example resource agent that impliments two states

Supported Parameters

`OCF_RESKEY_state=State file`
Location to store the resource state in

ocf:SysInfo (7)

ocf:SysInfo — Records various node attributes in the CIB

Synopsis

```
[OCF_RESKEY_pidfile=string] [OCF_RESKEY_delay=string] SysInfo [start  
| stop | monitor | meta-data | validate-all]
```

Description

This is a SysInfo Resource Agent. It records (in the CIB) various attributes of a node
Sample Linux output: arch: i686 os: Linux-2.4.26-gentoo-r14 free_swap: 1999 cpu_info:
Intel(R) Celeron(R) CPU 2.40GHz cpu_speed: 4771.02 cpu_cores: 1 cpu_load: 0.00
ram_total: 513 ram_free: 117 root_free: 2.4 Sample Darwin output: arch: i386 os:
Darwin-8.6.2 cpu_info: Intel Core Duo cpu_speed: 2.16 cpu_cores: 2 cpu_load: 0.18
ram_total: 2016 ram_free: 787 root_free: 13 Units: free_swap: Mb ram_*: Mb root_free:
Gb cpu_speed (Linux): bogomips cpu_speed (Darwin): Ghz

Supported Parameters

OCF_RESKEY_pidfile=PID file
PID file

OCF_RESKEY_delay=Dampening Delay
Interval to allow values to stabilize

ocf:syslog-ng (7)

ocf:syslog-ng — Syslog-ng resource agent

Synopsis

```
[OCF_RESKEY_configfile=string]  
[OCF_RESKEY_syslog_ng_binary=string]  
[OCF_RESKEY_start_opts=string]  
[OCF_RESKEY_kill_term_timeout=integer] syslog-ng [start | stop | status  
| monitor | meta-data | validate-all]
```

Description

This script manages a syslog-ng instance as an HA resource.

Supported Parameters

`OCF_RESKEY_configfile=`Configuration file

This parameter specifies a configuration file for a syslog-ng instance managed by this RA.

`OCF_RESKEY_syslog_ng_binary=`syslog-ng executable

This parameter specifies syslog-ng's executable file.

`OCF_RESKEY_start_opts=`Start options

This parameter specifies startup options for a syslog-ng instance managed by this RA. When no value is given, no startup options is used. Don't use option '-F'. It causes a stuck of a start action.

`OCF_RESKEY_kill_term_timeout=`Number of seconds to await to confirm a normal stop method

On a stop action, a normal stop method(`pkill -TERM`) is firstly used. And then the confirmation of its completion is waited for the specified seconds by this parameter. The default value is 10.

ocf:tomcat (7)

ocf:tomcat — Manages a Tomcat servlet environment instance

Synopsis

```
OCF_RESKEY_tomcat_name=string OCF_RESKEY_script_log=string
[OCF_RESKEY_tomcat_stop_timeout=integer]
[OCF_RESKEY_tomcat_suspend_trialcount=integer]
[OCF_RESKEY_tomcat_user=string] [OCF_RESKEY_statusurl=string]
[OCF_RESKEY_java_home=string] OCF_RESKEY_catalina_home=string
OCF_RESKEY_catalina_pid=string
[OCF_RESKEY_tomcat_start_opts=string]
[OCF_RESKEY_catalina_opts=string]
[OCF_RESKEY_catalina_rotate_log=string]
[OCF_RESKEY_catalina_rotatetime=integer] tomcat [start | stop | status |
monitor | meta-data | validate-all]
```

Description

Resource script for tomcat. It manages a Tomcat instance as an HA resource.

Supported Parameters

OCF_RESKEY_tomcat_name=The name of the resource
The name of the resource

OCF_RESKEY_script_log=A destination of the log of this script
A destination of the log of this script

OCF_RESKEY_tomcat_stop_timeout=Time-out at the time of the stop
Time-out at the time of the stop

OCF_RESKEY_tomcat_suspend_trialcount=The re-try number of times awaiting a stop

The re-try number of times awaiting a stop

OCF_RESKEY_tomcat_user=A user name to start a resource

A user name to start a resource

OCF_RESKEY_statusurl=URL for state confirmation

URL for state confirmation

OCF_RESKEY_java_home=Home directory of the Java

Home directory of the Java

OCF_RESKEY_catalina_home=Home directory of Tomcat

Home directory of Tomcat

OCF_RESKEY_catalina_pid=A PID file name of Tomcat

A PID file name of Tomcat

OCF_RESKEY_tomcat_start_opts=Tomcat start options

Tomcat start options

OCF_RESKEY_catalina_opts=Catalina options

Catalina options

OCF_RESKEY_catalina_rotate_log=Rotate catalina.out flag

Rotate catalina.out flag

OCF_RESKEY_catalina_rotatetime=Time span of the rotate catalina.out

Time span of the rotate catalina.out

ocf:VIPArip (7)

ocf:VIPArip — Manages a virtual IP address through RIP2

Synopsis

```
OCF_RESKEY_ip=string [OCF_RESKEY_nic=string]  
[OCF_RESKEY_zebra_binary=string] [OCF_RESKEY_ripd_binary=string]  
VIPArip [start | stop | monitor | validate-all | meta-data]
```

Description

Virtual IP Address by RIP2 protocol. This script manages IP alias in different subnet with quagga/ripd. It can add an IP alias, or remove one.

Supported Parameters

OCF_RESKEY_ip=The IP address in different subnet
The IPv4 address in different subnet, for example "192.168.1.1".

OCF_RESKEY_nic=The nic for broadcast the route information
The nic for broadcast the route information. The ripd uses this nic to broadcast the route informaton to others

OCF_RESKEY_zebra_binary=zebra binary
Absolute path to the zebra binary.

OCF_RESKEY_ripd_binary=ripd binary
Absolute path to the ripd binary.

ocf:VirtualDomain (7)

ocf:VirtualDomain — Manages virtual domains through the libvirt virtualization framework

Synopsis

```
OCF_RESKEY_config=string [OCF_RESKEY_hypervisor=string]
[OCF_RESKEY_force_stop=boolean]
[OCF_RESKEY_migration_transport=string]
[OCF_RESKEY_monitor_scripts=string] VirtualDomain [start | stop | status
| monitor | migrate_from | migrate_to | meta-data | validate-all]
```

Description

Resource agent for a virtual domain (a.k.a. domU, virtual machine, virtual environment etc., depending on context) managed by libvirtd.

Supported Parameters

OCF_RESKEY_config=Virtual domain configuration file
Absolute path to the libvirt configuration file, for this virtual domain.

OCF_RESKEY_hypervisor=Hypervisor URI
Hypervisor URI to connect to. See the libvirt documentation for details on supported URI formats. The default is system dependent.

OCF_RESKEY_force_stop=Always force shutdown on stop
Always forcefully shut down ("destroy") the domain on stop. The default behavior is to resort to a forceful shutdown only after a graceful shutdown attempt has failed. You should only set this to true if your virtual domain (or your virtualization backend) does not support graceful shutdown.

`OCF_RESKEY_migration_transport=Remote` hypervisor transport

Transport used to connect to the remote hypervisor while migrating. Please refer to the libvirt documentation for details on transports available. If this parameter is omitted, the resource will use libvirt's default transport to connect to the remote hypervisor.

`OCF_RESKEY_monitor_scripts=`space-separated list of monitor scripts

To additionally monitor services within the virtual domain, add this parameter with a list of scripts to monitor. Note: when monitor scripts are used, the start and migrate_from operations will complete only when all monitor scripts have completed successfully. Be sure to set the timeout of these operations to accommodate this delay.

ocf:vmware (7)

ocf:vmware — Manages VMWare Server 2.0 virtual machines

Synopsis

```
[OCF_RESKEY_vmxpath=string] [OCF_RESKEY_vimshbin=string] vmware  
[start | stop | monitor | meta-data]
```

Description

OCF compliant script to control vmware server 2.0 virtual machines.

Supported Parameters

OCF_RESKEY_vmxpath=VMX file path
VMX configuration file path

OCF_RESKEY_vimshbin=vmware-vim-cmd path
vmware-vim-cmd executable path

ocf:WAS6 (7)

ocf:WAS6 — Manages a WebSphere Application Server 6 instance

Synopsis

[OCF_RESKEY_profile=string] WAS6 [start | stop | status | monitor | validate-all | meta-data | methods]

Description

Resource script for WAS6. It manages a Websphere Application Server (WAS6) as an HA resource.

Supported Parameters

OCF_RESKEY_profile=profile name
The WAS profile name.

ocf:WAS (7)

ocf:WAS — Manages a WebSphere Application Server instance

Synopsis

[OCF_RESKEY_config=string] [OCF_RESKEY_port=integer] WAS [start | stop | status | monitor | validate-all | meta-data | methods]

Description

Resource script for WAS. It manages a Websphere Application Server (WAS) as an HA resource.

Supported Parameters

OCF_RESKEY_config=configuration file
The WAS-configuration file.

OCF_RESKEY_port=port
The WAS-(snoop)-port-number.

ocf:WinPopup (7)

ocf:WinPopup — Sends an SMB notification message to selected hosts

Synopsis

[OCF_RESKEY_hostfile=string] WinPopup [start | stop | status | monitor | validate-all | meta-data]

Description

Resource script for WinPopup. It sends WinPopups message to a sysadmin's workstation whenever a takeover occurs.

Supported Parameters

OCF_RESKEY_hostfile=Host file

The file containing the hosts to send WinPopup messages to.

ocf:Xen (7)

ocf:Xen — Manages Xen unprivileged domains (DomUs)

Synopsis

```
[OCF_RESKEY_xmfile=string] [OCF_RESKEY_name=string]  
[OCF_RESKEY_shutdown_timeout=boolean]  
[OCF_RESKEY_allow_mem_management=boolean]  
[OCF_RESKEY_reserved_Dom0_memory=string]  
[OCF_RESKEY_monitor_scripts=string] Xen [start | stop | migrate_from |  
migrate_to | monitor | meta-data | validate-all]
```

Description

Resource Agent for the Xen Hypervisor. Manages Xen virtual machine instances by mapping cluster resource start and stop, to Xen create and shutdown, respectively. A note on names We will try to extract the name from the config file (the xmfile attribute). If you use a simple assignment statement, then you should be fine. Otherwise, if there's some python acrobacy involved such as dynamically assigning names depending on other variables, and we will try to detect this, then please set the name attribute. You should also do that if there is any chance of a pathological situation where a config file might be missing, for example if it resides on a shared storage. If all fails, we finally fall back to the instance id to preserve backward compatibility. Para-virtualized guests can also be migrated by enabling the meta_attribute allow-migrate.

Supported Parameters

OCF_RESKEY_xmfile=Xen control file

Absolute path to the Xen control file, for this virtual machine.

OCF_RESKEY_name=Xen DomU name

Name of the virtual machine.

OCF_RESKEY_shutdown_timeout=Shutdown escalation timeout

The Xen agent will first try an orderly shutdown using `xm shutdown`. Should this not succeed within this timeout, the agent will escalate to `xm destroy`, forcibly killing the node. If this is not set, it will default to two-third of the stop action timeout. Setting this value to 0 forces an immediate destroy.

OCF_RESKEY_allow_mem_management=Use dynamic memory management

This parameter enables dynamic adjustment of memory for start and stop actions used for Dom0 and the DomUs. The default is to not adjust memory dynamically.

OCF_RESKEY_reserved_Dom0_memory=Minimum Dom0 memory

In case memory management is used, this parameter defines the minimum amount of memory to be reserved for the dom0. The default minimum memory is 512MB.

OCF_RESKEY_monitor_scripts=list of space separated monitor scripts

To additionally monitor services within the unprivileged domain, add this parameter with a list of scripts to monitor. NB: In this case make sure to set the start-delay of the monitor operation to at least the time it takes for the DomU to start all services.

ocf:Xinetd (7)

ocf:Xinetd — Manages an Xinetd service

Synopsis

```
[OCF_RESKEY_service=string] Xinetd [start | stop | restart | status | monitor |  
validate-all | meta-data]
```

Description

Resource script for Xinetd. It starts/stops services managed by xinetd. Note that the xinetd daemon itself must be running: we are not going to start it or stop it ourselves. Important: in case the services managed by the cluster are the only ones enabled, you should specify the -stayalive option for xinetd or it will exit on Heartbeat stop. Alternatively, you may enable some internal service such as echo.

Supported Parameters

OCF_RESKEY_service=service name
The service name managed by xinetd.

部分 V. 附录



设置简单测试资源的示例

本章提供了配置简单资源：IP 地址的基本示例。它演示了两种方法来完成资源配置：使用 Pacemaker GUI 或 `crm` 命令行工具。

对于以下示例，我们假定您已按第 3 章 *用 YaST 进行安装和基本设置*（第 19 页）中所述设置群集，且群集包括至少两个节点。有关如何使用 Pacemaker GUI 和 `crm` 外壳配置群集资源的简介和概述，请参阅以下章节：

- *配置和管理群集资源 (GUI)*（第 53 页）
- *配置和管理群集资源（命令行）*（第 83 页）

A.1 使用 GUI 配置资源

创建样本群集资源并将它迁移到其他服务器可帮助您进行测试，以确保群集运行正确。要配置和迁移的简单资源就是 IP 地址。

过程 A.1 创建 IP 地址群集资源

- 1 按第 5.1.1 节“连接到群集”（第 54 页）中所述，启动 Pacemaker GUI 并登录到群集。
- 2 在左窗格中，切换到 *Resources*（资源）视图；在右窗格中，选择要修改的组并单击 *Edit*（编辑）。下一个窗口将显示为该资源定义的基本组参数以及元属性和原始资源。
- 3 单击 *Primitives*（原始）选项卡并单击 *Add*（添加）。

4 在下一个对话框中，若要将 IP 地址添加为组的子资源，请设置以下参数：

4a 输入唯一的 ID。例如，myIP。

4b 从 *Class*（类）列表中，选择 *ocf* 作为资源代理类。

4c 在 OCF 资源代理的 *Provider*（提供程序）中，选择 *heartbeat*。

4d 从 *Type*（类型）列表中，选择 *IPaddr* 作为资源代理。

4e 单击 *Forward*（前进）。

4f 在实例属性选项卡中，选择 *IP* 项并单击 *编辑*（或双击 *IP* 项）。

4g 输入所需的 IP 地址作为值（例如，10.10.0.1）并单击 *确定*。

4h 添加新的实例属性，并将名称指定为 *nic*，将值指定为 *eth0*，然后单击 *确定*。

名称和值取决于您的硬件配置和安装 High Availability Extension 软件期间选择的媒体配置。

5 根据意愿设置所有参数后，请单击 *确定* 完成此资源的配置。配置对话框将关闭，主窗口将显示修改后的资源。

要使用 Pacemaker GUI 启动资源，请在左侧窗格中选择 *管理*。在右侧窗格中，右键单击资源并选择 *启动*（或从工具栏启动资源）。

要将 IP 地址资源迁移到其他节点 (*saturn*)，请按如下操作：

过程 A.2 将资源迁移到其他节点

- 1** 切换到左侧窗格中的 *管理* 视图，然后右键单击右侧窗格中的 IP 地址资源，并选择 *迁移资源*。
- 2** 在新窗口中，从 *目标节点* 下拉列表中选择 *saturn* 以将选定的资源移到节点 *saturn* 上。
- 3** 如果只想临时迁移资源，请激活 *持续时间* 并输入时间范围，在该时间段内资源应迁移到新的节点。

4 单击确定确认迁移。

A.2 手动配置资源

计算机提供的任何类型的服务都称为资源。资源能够由 High Availability 识别，当资源受 RA（资源代理）控制时，它们就是 LSB 脚本、OCF 脚本或旧式 Heartbeat 1 资源。所有资源都可以使用 `crm` 命令进行配置，或在 CIB（群集信息库）的 `resources` 部分配置为 XML。有关可用资源的概览，请查看第 19 章 *HA OCF Agents*（第 243 页）。

要将 IP 地址 10.10.0.1 作为资源添加到当前配置中，请使用 `crm` 命令：

过程 A.3 创建 IP 地址群集资源

- 1 打开壳层并成为 root。
- 2 输入 `crm configure` 打开内壳。
- 3 创建 IP 地址资源：

```
crm(live)configure# resource
primitive myIP ocf:heartbeat:IPaddr params ip=10.10.0.1
```

注意

使用 High Availability 配置资源时，不应通过 `init` 初始化相同的资源。高可用性负责所有服务的启动或停止操作。

如果配置成功，新资源将显示在 `crm_mon` 中，它在群集的随机节点上启动。

要将资源迁移到其他节点，请执行以下操作：

过程 A.4 将资源迁移到其他节点

- 1 启动壳层并成为 root 用户。
- 2 将资源 `myip` 迁移到节点 `saturn`：

```
crm resource migrate myIP saturn
```


将群集升级为最新产品版本

如果现有群集是基于 SUSE® Linux Enterprise Server 10 的，可以更新群集，使其与 SUSE Linux Enterprise Server 11 或 11 SP1 上的 High Availability Extension 一起运行。

要从 SUSE Linux Enterprise Server 10 迁移到 SUSE Linux Enterprise Server 11 或 11 SP1，所有群集节点都必须处于脱机状态，并将群集作为一个整体来迁移 - 不支持运行在 sls; 10/SUSE Linux Enterprise Server 11 上的混合群集。

B.1 从 SLES 10 升级到 SLEHA 11

为方便起见，SUSE® Linux Enterprise High Availability Extension 包括一个 `hb2openais.sh` 脚本，使用它可在从 Heartbeat 移动到 OpenAIS 群集堆栈的同时转换数据。脚本会分析储存在 `/etc/ha.d/ha.cf` 中的配置，并为 OpenAIS 群集堆栈生成新的配置文件。此外，它调整 CIB 以匹配 OpenAIS 约定、转换 OCFS2 文件系统以及用 cLVM 替换 EVMS。任何 EVMS2 容器都将转换为 cLVM2 卷。对于 CIB 中现有资源所引用的卷组，将创建新的 LVM 资源。

要成功地将群集从 SUSE Linux Enterprise Server 10 SP3 迁移到 SUSE Linux Enterprise Server 11，需要执行以下步骤：

1. 准备 SUSE Linux Enterprise Server 10 SP3 群集（第 342 页）
2. 更新到 SUSE Linux Enterprise 11（第 343 页）
3. 测试转换（第 344 页）

4. 转换数据（第 344 页）

成功完成转换后，可以重新使更新后的群集联机。

注意：更新后还原

更新到 SUSE Linux Enterprise Server 11 之后，不支持恢复回 SUSE Linux Enterprise Server 10。

B.1.1 准备和备份

将群集更新为下一个产品版本并相应地转换数据之前，需要准备当前群集。

过程 B.1 准备 SUSE Linux Enterprise Server 10 SP3 群集

- 1 登录到群集。
- 2 查看检测信号配置文件 `/etc/ha.d/ha.cf` 并检查是否所有通讯媒体都支持多路广播。
- 3 确保以下文件在所有节点上都是相同的：`/etc/ha.d/ha.cf` 和 `/var/lib/heartbeat/crm/cib.xml`。
- 4 通过在每个节点上执行 `rheartbeat stop` 使所有节点都处于脱机状态。
- 5 除了进行更新到最新版本之前推荐的常规系统备份外，另请备份以下文件，因为更新到 SUSE Linux Enterprise Server 11 之后需要它们来运行转换脚本：

- `/var/lib/heartbeat/crm/cib.xml`
- `/var/lib/heartbeat/hostcache`
- `/etc/ha.d/ha.cf`
- `/etc/logd.cf`

- 6 如果有 EVMS2 资源，请将非 LVMEVMS2 卷转换为 SUSE Linux Enterprise Server 10 上的兼容卷。在转换过程中（请参见第 B.1.3 节“数据转换”（第 343 页）），它们会随即转换为 LVM2 卷组。转换后，请务必使用 `vgchange -c y` 将每个卷组都标记为 High Availability 群集的成员。

B.1.2 更新/安装

准备好群集并备份文件后，就可以开始将群集节点更新到下一个产品版本了。除了运行更新外，还可以在群集节点上执行 SUSE Linux Enterprise 11 全新安装。

过程 B.2 更新到 SUSE Linux Enterprise 11

- 1 在所有群集节点上，执行从 SUSE Linux Enterprise Server 10 SP3 到 SUSE Linux Enterprise Server 11 的更新。有关如何更新产品的信息，请参阅 SUSE Linux Enterprise Server 11 部署指南的更新 *SUSE Linux Enterprise* 一章。

或者，也可以在所有群集节点上全新安装 SUSE Linux Enterprise Server 11。

- 2 在所有群集节点上将 SUSE Linux Enterprise High Availability Extension 11 作为外接式附件安装在 SUSE Linux Enterprise Server 上。有关详细信息，请参见第 3.1 节“安装 High Availability Extension”（第 19 页）。

B.1.3 数据转换

安装了 SUSE Linux Enterprise Server 11 和 High Availability Extension 后，就可以开始数据转换了。High Availability Extension 附带的转换脚本已谨慎设置过，但是它无法在全自动模式下处理所有设置。它会其所作更改显示警报，但是需要您进行干预和决策。您需要详细地了解群集—因为由您来校验更改是否有意义。转换脚本位于 `/usr/lib/heartbeat`（如果使用 64 位系统，则位于 `/usr/lib64/heartbeat`）。

注意：执行测试运行

要熟悉转换进程，我们强烈建议您首先测试一下转换（不作任何更改）。可以使用同一测试目录执行重复的测试运行，但是只需要复制一次文件。

过程 B.3 测试转换

- 1 在某个节点上创建测试目录，并将备份文件复制到此测试目录：

```
$ 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 使用以下命令开始测试运行

```
$ /usr/lib/heartbeat/hb2openais.sh -T /tmp/hb2openais-testdir -U
```

如果使用 64 位系统，请使用以下命令：

```
$ /usr/lib64/heartbeat/hb2openais.sh -T /tmp/hb2openais-testdir -U
```

- 3 阅读并校验生成的 openais.conf 和 cib-out.xml 文件：

```
$ cd /tmp/hb2openais-testdir
$ less openais.conf
$ crm_verify -V -x cib-out.xml
```

有关转换阶段的详细信息，请参阅安装的 High Availability Extension 中的 /usr/share/doc/packages/pacemaker/README.hb2openais。

过程 B.4 转换数据

执行测试运行并检查输出后，可以立即开始数据转换。只需在一个节点上运行转换。主群集配置(CIB)会自动复制到其他节点。需要复制的所有其他文件会由转换脚本自动进行复制。

- 1 确保 sshd 运行于 root 有权访问的所有节点上，以便转换脚本成功地将文件复制到其他群集节点。
- 2 确保所有 ocfs2 文件系统都已卸载。
- 3 High Availability Extension 附带了一个默认的 OpenAIS 配置文件。如果要防止在后面的步骤中重写此默认配置，请制作 /etc/ais/openais.conf 配置文件的副本。
- 4 以 root 身份启动转换脚本。如果使用 sudo，请使用 -u 选项指定特权用户：

```
$ /usr/lib/heartbeat/hb2openais.sh -u root
```

基于储存在 `/etc/ha.d/ha.cf` 中的配置，脚本将为 OpenAIS 群集堆栈生成新的配置文件，`/etc/ais/openais.conf`。由于从 Heartbeat 更改到 OpenAIS，它还将分析 CIB 配置并让您了解群集配置是否需要更改。在转换运行的节点上完成所有文件处理，并将文件处理复制到其他节点。

5 按照屏幕指导执行操作。

成功完成转换后，按照第 3.3 节“使群集联机”（第 27 页）中所述启动新的群集堆栈。

升级进程完成后，就不支持还原回 SUSE Linux Enterprise Server 10。

B.1.4 更多信息

有关转换脚本和转换阶段的更多细节，请参阅安装的 High Availability Extension 中的 `/usr/share/doc/packages/pacemaker/README.hb2openais`。

B.2 从 SLEHA 11 升级到 SLEHA 11 SP1

为成功地将现有群集从 SUSE Linux Enterprise High Availability Extension 11 迁移到 11 SP1，可以执行“滚动升级”，即一个接一个地升级节点。随着主群集配置文件从 `/etc/ais/openais.conf` 更改为 SUSE Linux Enterprise High Availability Extension 11 SP1 的 `/etc/corosync/corosync.conf`，脚本会负责进行必要的转换。它们会在更新 `openais` 包时自动执行。

过程 B.5 执行滚动升级

重要：更新软件包

如果要更新作为运行中群集一部分的节点上的任何软件包，请先停止此节点上的群集堆栈，再启动软件更新。要停止群集堆栈，请以 `root` 用户身份登录节点并输入 `rcopenais stop`。

如果 OpenAIS/Corosync 在软件更新过程中正在运行，这可能会导致不可预料的结果，如活动节点被屏蔽。

- 1 以 root 用户身份登录要更新和停止 OpenAIS 的节点：

```
rcopenais stop
```

- 2 检查系统备份是否为最新并且可恢复。
- 3 执行从 SUSE Linux Enterprise Server 11 到 SUSE Linux Enterprise Server 11 SP1 和从 SUSE Linux Enterprise High Availability Extension 11 到 SUSE Linux Enterprise High Availability Extension 11 SP1 的升级。有关如何更新产品的信息，请参阅 SUSE Linux Enterprise Server 11 SP1 *部署指南* 的更新 *SUSE Linux Enterprise* 一章。

- 4 在升级后的节点上重新启动 OpenAIS/Corosync，使此节点重新加入群集：

```
rcopenais start
```

- 5 使下一个节点处于脱机状态，并对此节点重复上述过程。

新功能?

以下部分详述了不同版本之间的软件修改。此摘要指出了基本设置是否已完全重配置、配置文件是否已移至其他位置或者发生的其他重要更改等信息。

C.1 版本 10 SP3 到版本 11

SUSE Linux Enterprise Server 11 的群集堆栈已从 Heartbeat 更改为 OpenAIS。OpenAIS 实行业标准 API，应用程序界面规范 (AIS)，由服务可用性论坛 (Service Availability Forum) 发布。SUSE Linux Enterprise Server 10 的群集资源管理器得以保留但有了显著增强，它已转换为 OpenAIS 且现在称为 Pacemaker。

有关从 SUSE® Linux Enterprise Server 10 SP3 到 SUSE Linux Enterprise Server 11 High Availability 组件所更改内容的更多细节，请参阅以下部分。

C.1.1 新增功能

迁移阈值和故障超时

High Availability Extension 现在新增了迁移阈值和故障超时的概念。可以对资源定义一个故障数量，达到此数量后资源将迁移到新节点。默认情况下，将不再允许节点运行出现故障的资源，直到管理员手动重置资源的故障计数。但也可以通过设置资源的 `failure-timeout` 选项来使资源失效。

资源和操作默认值

现在可以设置资源选项和操作的全局默认值。

支持脱机配置更改

在以原子方式更新配置之前，通常希望预览一系列更改的效果。现在，您可以先创建可使用命令行界面编辑的配置的“阴影”副本，然后再提交它，从而以原子方式更改活动群集配置。

重用规则、选项和操作集

规则、`instance_attributes`、`meta_attributes` 和操作集可定义一次并在多处引用。

对 CIB 中的某些操作使用 XPath 表达式

现在 CIB 接受基于 XPath 的 `create`、`modify`、`delete` 操作。有关更多信息，请参见 `cibadmin` 帮助文本。

多维排列和排序约束

为创建一个排列资源集，以前可以定义一个资源组（无法总是准确地表达设计意图）或将每个关系定义为单独的约束，导致约束随着资源和组合的数量增长而激增。现在还可以使用排列约束的另一种形式，即定义

`resource_sets`。

从非群集的服务器连接到 CIB

如果服务器上安装了 `Pacemaker`，则即使服务器本身不是群集的一部分，也可以连接到群集。

在已知时间触发重现操作

默认情况下，重现操作是根据资源启动的时间来计划的，但这并不总令人满意。要指定操作应根据的日期/时间，请设置操作的间隔-起始时间。群集使用此时间计算正确的启动-延迟，这样操作将在起始时间 + (间隔 * N) 时发生。

C.1.2 变更功能

资源和群集选项的命名约定

现在所有资源和群集选项都使用连字符 (-) 代替下划线 (_)。例如，`master_max` 元选项已重命名为 `master-max`。

重命名 `master_slave` 资源

`master_slave` 资源已重命名为 `master`。主资源是一种特殊类型的克隆，可按两种模式之一运行。

属性的容器标记

`attributes` 容器标记已删除。

先决条件的操作字段

`pre-req` 操作字段已重命名为 `requires`。

操作间隔

所有操作都必须有间隔。对于启动/停止操作，间隔必须设置为 0。

排列和排序约束的属性

为了清晰起见，已重命名排列和排序约束的属性。

因故障而迁移的群集选项

`resource-failure-stickiness` 群集选项已替换为 `migration-threshold` 群集选项。另请参见迁移阈值和故障超时（第 347 页）。

命令行工具的自变量

已使命令行工具的自变量保持一致。另请参阅资源和群集选项的命名约定（第 348 页）。

验证和分析 XML

群集配置是用 XML 编写的。现在，一种更强大的 RELAX-NG 纲要已取代文档类型定义(DTD)，用于定义结构和内容的模式。`libxml2` 用作分析器。

id 字段

`id` 字段现在是具有以下限制的 XML ID：

- ID 不能包含冒号。
- ID 不能以数字开始。
- ID 必须是全局唯一的（不只是对标记唯一）。

参考其他对象

某些字段（如引用资源的限制中的字段）是 `IDREF`。这意味着它们必须引用现有资源或对象才能使配置有效。无法删除在别处作为参考的对象。

C.1.3 删除功能

设置资源元选项

不再能将资源元选项设置为顶级属性。改为使用元属性。另请参见 `crm_resource(8)`（第 217 页）。

设置全局默认值

不再从 `crm_config` 读取资源和操作默认值。

C.2 版本 11 到版本 11 SP1

群集配置文件

主群集配置文件已从 `/etc/ais/openais.conf` 更改为 `/etc/corosync/corosync.conf`。这两个文件很相似。从 SUSE Linux Enterprise High Availability Extension 11 升级到 SP1 时，脚本会负责处理这些文件之间的小差异。有关 OpenAIS 和 Corosync 之间关系的更多信息，请参见。[\[http://www.corosync.org/doku.php?id=faq:why\]](http://www.corosync.org/doku.php?id=faq:why)

滚动升级

为了在最短停机时间内完成现有群集的迁移，SUSE Linux Enterprise High Availability Extension 允许执行从 SUSE Linux Enterprise High Availability Extension 11 到 11 SP1 的“滚动升级”。一个接一个地升级节点时，群集仍处于联机状态。

自动群集部署

为了方便群集部署，AutoYaST 允许克隆现有节点。AutoYaST 是使用包含安装和配置数据的 AutoYaST 配置文件自动安装一个或多个 SUSE Linux Enterprise 系统而无需用户干预的系统。此配置文件将告知 AutoYaST 要安装的内容以及如何配置安装好的系统，以最终获得一个即用型的系统。此配置文件可用于以不同方式进行大批量部署。

配置文件的传送

SUSE Linux Enterprise High Availability Extension 附带 `Csync2`，后者是用于在群集中所有节点之间复制配置文件的工具。它能处理任意数量的主机，还可以只在特定主机子组间同步文件。使用 YaST 配置应通过 `Csync2` 同步的主机名和文件。

群集管理的 Web 界面

High Availability Extension 现在还包含基于 Web 的用户界面 HA Web Konsole 用于执行管理任务。它还可用于从非 Linux 计算机监视和管理 Linux 群集。如果系统未提供或不支持图形用户界面，它还是理想的解决方案。

资源配置模板

使用命令行界面创建和配置资源时，现在可以从各种资源模板中进行选择，以更快、更方便地进行配置。

根据负载放置资源

通过定义特定节点提供的容量、特定资源需要的容量，以及通过选择群集中的若干放置策略之一，可以根据资源负载影响来布置资源，以避免降低群集性能。

群集感知的主动/主动 RAID1

现在可以使用 cmirrord 从两个独立的 SAN 创建能迅速从灾难中恢复的储存配置。

只读 GFS2 支持

为了便于从 GFS2 迁移到 OCFS2，可以采用只读模式装入 GFS2 文件系统，以将数据复制到 OCFS2 文件系统。SUSE Linux Enterprise High Availability Extension 完全支持 OCFS2。

OCFS2 的 SCTP 支持

如果配置了冗余环，OCFS2 和 DLM 可通过独立于网络设备绑定的 SCIP 自动使用冗余通讯路径。

储存保护

为提供附加安全层来保护储存的数据免受损坏，可以使用 IO 屏蔽（使用 external/sbd 屏蔽设备）和 sfex 资源代理的组合来确保对储存内容的排它访问。

Samba 群集

High Availability Extension 现在支持普通数据库的群集实现：CTDB。这样您就可以配置群集 Samba 服务器 - 也为异构环境提供 High Availability 解决方案。

用于 IP 负载平衡的 YaST 模块

此新模块允许使用图形用户界面配置基于内核的负载平衡。它是用于管理 Linux Virtual Server 和监视真实服务器的用户空间守护程序 ldirectord 的前端。

GNU 许可证



此附件包含 GNU 通用公共许可证和 GNU 自由文档许可证。

GNU General Public License

Version 2, June 1991

Copyright (C) 1989, 1991 Free Software Foundation, Inc. 59 Temple Place - Suite 330, Boston, MA 02111-1307, USA

Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

Preamble

The licenses for most software are designed to take away your freedom to share and change it. By contrast, the GNU General Public License is intended to guarantee your freedom to share and change free software--to make sure the software is free for all its users. This General Public License applies to most of the Free Software Foundation's software and to any other program whose authors commit to using it. (Some other Free Software Foundation software is covered by the GNU Library General Public License instead.) You can apply it to your programs, too.

When we speak of free software, we are referring to freedom, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for this service if you wish), that you receive source code or can get it if you want it, that you can change the software or use pieces of it in new free programs; and that you know you can do these things.

To protect your rights, we need to make restrictions that forbid anyone to deny you these rights or to ask you to surrender the rights. These restrictions translate to certain responsibilities for you if you distribute copies of the software, or if you modify it.

For example, if you distribute copies of such a program, whether gratis or for a fee, you must give the recipients all the rights that you have. You must make sure that they, too, receive or can get the source code. And you must show them these terms so they know their rights.

We protect your rights with two steps: (1) copyright the software, and (2) offer you this license which gives you legal permission to copy, distribute and/or modify the software.

Also, for each author's protection and ours, we want to make certain that everyone understands that there is no warranty for this free software. If the software is modified by someone else and passed on, we want its recipients to know that what they have is not the original, so that any problems introduced by others will not reflect on the original authors' reputations.

Finally, any free program is threatened constantly by software patents. We wish to avoid the danger that redistributors of a free program will individually obtain patent licenses, in effect making the program proprietary. To prevent this, we have made it clear that any patent must be licensed for everyone's free use or not licensed at all.

The precise terms and conditions for copying, distribution and modification follow.

GNU GENERAL PUBLIC LICENSE TERMS AND CONDITIONS FOR COPYING, DISTRIBUTION AND MODIFICATION

0. This License applies to any program or other work which contains a notice placed by the copyright holder saying it may be distributed under the terms of this General Public License. The “Program”, below, refers to any such program or work, and a “work based on the Program” means either the Program or any derivative work under copyright law: that is to say, a work containing the Program or a portion of it, either verbatim or with modifications and/or translated into another language. (Hereinafter, translation is included without limitation in the term “modification”.) Each licensee is addressed as “you”.

Activities other than copying, distribution and modification are not covered by this License; they are outside its scope. The act of running the Program is not restricted, and the output from the Program is covered only if its contents constitute a work based on the Program (independent of having been made by running the Program). Whether that is true depends on what the Program does.

1. You may copy and distribute verbatim copies of the Program’s source code as you receive it, in any medium, provided that you conspicuously and appropriately publish on each copy an appropriate copyright notice and disclaimer of warranty; keep intact all the notices that refer to this License and to the absence of any warranty; and give any other recipients of the Program a copy of this License along with the Program.

You may charge a fee for the physical act of transferring a copy, and you may at your option offer warranty protection in exchange for a fee.

2. You may modify your copy or copies of the Program or any portion of it, thus forming a work based on the Program, and copy and distribute such modifications or work under the terms of Section 1 above, provided that you also meet all of these conditions:

a) You must cause the modified files to carry prominent notices stating that you changed the files and the date of any change.

b) You must cause any work that you distribute or publish, that in whole or in part contains or is derived from the Program or any part thereof, to be licensed as a whole at no charge to all third parties under the terms of this License.

c) If the modified program normally reads commands interactively when run, you must cause it, when started running for such interactive use in the most ordinary way, to print or display an announcement including an appropriate copyright notice and a notice that there is no warranty (or else, saying that you provide a warranty) and that users may redistribute the program under these conditions, and telling the user how to view a copy of this License. (Exception: if the Program itself is interactive but does not normally print such an announcement, your work based on the Program is not required to print an announcement.)

These requirements apply to the modified work as a whole. If identifiable sections of that work are not derived from the Program, and can be reasonably considered independent and separate works in themselves, then this License, and its terms, do not apply to those sections when you distribute them as separate works. But when you distribute the same sections as part of a whole which is a work based on the Program, the distribution of the whole must be on the terms of this License, whose permissions for other licensees extend to the entire whole, and thus to each and every part regardless of who wrote it.

Thus, it is not the intent of this section to claim rights or contest your rights to work written entirely by you; rather, the intent is to exercise the right to control the distribution of derivative or collective works based on the Program.

In addition, mere aggregation of another work not based on the Program with the Program (or with a work based on the Program) on a volume of a storage or distribution medium does not bring the other work under the scope of this License.

3. You may copy and distribute the Program (or a work based on it, under Section 2) in object code or executable form under the terms of Sections 1 and 2 above provided that you also do one of the following:

a) Accompany it with the complete corresponding machine-readable source code, which must be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange; or,

b) Accompany it with a written offer, valid for at least three years, to give any third party, for a charge no more than your cost of physically performing source distribution, a complete machine-readable copy of the corresponding source code, to be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange; or,

c) Accompany it with the information you received as to the offer to distribute corresponding source code. (This alternative is allowed only for noncommercial distribution and only if you received the program in object code or executable form with such an offer, in accord with Subsection b above.)

The source code for a work means the preferred form of the work for making modifications to it. For an executable work, complete source code means all the source code for all modules it contains, plus any associated interface definition files, plus the scripts used to control compilation and installation of the executable. However, as a special exception, the source code distributed need not include anything that is normally distributed (in either source or binary form) with the major components (compiler, kernel, and so on) of the operating system on which the executable runs, unless that component itself accompanies the executable.

If distribution of executable or object code is made by offering access to copy from a designated place, then offering equivalent access to copy the source code from the same place counts as distribution of the source code, even though third parties are not compelled to copy the source along with the object code.

4. You may not copy, modify, sublicense, or distribute the Program except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense or distribute the Program 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.

5. You are not required to accept this License, since you have not signed it. However, nothing else grants you permission to modify or distribute the Program or its derivative works. These actions are prohibited by law if you do not accept this License. Therefore, by modifying or distributing the Program (or any work based on the Program), you indicate your acceptance of this License to do so, and all its terms and conditions for copying, distributing or modifying the Program or works based on it.

6. Each time you redistribute the Program (or any work based on the Program), the recipient automatically receives a license from the original licensor to copy, distribute or modify the Program subject to these terms and conditions. You may not impose any further restrictions on the recipients' exercise of the rights granted herein. You are not responsible for enforcing compliance by third parties to this License.

7. If, as a consequence of a court judgment or allegation of patent infringement or for any other reason (not limited to patent issues), conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot distribute so as to satisfy simultaneously your obligations under this License and any other pertinent obligations, then as a consequence you may not distribute the Program at all. For example, if a patent license would not permit royalty-free redistribution of the Program by all those who receive copies directly or indirectly through you, then the only way you could satisfy both it and this License would be to refrain entirely from distribution of the Program.

If any portion of this section is held invalid or unenforceable under any particular circumstance, the balance of the section is intended to apply and the section as a whole is intended to apply in other circumstances.

It is not the purpose of this section to induce you to infringe any patents or other property right claims or to contest validity of any such claims; this section has the sole purpose of protecting the integrity of the free software distribution system, which is implemented by public license practices. Many people have made generous contributions to the wide range of software distributed through that system in reliance on consistent application of that system; it is up to the author/donor to decide if he or she is willing to distribute software through any other system and a licensee cannot impose that choice.

This section is intended to make thoroughly clear what is believed to be a consequence of the rest of this License.

8. If the distribution and/or use of the Program is restricted in certain countries either by patents or by copyrighted interfaces, the original copyright holder who places the Program under this License may add an explicit geographical distribution limitation excluding those countries, so that distribution is permitted only in or among countries not thus excluded. In such case, this License incorporates the limitation as if written in the body of this License.

9. The Free Software Foundation may publish revised and/or new versions of the General Public 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.

Each version is given a distinguishing version number. If the Program specifies a version number of this License which applies to it and “any later version”, you have the option of following the terms and conditions either of that version or of any later version published by the Free Software Foundation. If the Program does not specify a version number of this License, you may choose any version ever published by the Free Software Foundation.

10. If you wish to incorporate parts of the Program into other free programs whose distribution conditions are different, write to the author to ask for permission. For software which is copyrighted by the Free Software Foundation, write to the Free Software Foundation; we sometimes make exceptions for this. Our decision will be guided by the two goals of preserving the free status of all derivatives of our free software and of promoting the sharing and reuse of software generally.

NO WARRANTY

11. BECAUSE THE PROGRAM IS LICENSED FREE OF CHARGE, THERE IS NO WARRANTY FOR THE PROGRAM, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE PROGRAM “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE PROGRAM IS WITH YOU. SHOULD THE PROGRAM PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

12. IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MAY MODIFY AND/OR REDISTRIBUTE THE PROGRAM AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE PROGRAM (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE PROGRAM TO OPERATE WITH ANY OTHER PROGRAMS), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

END OF TERMS AND CONDITIONS

How to Apply These Terms to Your New Programs

If you develop a new program, and you want it to be of the greatest possible use to the public, the best way to achieve this is to make it free software which everyone can redistribute and change under these terms.

To do so, attach the following notices to the program. It is safest to attach them to the start of each source file to most effectively convey the exclusion of warranty; and each file should have at least the “copyright” line and a pointer to where the full notice is found.

one line to give the program's name and an idea of what it does. Copyright (C) yyyy name of author

This program is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with this program; if not, write to the Free Software Foundation, Inc., 59 Temple Place - Suite 330, Boston, MA 02111-1307, USA.

Also add information on how to contact you by electronic and paper mail.

If the program is interactive, make it output a short notice like this when it starts in an interactive mode:

```
Gnomovision version 69, Copyright (C) year name of author
Gnomovision comes with ABSOLUTELY NO WARRANTY; for details
type `show w'. This is free software, and you are welcome
to redistribute it under certain conditions; type `show c'
for details.
```

The hypothetical commands 'show w' and 'show c' should show the appropriate parts of the General Public License. Of course, the commands you use may be called something other than 'show w' and 'show c'; they could even be mouse-clicks or menu items--whatever suits your program.

You should also get your employer (if you work as a programmer) or your school, if any, to sign a "copyright disclaimer" for the program, if necessary. Here is a sample; alter the names:

```
Yoyodyne, Inc., hereby disclaims all copyright
interest in the program `Gnomovision'
(which makes passes at compilers) written
by James Hacker.
```

```
signature of Ty Coon, 1 April 1989
Ty Coon, President of Vice
```

This General Public License does not permit incorporating your program into proprietary programs. If your program is a subroutine library, you may consider it more useful to permit linking proprietary applications with the library. If this is what you want to do, use the GNU Lesser General Public License [<http://www.fsf.org/licenses/lgpl.html>] instead of this License.

GNU Free Documentation License

Version 1.2, November 2002

Copyright (C) 2000,2001,2002 Free Software Foundation, Inc. 59 Temple Place, Suite 330, Boston, MA 02111-1307 USA

Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

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.

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.

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.

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.

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 it was based on. 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.

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”.

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.

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.

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.

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.

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

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

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 only as published by the Free Software Foundation; with the Invariant Section being this copyright notice and license. 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.

术语

主动/主动、主动/被动

一个有关服务在节点上如何运行的概念。主动-被动方案表示一个或多个服务正在主动节点上运行，而被动节点则等待主动节点出现故障。主动-主动方案表示每个节点既是主动节点同时也是被动节点。

群集

高性能群集是一组为实现更快结果而共享应用负载的计算机（实际或虚拟）。高可用性群集主要用于确保服务的最大可用性。

群集信息库 (CIB)

整个群集配置和状态（节点成员资格、资源、约束等等）的表示。它用XML编写，位于内存中。主 CIB 保留并在指定协调器 (DC)（第 361 页）上进行维护，并复制到其他节点。

群集分区

当一个或多个节点与群集的剩余节点之间的通讯失败时，即会发生群集分区。群集分区的节点仍是活动的且能够相互通讯，但它们无法感知不能与其通讯的节点。由于无法确认其他分区的丢失，所以开发了一种节点分裂方案（另请参见节点分裂（第 363 页））。

群集资源管理器 (CRM)

负责协调所有非本地交互的主要管理实体。群集的每个节点都有自己的 CRM，但在 DC 上运行的 CRM 会将决策转发给其他非本地 CRM 并处理其输入。CRM 会与多个组件交互：其自己的节点和其他节点上的本地资源管理器、非本地 CRM、管理命令、屏蔽功能以及成员资格层。

一致群集成员资格 (CCM)

CCM 确定组成群集的节点并在群集中共享此信息。任何节点或仲裁人数的新增和丢失都由 CCM 提供。群集的每个节点上都运行 CCM 模块。

指定协调器 (DC)

“主”节点。在此节点上保存着 CIB 的主副本。所有其他节点都从当前 DC 获取他们的配置和资源分配信息。DC 是在成员资格更改后从群集的所有节点中选出的。

分布式锁管理器 (DLM)

DLM 协调群集文件系统的磁盘访问和管理文件锁定以提高性能和可用性。

分布式复制块设备 (drbd)

DRBD 是为构建高可用性群集而设计的块设备。整个块设备通过专用网络镜像，且视作网络 RAID-1。

故障转移

指资源或节点在某台服务器上出现故障、受影响的资源在另一个节点上启动的情况。

屏障

描述了防止非群集成员访问共享资源的概念。通过终止（关闭）“有故障”的节点以防止其引起问题、使资源远离状态不确定的节点或多种其他方式均可以达到此目的。此外，屏蔽分为节点屏蔽和资源屏蔽。

Heartbeat 资源代理

Heartbeat 第 1 版中广泛地使用了 Heartbeat 资源代理。第 2 版中已废弃对它们的使用，但仍然支持。Heartbeat 资源代理可以执行启动、停止和状态操作，它位于 `/etc/ha.d/resource.d` 或 `/etc/init.d` 下。有关 Heartbeat 资源代理的更多信息，请参阅 <http://www.linux-ha.org/HeartbeatResourceAgent>（另请参见 OCF 资源代理（第 363 页））。

本地资源管理器 (LRM)

本地资源管理器 (LRM) 负责对资源执行操作。它使用资源代理脚本执行工作。LRM 是“哑”的，它自己无法了解任何策略。它需要 DC 告诉它做什么。

LSB 资源代理

LSB 资源代理是标准 LSB init 脚本。LSB init 脚本不仅用于高可用性环境中。任何兼容 LSB 的 Linux 系统使用 LSB init 脚本控制服务。任何 LSB 资源代理支持 `start`、`stop`、`restart`、`status` 和 `force-reload` 选项，并可能可选地提供 `try-restart` 和 `reload`。LSB 资源代理位于 `/etc/init.d`。在 <http://www.linux-ha.org/LSBResourceAgent> 和 http://www.linux-foundation.org/spec/refspecs/LSB_3.0.0/LSB-Core-generic/LSB-Core-generic/iniscriptact.html 可了解有关 LSB 资源代理和实际规范的更多信息。（另请参见 OCF 资源代理（第 363 页）和 Heartbeat 资源代理（第 362 页））。

节点

是群集成员并对用户不可见的任何计算机（实际或虚拟）。

pingd

ping 守护程序。它使用 ICMP ping 持续联系一个或多个群集外的服务器。

策略引擎 (PE)

策略引擎计算要实现 CIB 中的策略更改而需要执行的操作。此信息随后传递到事务引擎，它在群集设置中依次实施策略更改。PE 始终在 DC 上运行。

OCF 资源代理

OCF 资源代理类似于 LSB 资源代理 (init 脚本)。任何 OCF 资源代理必须支持 start、stop 和 status (有时候称为 monitor) 选项。另外，它支持以 XML 返回资源代理类型描述的元数据选项。它可能支持更多选项，但不是强制的。OCF 资源代理位于 /usr/lib/ocf/resource.d/ 提供程序在 <http://www.linux-ha.org/OCFResourceAgent> 和 <http://www.opencf.org/cgi-bin/viewcvs.cgi/specs/ra/resource-agent-api.txt?rev=HEAD> 上可以找到有关 OCF 资源代理的更多信息及规范草稿 (另请参见 Heartbeat 资源代理 (第 362 页))。

仲裁人数

在群集中，如果群集分区具有多数节点 (或投票)，则它定义为具有仲裁人数 (是“具有仲裁人数的”)。仲裁人数准确地区分了一个分区。它是算法的组成部分，用于防止多个断开的分区或节点继续运行而导致数据和服务损坏 (节点分裂)。仲裁人数是屏障的先决条件，而屏障随后确保仲裁人数确实是唯一的。

资源

Heartbeat 已知的任何类型的服务或应用程序。例如，IP 地址、文件系统或数据库。

资源代理 (RA)

资源代理 (RA) 是一种脚本，作为代理管理资源。有三种不同的资源代理：OCF (开放群集框架) 资源代理、LSB 资源代理 (标准 LSB init 脚本) 和 Heartbeat 资源代理 (Heartbeat v1 资源)。

单一故障点 (SPOF)

单一故障点 (SPOF) 是群集的任何如下的组件：如果它出现故障，则会触发整个群集的故障。

节点分裂

一种将群集节点分为两个或多个互不了解的组的方案 (通过软件或硬件故障)。STONITH 防止节点分裂情况对整个群集产生不利影响。也称为“分区的群集”方案。

术语“节点分裂”还用于 DRBD 中，但在 DRBD 中，它表示两个节点包含不同的数据。

STONITH

“Shoot the other node in the head（关闭其他节点）”的首字母缩写，它关闭功能不正常的节点以防止其在群集中造成故障。

事务引擎 (TE)

事务引擎 (TE) 从 PE 取得策略指令并执行它们。TE 始终在 DC 上运行。它从此处指示其他节点上的本地资源管理器执行哪些操作。