Pacemaker 1.1

从头开始搭建集群

在Fedora上面创建主/主和主备集群

Andrew Beekhof

Pacemaker 1.1 从头开始搭建集群
在Fedora上面创建主/主和主备集群
版 5

作者	Andrew Beekhof	*andrew@beekhof.net*
译者	Raoul Scarazzini	*rasca@miamammausalinux.org*
译者	Dan Frîncu	*df.cluster@gmail.com*

本文档的主要目的是提供一站式指南，教您如何使用Pacemaker创建一个主/备模式的集群并把它转换到主/主模式。

示例集群会使用以下软件:

1. Fedora 13 as the host operating system
2. Corosync作为通信层和提供关系管理服务
3. Pacemaker来实现资源管理
4. DRBD 作为一个经济的共享存储方案
5. GFS2 作为集群文件系统（主/主模式中）
6. crm shell 来显示并修改配置文件

虽然给出了图形化安装Fedora的过程，并且有很多截图，但是本文的主要是靠命令来操作，包括为什么要运行这个命令和这些操作产生的结果。(译者注:本文中基本是crm shell来操作的，这里应该是老版本文档的遗留)

[1] An explanation of CC-BY-SA is available at *http://creativecommons.org/licenses/by-sa/3.0/*

目录

插图清单

前言

目录

1. 文档约定

本手册使用几个约定来突出某些用词和短语以及信息的某些片段。

在 PDF 版本以及纸版中，本手册使用在 *Liberation ##*[1]套件中选出的字体。如果您在您的系统中安装了 Liberation 字体套件，它还可用于 HTML 版本。如果没有安装，则会显示可替换的类似字体。请注意：红帽企业 Linux 5 以及其后的版本默认包含 Liberation 字体套件。

1.1. 排版约定

我们使用四种排版约定突出特定用词和短语。这些约定及其使用环境如下。

单行粗体

用来突出系统输入，其中包括 shell 命令、文件名以及路径。还可用来突出按键以及组合键。例如：

> 要看到文件您当前工作目录中文件 my_next_bestselling_novel 的内容，请在 shell 提示符后输入 cat my_next_bestselling_novel 命令并按 Enter 键执行该命令。

以上内容包括一个文件名，一个 shell 命令以及一个按键，它们都以固定粗体形式出现，且全部与上下文有所区别。

按键组合与单独按键之间的区别是按键组合是使用加号将各个按键连在一起。例如：

> 按 Enter 执行该命令。
>
> 按 Ctrl+Alt+F2 切换到虚拟终端。

第一个示例突出的是要按的特定按键。第二个示例突出了按键组合：一组要同时按下的三个按键。

如果讨论的是源码、等级名称、方法、功能、变量名称以及在段落中提到的返回的数值，那么都会以上述形式出现，即固定粗体。例如：

> 与文件相关的等级包括用于文件系统的 filesystem、用于文件的 file 以及用于目录的 dir。每个等级都有其自身相关的权限。

比例粗体

这是指在系统中遇到的文字或者短语，其中包括应用程序名称、对话框文本、标记的按钮、复选框以及单选按钮标签、菜单标题以及子菜单标题。例如：

[1] https://fedorahosted.org/liberation-fonts/

在主菜单条中选择「系统」→「首选项」→「鼠标」启动 鼠标首选项。在「按钮」标签中点击「惯用左手鼠标」 复选框并点击 关闭切换到主鼠标按钮从左向右（让鼠标适合左手使用）。

要在 gedit 文件中插入特殊字符，请在主菜单栏中选择「应用程序」→「附件」→「字符映射表」。接下来选择从 Character Map 菜单中选择Search →「查找.....」，在「搜索」字段输入字符名称并点击「下一个」按钮。此时会在「字符映射表」中突出您搜索的字符。双击突出的字符将其放在「要复制的文本」字段中，然后点击「复制」按钮。现在返回您的文档，并选择 gedit 菜单中的「编辑」→「粘贴」。

以上文本包括应用程序名称、系统范围菜单名称及项目、应用程序特定菜单名称以及按钮和 GUI 界面中的文本，所有都以比例粗体出现并与上下文区别。

或者

无论固定粗体或者比例粗体，附加的斜体表示是可替换或者变量文本。斜体表示那些不直接输入的文本或者那些根据环境改变的文本。例如：

要使用 ssh 连接到远程机器，请在 shell 提示符后输入 ssh ***username@domain.name***。如果远程机器是 example.com 且您在该其机器中的用户名为 john，请输入 ssh john@example.com。

mount -o remount ***file-system*** 命令会重新挂载命名的文件系统。例如：要重新挂载 /home 文件系统，则命令为 mount -o remount /home。

要查看目前安装的软件包版本，请使用 rpm -q ***package*** 命令。它会返回以下结果：***package-version-release***。

请注意上述使用黑斜体的文字 -- username、domain.name、file-system、package、version 和 release。每个字都是一个站位符，可用作您执行命令时输入的文本，也可作为该系统显示的文本。

不考虑工作中显示标题的标准用法，斜体表示第一次使用某个新且重要的用语。例如：

Publican 是一个 *DocBook* 发布系统。

1.2. 抬升式引用约定

终端输出和源代码列表要与周围文本明显分开。

将发送到终端的输出设定为 Mono-spaced Roman 并显示为：

```
books        Desktop   documentation  drafts  mss    photos   stuff  svn
books_tests  Desktopl  downloads      images  notes  scripts  svgs
```

源码列表也设为 Mono-spaced Roman，但添加下面突出的语法：

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

import javax.naming.InitialContext;

public class ExClient
{
   public static void main(String args[])
       throws Exception
   {
      InitialContext iniCtx = new InitialContext();
```

```
        Object          ref     = iniCtx.lookup("EchoBean");
        EchoHome        home    = (EchoHome) ref;
        Echo            echo    = home.create();

        System.out.println("Created Echo");

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

1.3. 备注及警告

最后，我们使用三种视觉形式来突出那些可能被忽视的信息。

注意

备注是对手头任务的提示、捷径或者备选的解决方法。忽略提示不会造成负面后果，但您可能会错过一个更省事的诀窍。

重要

重要框中的内容是那些容易错过的事情：配置更改只可用于当前会话，或者在应用更新前要重启的服务。忽略‘重要’框中的内容不会造成数据丢失但可能会让您抓狂。

警告

警告是不应被忽略的。忽略警告信息很可能导致数据丢失。

2. We Need Feedback!

If you find a typographical error in this manual, or if you have thought of a way to make this manual better, we would love to hear from you! Please submit a report in Bugzilla[2] against the product Pacemaker.

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

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

[2] http://bugs.clusterlabs.org

Read-Me-First

目录

1.1. 本文范围

Computer clusters can be used to provide highly available services or resources. The redundancy of multiple machines is used to guard against failures of many types.

This document will walk through the installation and setup of simple clusters using the Fedora distribution, version 14.

The clusters described here will use Pacemaker and Corosync to provide resource management and messaging. Required packages and modifications to their configuration files are described along with the use of the Pacemaker command line tool for generating the XML used for cluster control.

Pacemaker is a central component and provides the resource management required in these systems. This management includes detecting and recovering from the failure of various nodes, resources and services under its control.

When more in depth information is required and for real world usage, please refer to the *Pacemaker Explained*[1] manual.

1.2. 什么是Pacemaker?

Pacemaker is a cluster resource manager. It achieves maximum availability for your cluster services (aka. resources) by detecting and recovering from node and resource-level failures by making use of the messaging and membership capabilities provided by your preferred cluster infrastructure (either Corosync or Heartbeat).

Pacemaker's key features include:

- 监测并恢复节点和服务级别的故障
- 存储无关，并不需要共享存储
- 资源无关，任何能用脚本控制的资源都可以作为服务
- Supports STONITH for ensuring data integrity
- 支持大型或者小型的集群

[1] http://www.clusterlabs.org/doc/

- Supports both quorate and resource driven clusters
- Supports practically any redundancy configuration
- 自动同步各个节点的配置文件
- 可以设定集群范围内的ordering, colocation and anti-colocation
- Support for advanced service types
  - Clones:为那些要在多个节点运行的服务所准备的
  - Multi-state:为那些有多种模式的服务准备的。(比如.主从，主备)
- 统一的，可脚本控制的cluster shell

1.3. Pacemaker 架构

在最高一个层次，集群由三个部分组成:

- Non-cluster aware components (illustrated in green). These pieces include the resources themselves, scripts that start, stop and monitor them, and also a local daemon that masks the differences between the different standards these scripts implement.
- Resource management Pacemaker provides the brain (illustrated in blue) that processes and reacts to events regarding the cluster. These events include nodes joining or leaving the cluster; resource events caused by failures, maintenance, scheduled activities; and other administrative actions. Pacemaker will compute the ideal state of the cluster and plot a path to achieve it after any of these events. This may include moving resources, stopping nodes and even forcing them offline with remote power switches.
- Low level infrastructure Corosync provides reliable messaging, membership and quorum information about the cluster (illustrated in red).

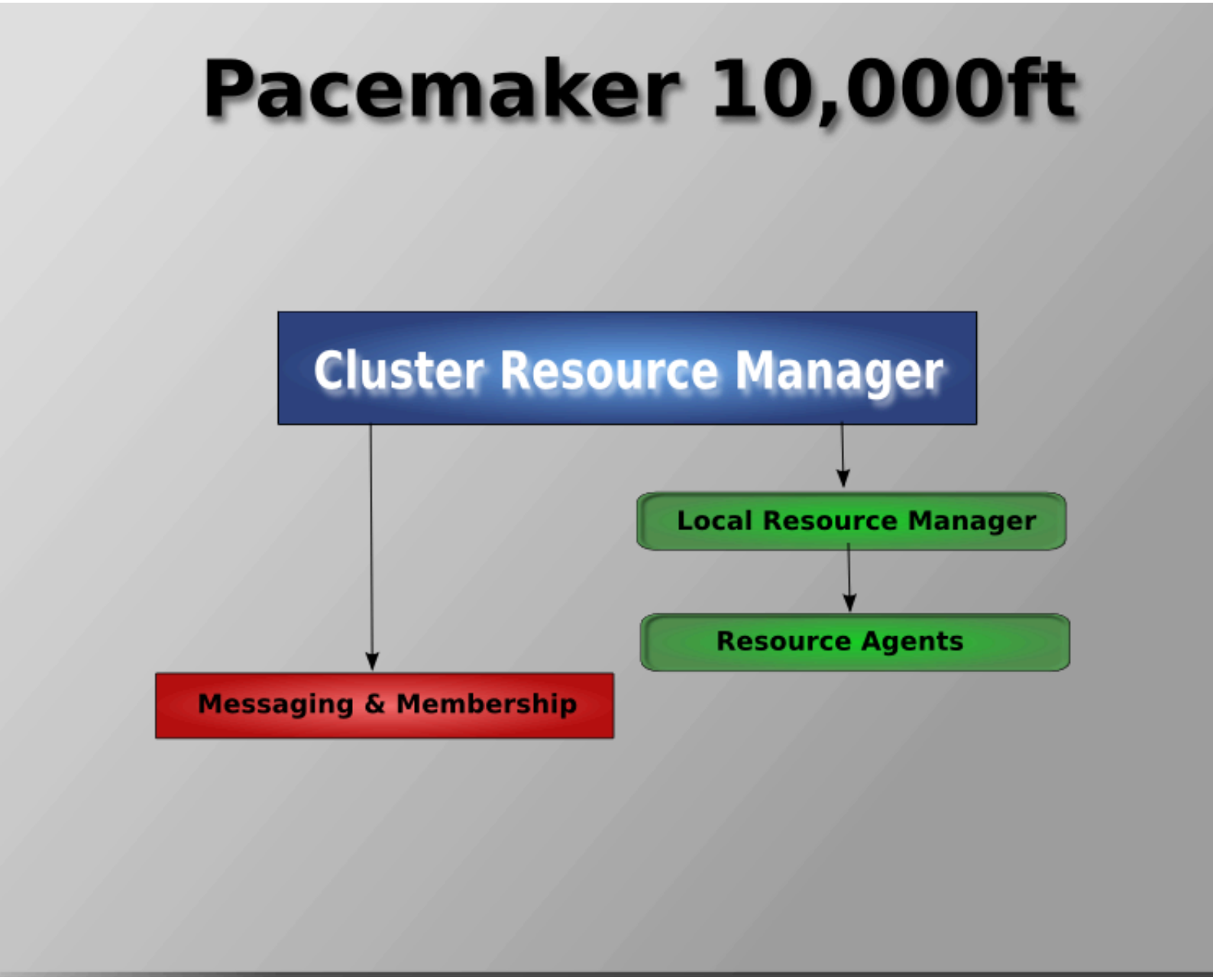


图 1.1. 概念层次总览

When combined with Corosync, Pacemaker also supports popular open source cluster filesystems. [2]

Due to recent standardization within the cluster filesystem community, they make use of a common distributed lock manager which makes use of Corosync for its messaging capabilities and Pacemaker for its membership (which nodes are up/down) and fencing services.

[2] Even though Pacemaker also supports Heartbeat, the filesystems need to use the stack for messaging and membership and Corosync seems to be what they' re standardizing on. Technically it would be possible for them to support Heartbeat as well, however there seems little interest in this.

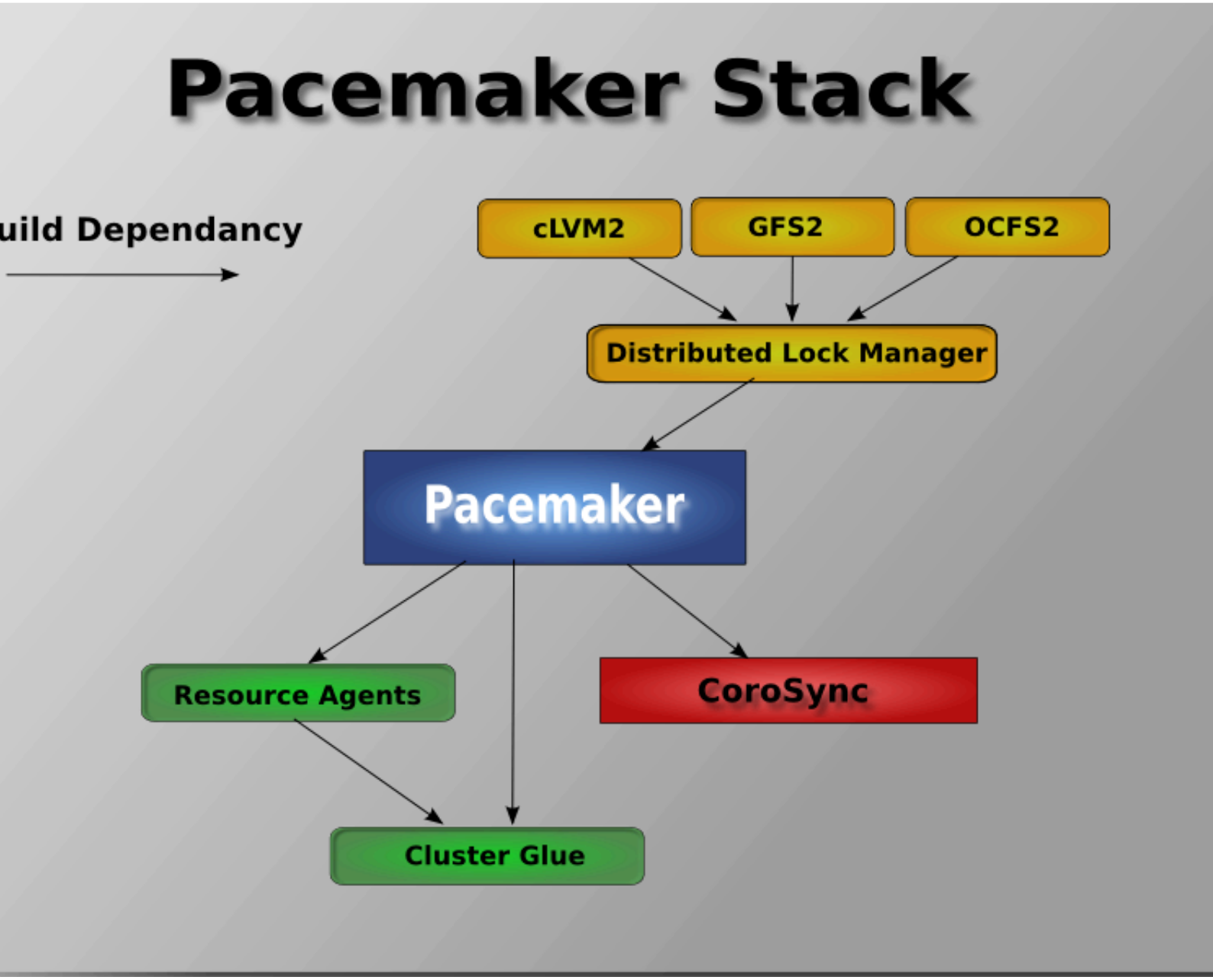


图 1.2. Pacemaker 层次

1.3.1. 内部组件

Pacemaker本身由四个关键组件组成:

- CIB (aka. 集群信息基础)
- CRMd (aka. 集群资源管理守护进程)
- PEngine (aka. PE or 策略引擎)
- STONITHd

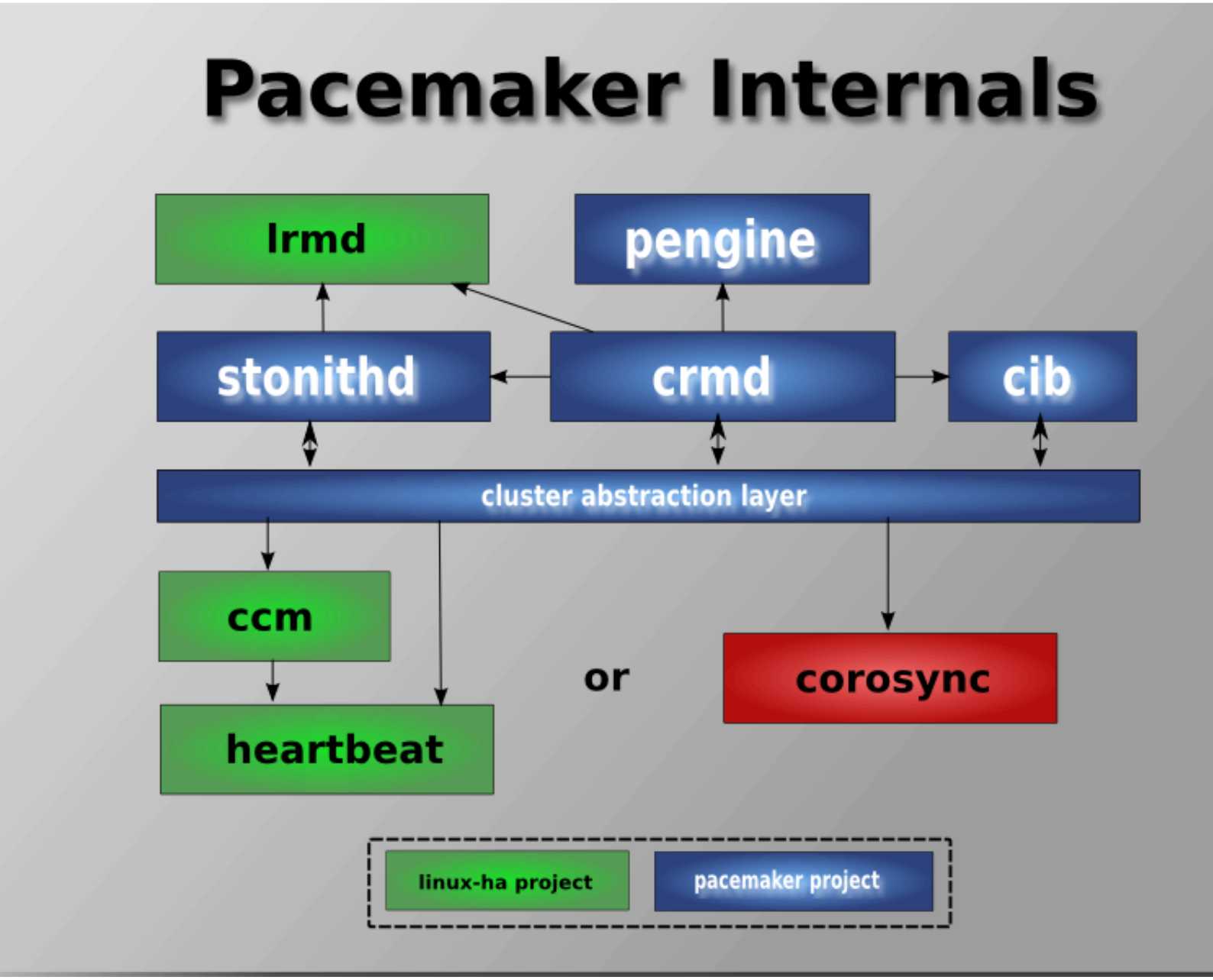


图 1.3. 内部组件

The CIB uses XML to represent both the cluster's configuration and current state of all resources in the cluster. The contents of the CIB are automatically kept in sync across the entire cluster and are used by the PEngine to compute the ideal state of the cluster and how it should be achieved.

This list of instructions is then fed to the DC (Designated Co-ordinator). Pacemaker centralizes all cluster decision making by electing one of the CRMd instances to act as a master. Should the elected CRMd process, or the node it is on, fail… a new one is quickly established.

The DC carries out the PEngine's instructions in the required order by passing them to either the LRMd (Local Resource Management daemon) or CRMd peers on other nodes via the cluster messaging infrastructure (which in turn passes them on to their LRMd process).

节点会把他们所有操作的日志发给DC，然后根据预期的结果和实际的结果(之间的差异)， 执行下一个等待中的命令，或者取消操作，并让PEngine根据非预期的结果重新计算集群的理想状态。

在某些情况下，可能会需要关闭节点的电源来保证共享数据的完整性或是完全地恢复资源。为此Pacemaker引入了STONITHd。STONITH是 Shoot-The-Other-Node-In-The-Head(爆其他节点的头)的缩写，并且通常是靠远程电源开关来实现的。在Pacemaker中，STONITH设备被当成资源(并且是在CIB中配置)从而轻松地监控，然而STONITHd会注意理解STONITH拓扑，比如它的客户端请求隔离一个节点，它会重启那个机器。(译者注:就是说不同的爆头设备驱动会对相同的请求有不同的理解，这些都是在驱动中定义的。)

1.4. Pacemaker 集群的种类

Pacemaker对你的环境没有特定的要求，这使得它支持任何的冗余配置，包括 Active/Active，Active/Passive，N+1，N+M，N-to-1 and N-to-N。

In this document we will focus on the setup of a highly available Apache web server with an Active/Passive cluster using DRBD and Ext4 to store data. Then, we will upgrade this cluster to Active/Active using GFS2.

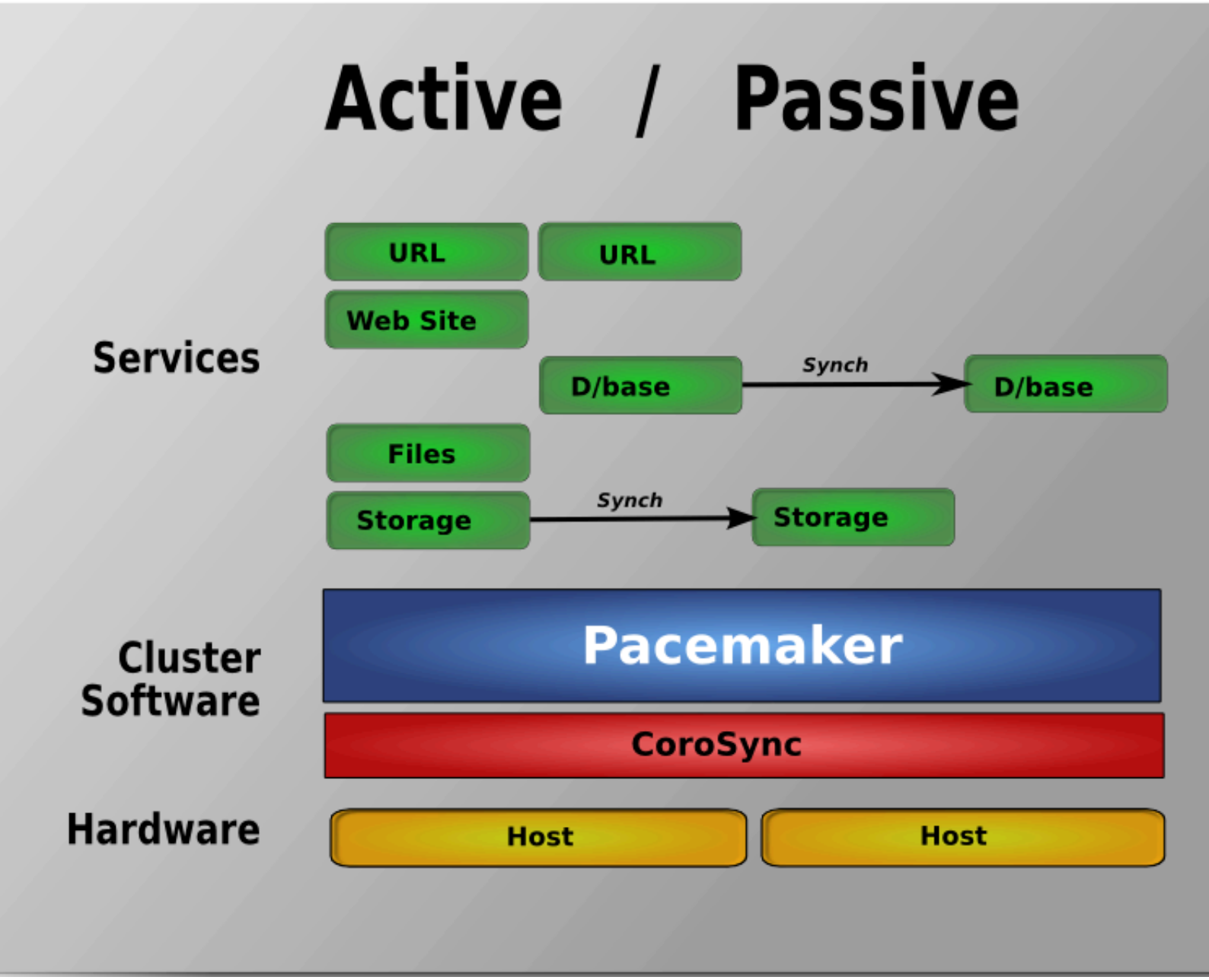


图 1.4. Active/Passive 冗余

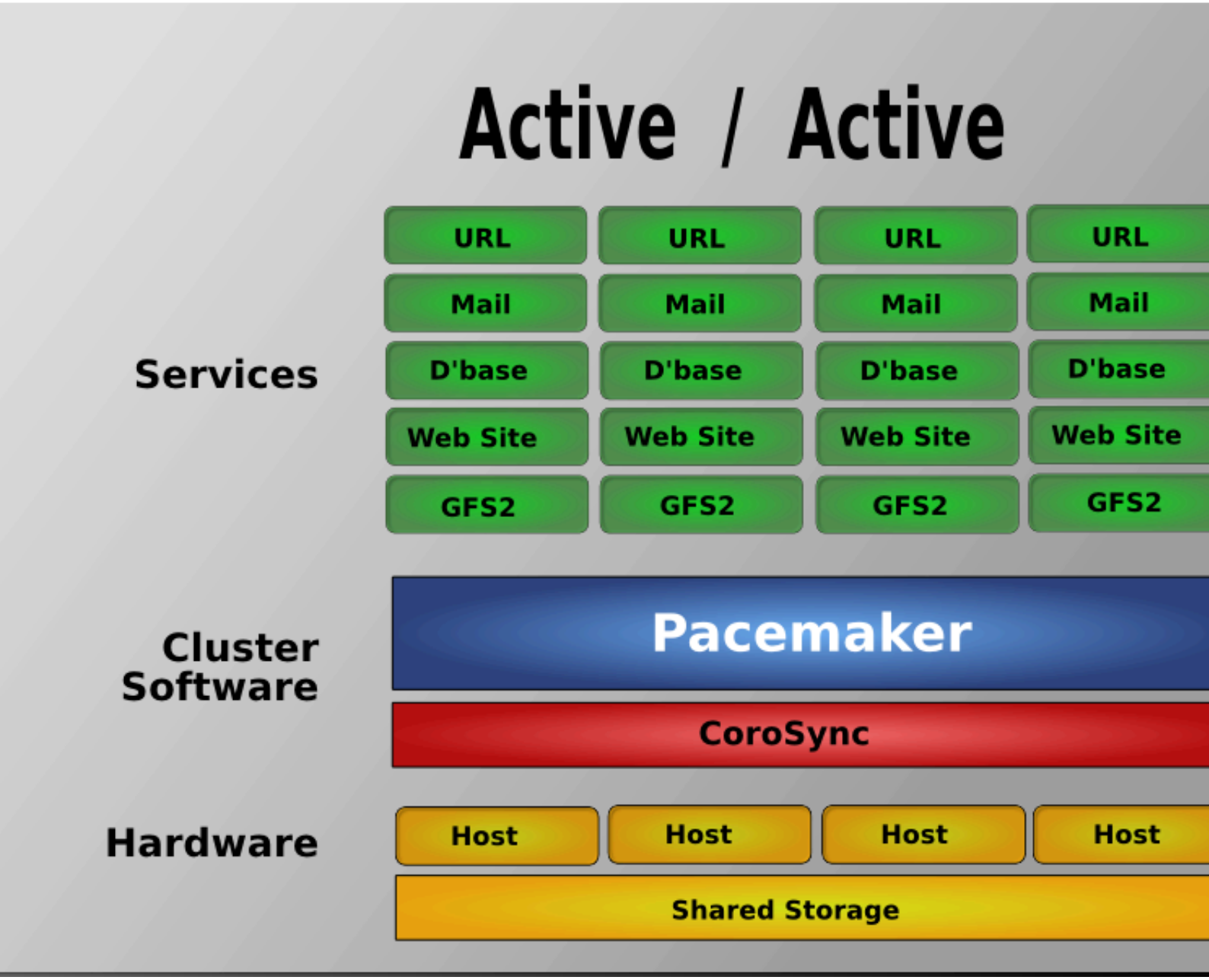

Active / Active
Services
URL
URL
URL
URL
Mail
Mail
Mail
Mail
D'base
D'base
D'base
D'base
Web Site
Web Site
Web Site
Web Site
GFS2
GFS2
GFS2
GFS2
Cluster Software
Pacemaker
CoroSync
Hardware
Host
Host
Host
Host
Shared Storage

图 1.5. N to N 冗余

安装

目录

2.1. 安装操作系统

Detailed instructions for installing Fedora are available at *http://docs.fedoraproject.org/install-guide/f13/* in a number of languages. The abbreviated version is as follows…

Point your browser to *http://fedoraproject.org/en/get-fedora-all*, locate the Install Media section and download the install DVD that matches your hardware.

Burn the disk image to a DVD [1] and boot from it. Or use the image to boot a virtual machine as I have done here. After clicking through the welcome screen, select your language and keyboard layout [2]

[1] *http://docs.fedoraproject.org/readme-burning-isos/en-US.html*
[2] *http://docs.fedoraproject.org/install-guide/f13/en-US/html/s1-langselection-x86.html*

图 2.1. Installation: Good choice

图 2.2. 安装Fedora - 存储设备

Assign your machine a host name. [3] I happen to control the clusterlabs.org domain name, so I will use that here.

[3] http://docs.fedoraproject.org/install-guide/f13/en-US/html/sn-networkconfig-fedora.html

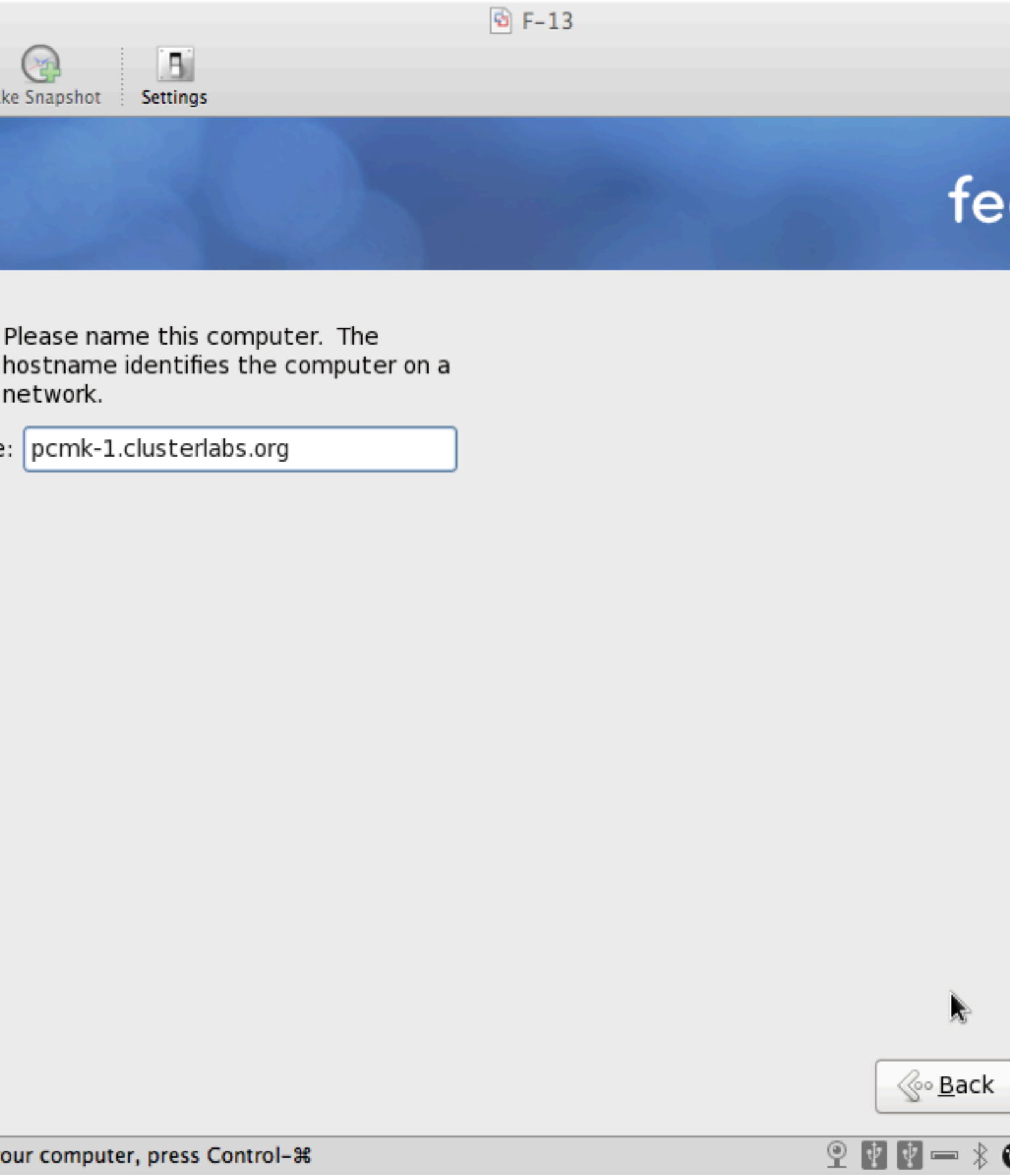


图 2.3. 安装Fedora -机器名

You will then be prompted to indicate the machine's physical location and to supply a root password. [4]

Now select where you want Fedora installed. [5]

As I don't care about any existing data, I will accept the default and allow Fedora to use the complete drive. However I want to reserve some space for DRBD, so I'll check the Review and modify partitioning layout box.

[4] *http://docs.fedoraproject.org/install-guide/f13/en-US/html/sn-account_configuration.html*
[5] *http://docs.fedoraproject.org/install-guide/f13/en-US/html/s1-diskpartsetup-x86.html*

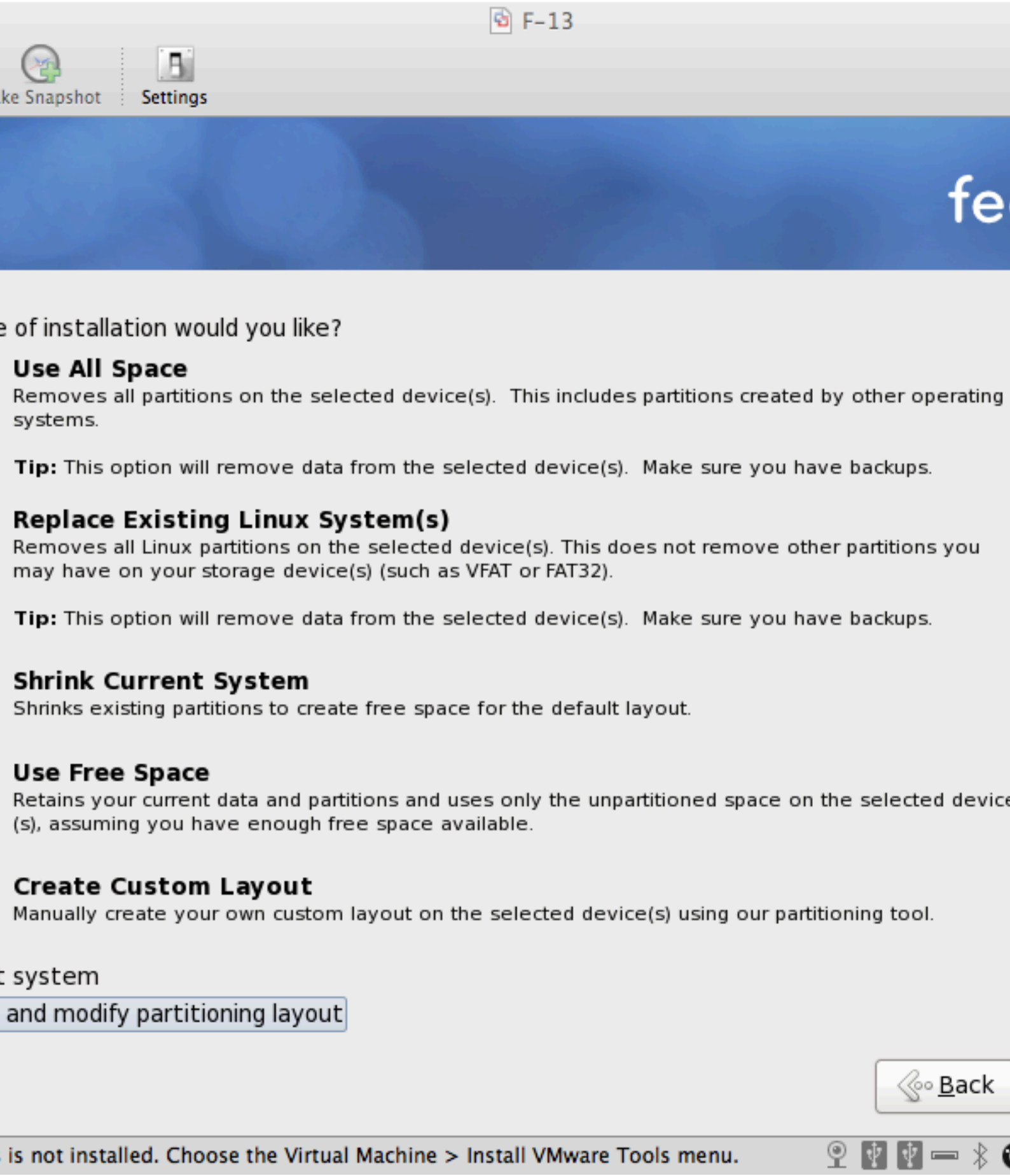


图 2.4. 安装Fedora - 安装类型

By default, Fedora will give all the space to the / (aka. root) partition. We'll take some back so we can use DRBD.

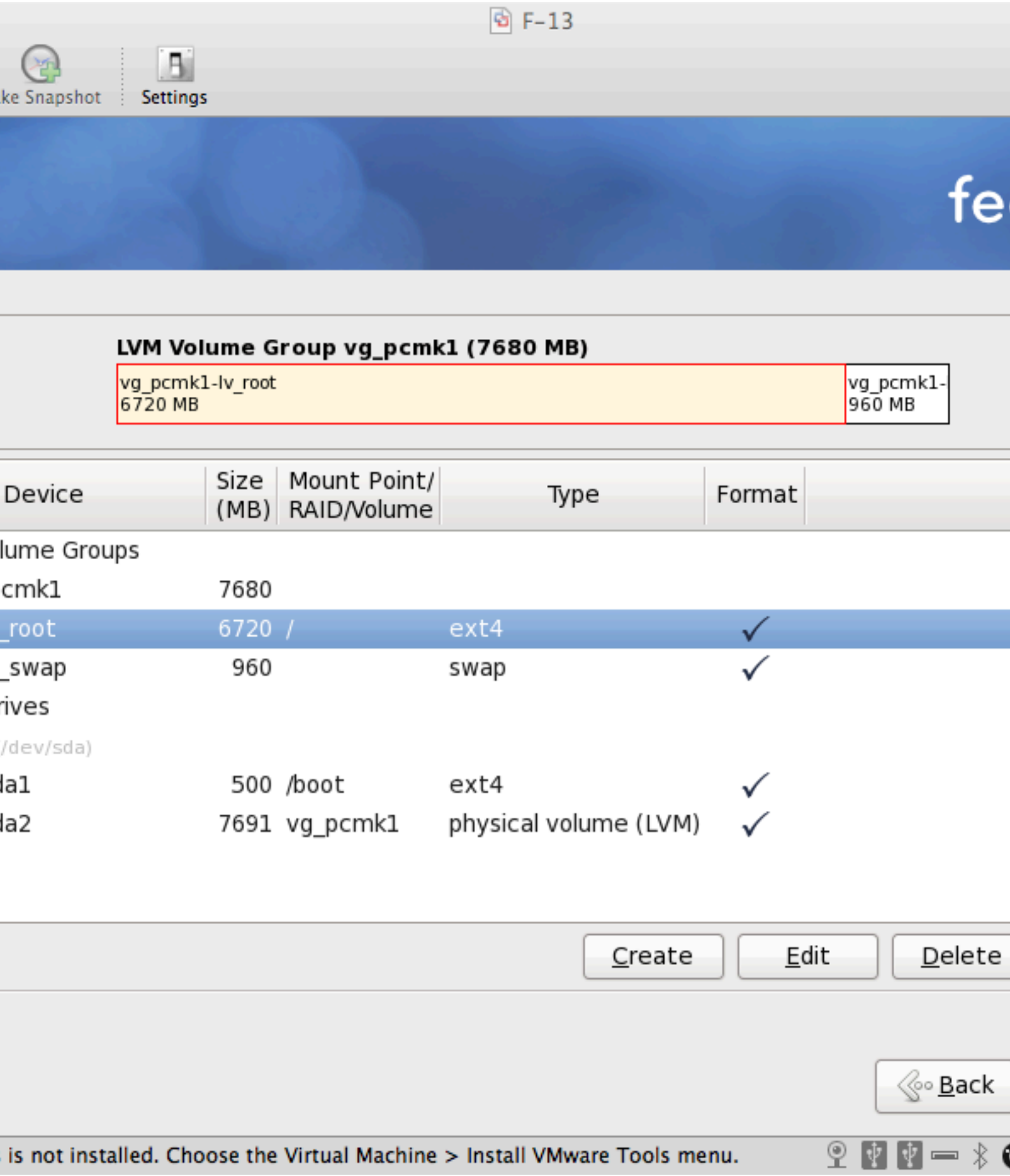


图 2.5. 安装Fedora - 默认分区

完整的分区应该像下面一样。

重要

If you plan on following the DRBD or GFS2 portions of this guide, you should reserve at least 1Gb of space on each machine from which to create a shared volume. Fedora Installation - Customize PartitioningFedora Installation: Create a partition to use (later) for website data

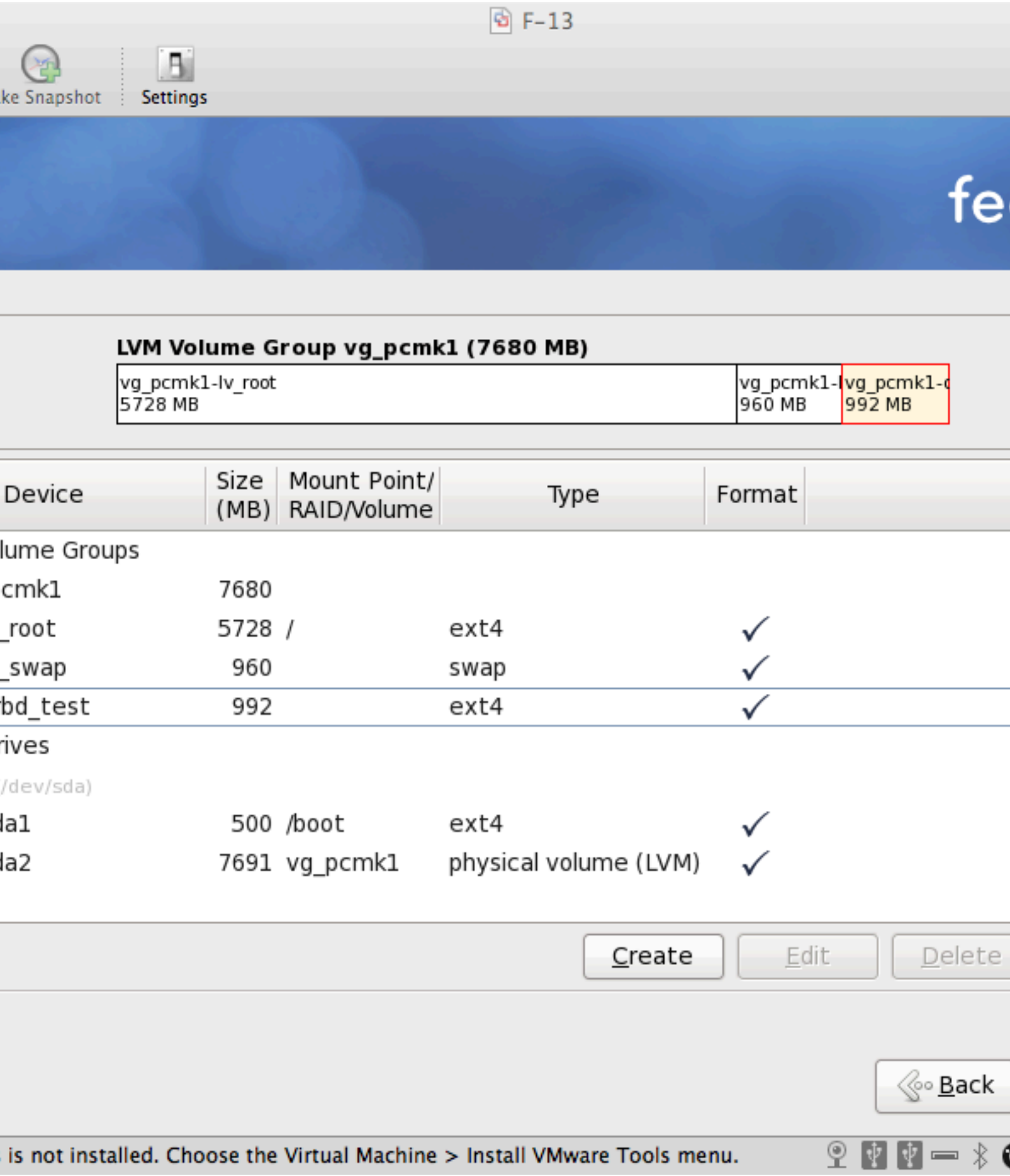


图 2.6. 安装Fedora - 自定义分区

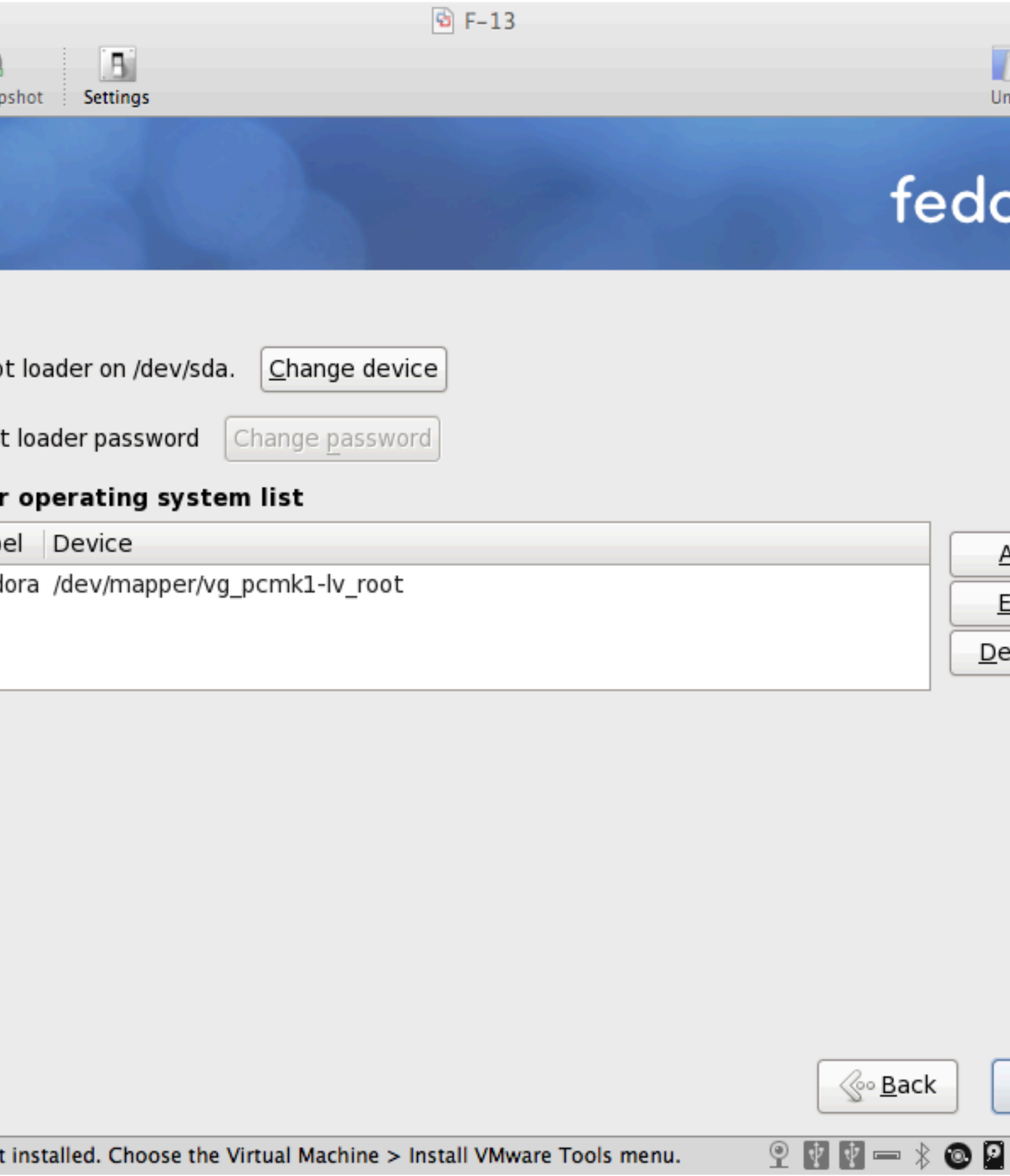


图 2.7. 安装Fedora - Bootloader

Next choose which software should be installed. Change the selection to Web Server since we plan on using Apache. Don't enable updates yet, we'll do that (and install any extra software we need) later. After you click next, Fedora will begin installing.

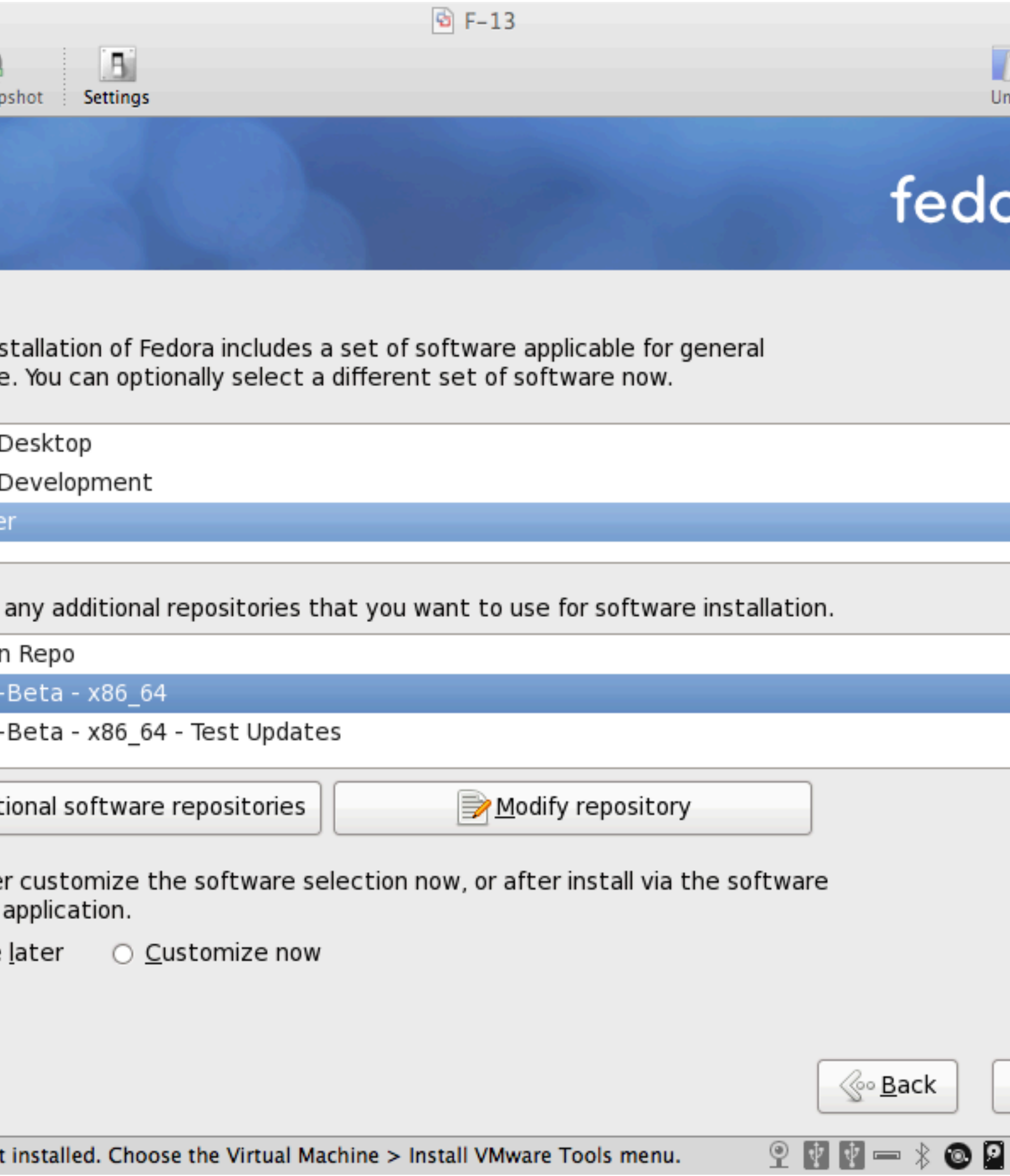


图 2.8. 安装Fedora - 软件

Go grab something to drink, this may take a while

图 2.9. 安装Fedora - 安装中

图 2.10. 安装Fedora - 安装完成

Once the node reboots, follow the on screen instructions [6] to create a system user and configure the time.

[6] http://docs.fedoraproject.org/install-guide/f13/en-US/html/ch-firstboot.html

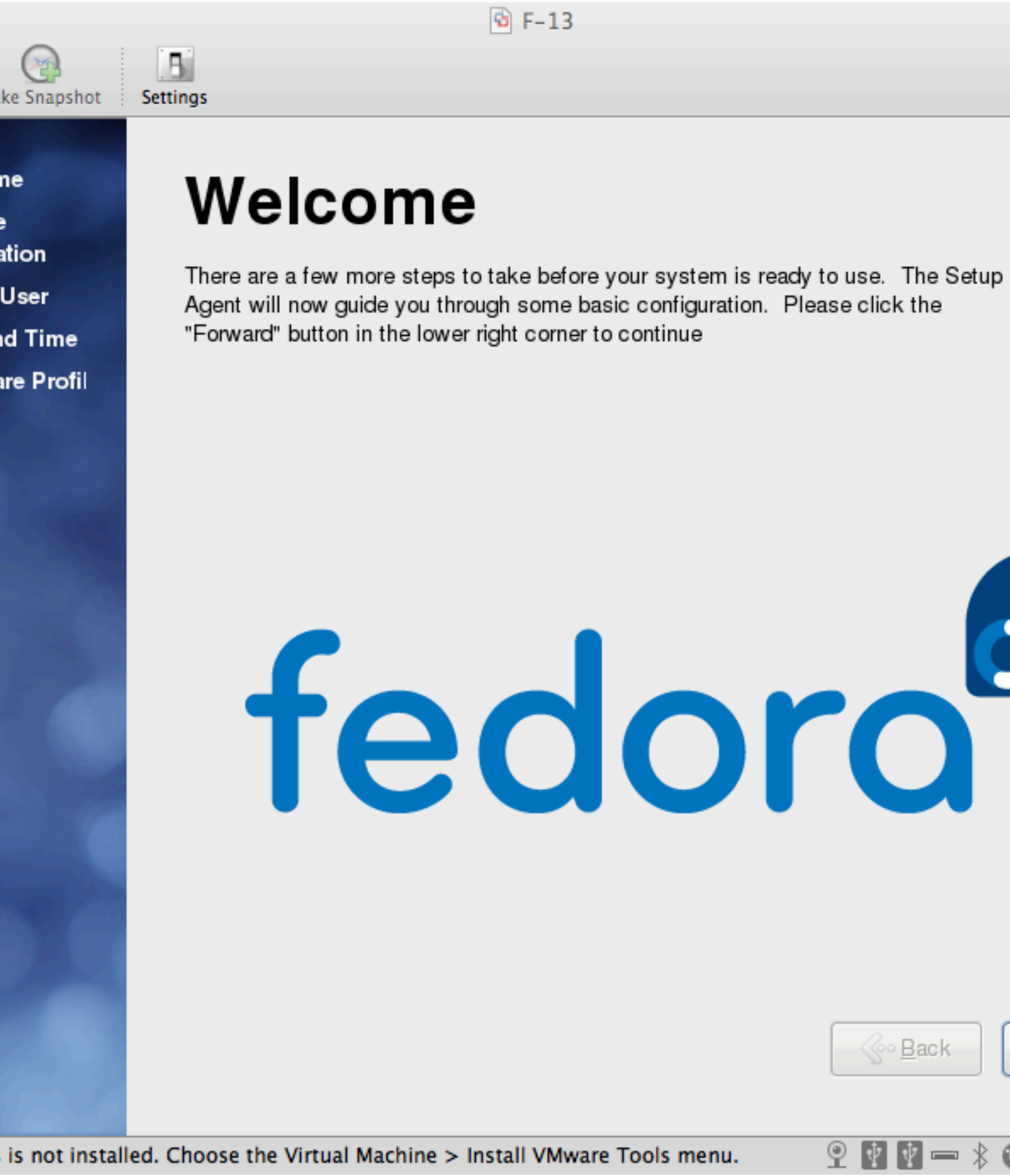


图 2.11. 安装Fedora - 第一次启动

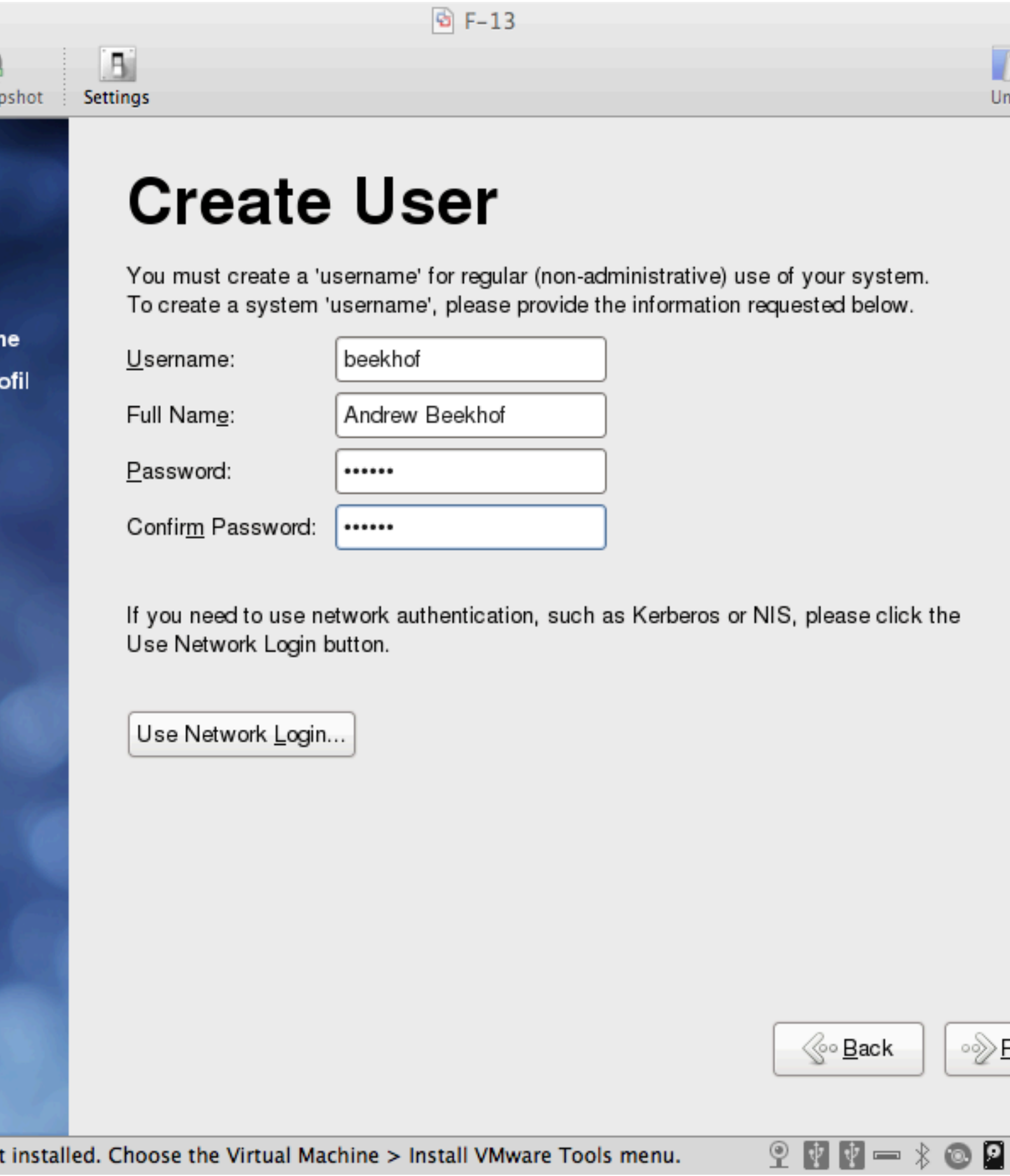


图 2.12. 安装Fedora - 创建非特权用户

注意

It is highly recommended to enable NTP on your cluster nodes. Doing so ensures all nodes agree on the current time and makes reading log files significantly easier. Fedora Installation - Date and TimeFedora Installation: Enable NTP to keep the times on all your nodes consistent

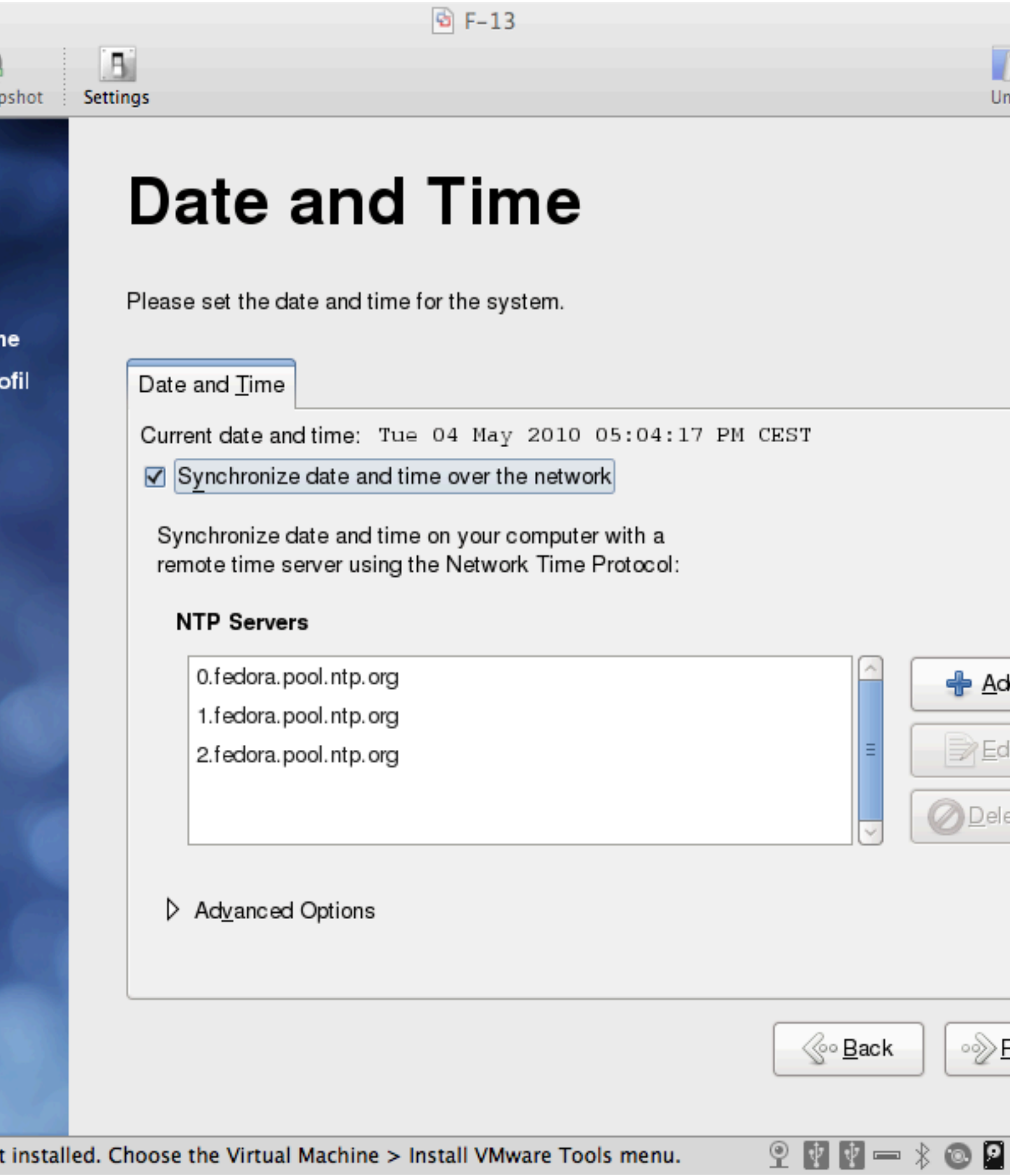
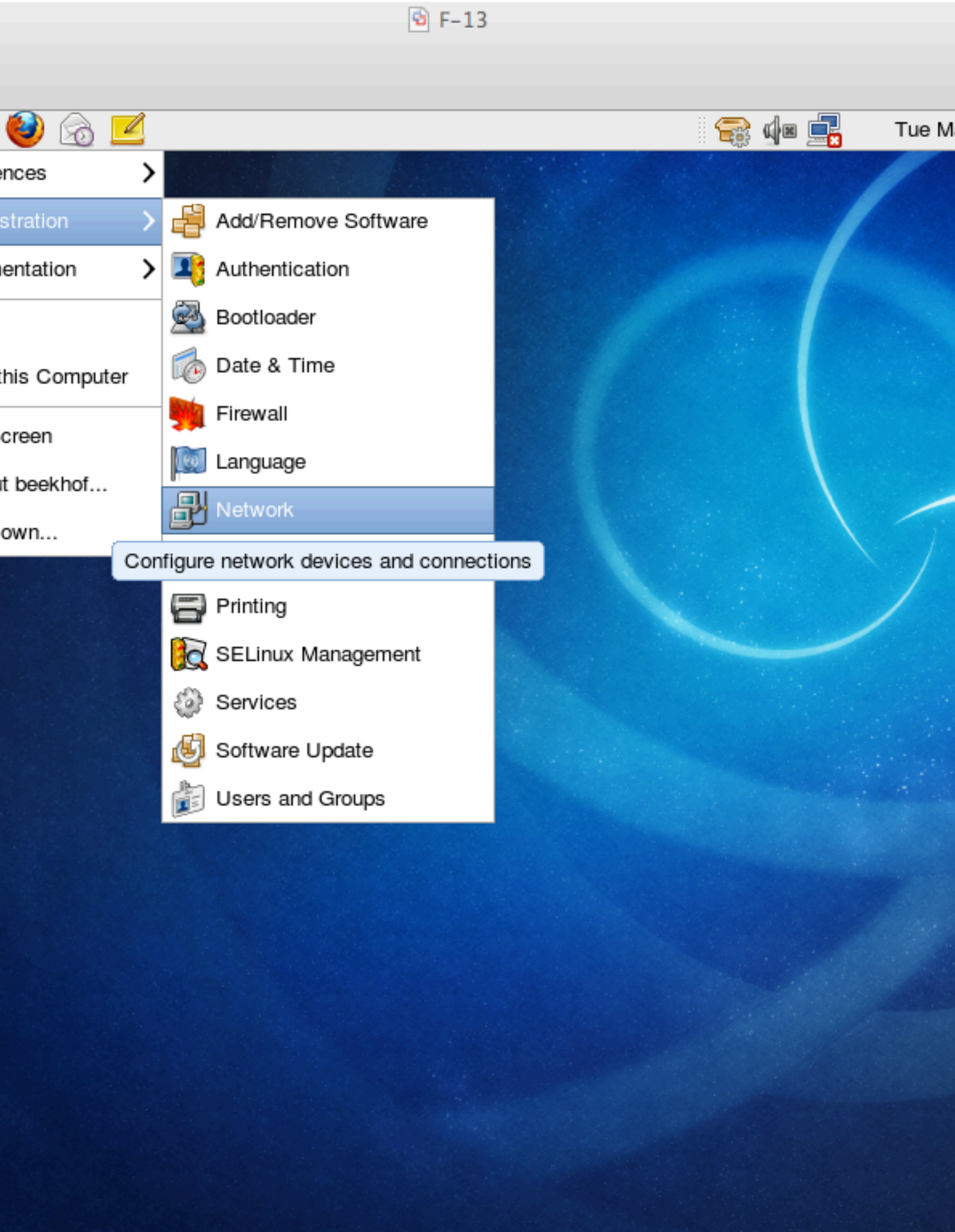


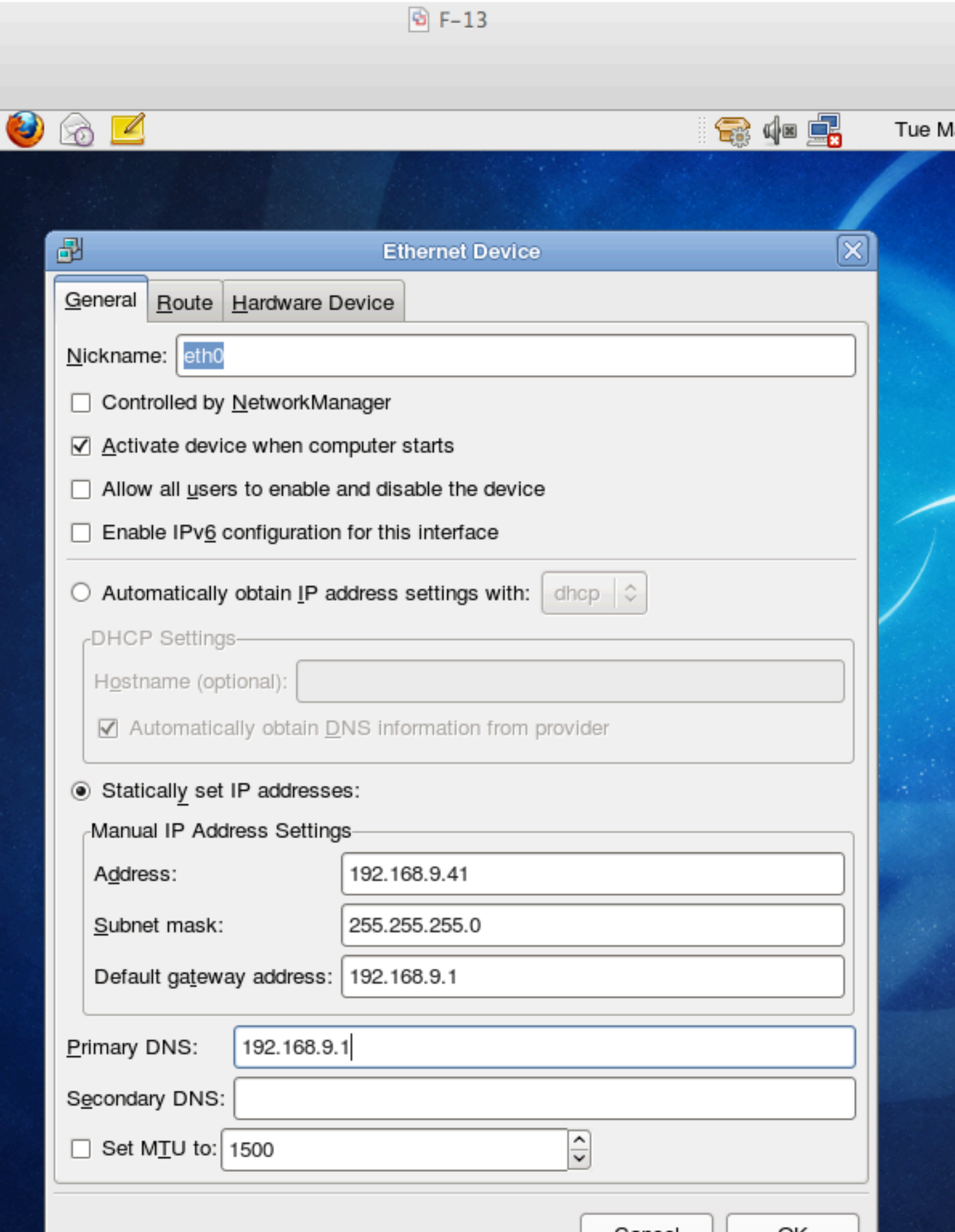
图 2.13. 安装Fedora - 日期和时间

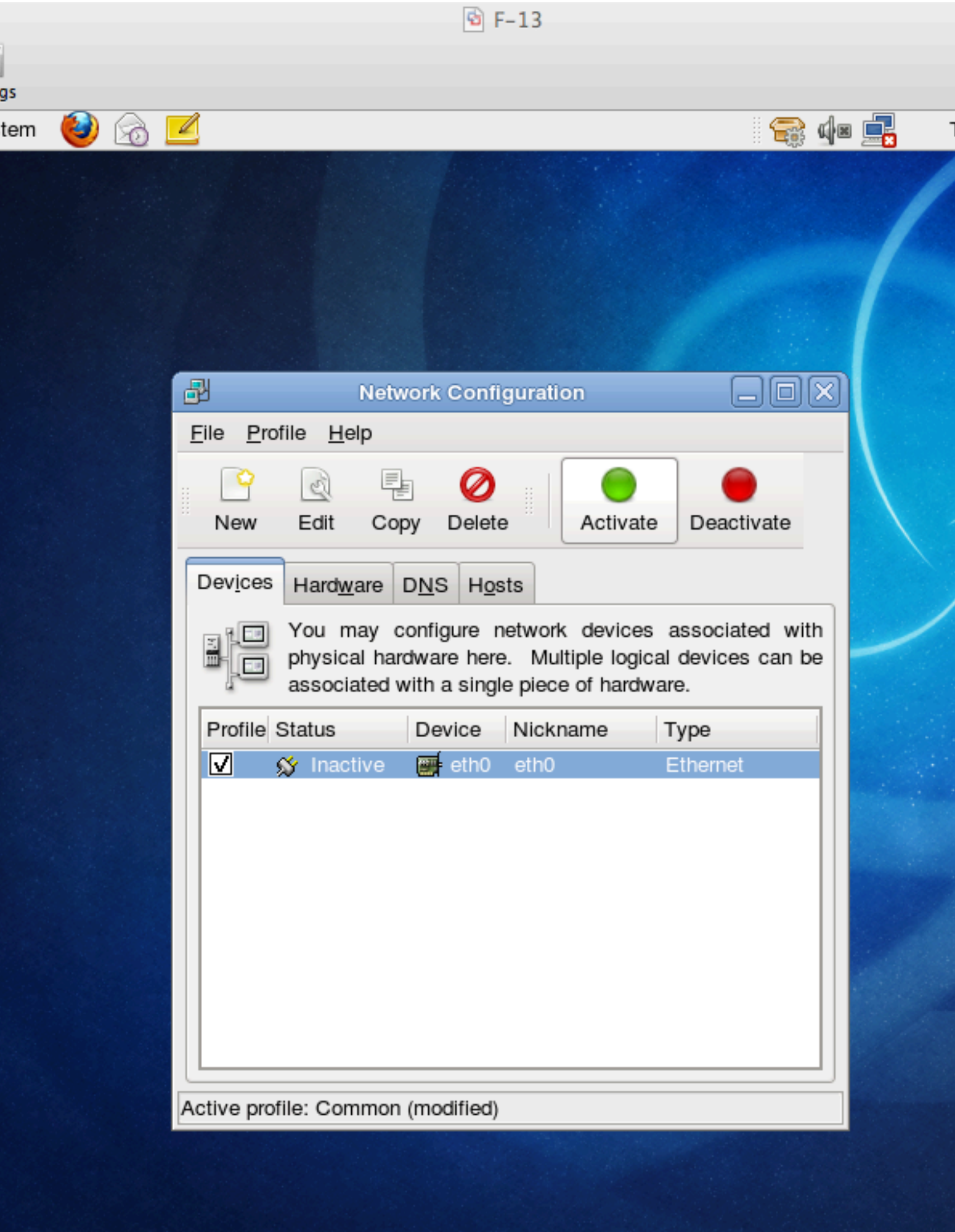
点击next会进入登入界面，点击你创建的用户并输入之前设定的密码。

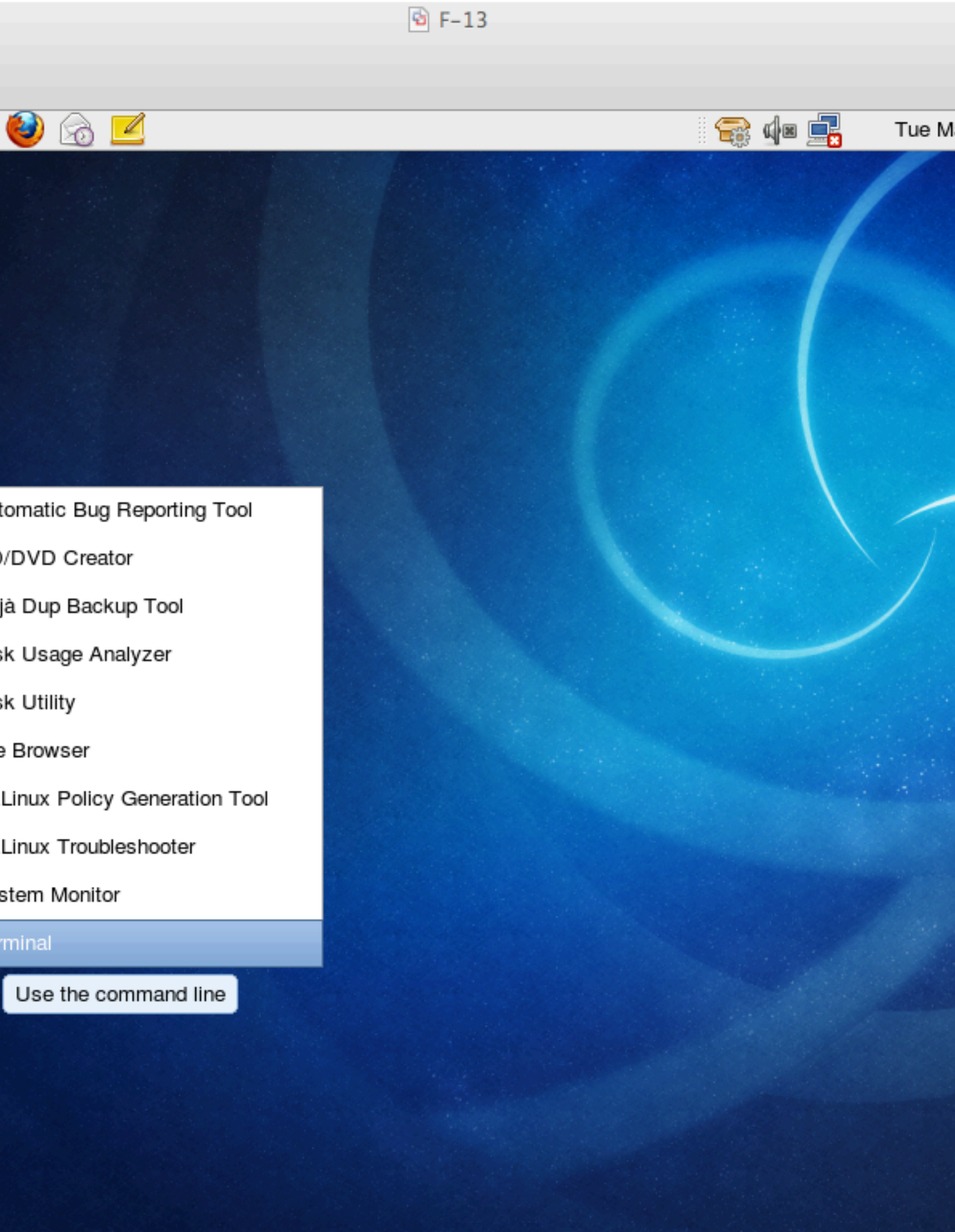

F-13
Tue M
nces
stration
entation
his Computer
creen
t beekhof...
own...
Add/Remove Software
Authentication
Bootloader
Date & Time
Firewall
Language
Network
Configure network devices and connections
Printing
SELinux Management
Services
Software Update
Users and Groups

重要

Do not accept the default network settings. Cluster machines should never obtain an ip address via DHCP. Here I will use the internal addresses for the clusterlab.org network.


F-13
Tue M
Ethernet Device
General
Route
Hardware Device
Nickname:
eth0
Controlled by NetworkManager
Activate device when computer starts
Allow all users to enable and disable the device
Enable IPv6 configuration for this interface
Automatically obtain IP address settings with:
dhcp
DHCP Settings
Hostname (optional):
Automatically obtain DNS information from provider
Statically set IP addresses:
Manual IP Address Settings
Address:
192.168.9.41
Subnet mask:
255.255.255.0
Default gateway address:
192.168.9.1
Primary DNS:
192.168.9.1
Secondary DNS:
Set MTU to:
1500


F-13
Network Configuration
File Profile Help
New Edit Copy Delete Activate Deactivate
Devices Hardware DNS Hosts
You may configure network devices associated with physical hardware here. Multiple logical devices can be associated with a single piece of hardware.
Profile Status Device Nickname Type
Inactive eth0 eth0 Ethernet
Active profile: Common (modified)


F-13
Tue M
tomatic Bug Reporting Tool
/DVD Creator
jà Dup Backup Tool
sk Usage Analyzer
sk Utility
e Browser
Linux Policy Generation Tool
Linux Troubleshooter
stem Monitor
rminal
Use the command line

注意

这是最后一个截屏了，剩下的我们都用命令行来操作。

2.2. 集群软件安装

Go to the terminal window you just opened and switch to the super user (aka. "root") account with the su command. You will need to supply the password you entered earlier during the installation process.

```
[beekhof@pcmk-1 ~]$ su -
Password:
[root@pcmk-1 ~]#
```

注意

注意用户名（@符号左边的字符串）显示我们现在使用的是root用户.

```
# ip addr
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 16436 qdisc noqueue state UNKNOWN
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
    inet6 ::1/128 scope host
       valid_lft forever preferred_lft forever
2: eth0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast state UNKNOWN qlen 1000
    link/ether 00:0c:29:6f:e1:58 brd ff:ff:ff:ff:ff:ff
    inet 192.168.9.41/24 brd 192.168.9.255 scope global eth0
    inet6 ::20c:29ff:fe6f:e158/64 scope global dynamic
       valid_lft 2591667sec preferred_lft 604467sec
    inet6 2002:57ae:43fc:0:20c:29ff:fe6f:e158/64 scope global dynamic
       valid_lft 2591990sec preferred_lft 604790sec
    inet6 fe80::20c:29ff:fe6f:e158/64 scope link
       valid_lft forever preferred_lft forever
# ping -c 1 www.google.com
PING www.l.google.com (74.125.39.99) 56(84) bytes of data.
64 bytes from fx-in-f99.1e100.net (74.125.39.99): icmp_seq=1 ttl=56 time=16.7 ms

--- www.l.google.com ping statistics ---
1 packets transmitted, 1 received, 0% packet loss, time 20ms
rtt min/avg/max/mdev = 16.713/16.713/16.713/0.000 ms
# /sbin/chkconfig network on
#
```

2.2.1. 安全提示

为了简化本文档并更好的关注集群方面的问题，我们现在在先禁用防火墙和SELinux。这些操作都会导致重大的安全问题,并不推荐对公网上的集群这样做。

重要

TODO: Create an Appendix that deals with (at least) re-enabling the firewall.

```
# sed -i.bak "s/SELINUX=enforcing/SELINUX=permissive/g" /etc/selinux/config
# /sbin/chkconfig --del iptables
# service iptables stop
iptables: Flushing firewall rules:                         [  OK  ]
iptables: Setting chains to policy ACCEPT: filter          [  OK  ]
iptables: Unloading modules:                               [  OK  ]
```

注意

你需要重启来保证SELinux正确关闭。不然你启动corosync的时候将看到以下提示：

```
May  4 19:30:54 pcmk-1 setroubleshoot: SELinux is preventing /usr/sbin/corosync "getattr" access on /. For
 complete SELinux messages. run sealert -l 6e0d4384-638e-4d55-9aaf-7dac011f29c1
May  4 19:30:54 pcmk-1 setroubleshoot: SELinux is preventing /usr/sbin/corosync "getattr" access on /. For
 complete SELinux messages. run sealert -l 6e0d4384-638e-4d55-9aaf-7dac011f29c1
```

2.2.2. 安装集群软件

从Fedora 12开始，你需要的东西都已经准备好了，只需在终端命令行运行以下命令：

```
# sed -i.bak "s/enabled=0/enabled=1/g"
/etc/yum.repos.d/fedora.repo
# sed -i.bak "s/enabled=0/enabled=1/g"
/etc/yum.repos.d/fedora-updates.repo
# yum install -y pacemaker corosync
Loaded plugins: presto, refresh-packagekit
fedora/metalink                                          |  22 kB     00:00
fedora-debuginfo/metalink                                |  16 kB     00:00
fedora-debuginfo                                         | 3.2 kB     00:00
fedora-debuginfo/primary_db                              | 1.4 MB     00:04
fedora-source/metalink                                   |  22 kB     00:00
fedora-source                                            | 3.2 kB     00:00
fedora-source/primary_db                                 | 3.0 MB     00:05
updates/metalink                                         |  26 kB     00:00
updates                                                  | 2.6 kB     00:00
updates/primary_db                                       | 1.1 kB     00:00
updates-debuginfo/metalink                               |  18 kB     00:00
updates-debuginfo                                        | 2.6 kB     00:00
updates-debuginfo/primary_db                             | 1.1 kB     00:00
updates-source/metalink                                  |  25 kB     00:00
updates-source                                           | 2.6 kB     00:00
updates-source/primary_db                                | 1.1 kB     00:00
Setting up Install Process
Resolving Dependencies
--> Running transaction check
---> Package corosync.x86_64 0:1.2.1-1.fc13 set to be updated
--> Processing Dependency: corosynclib = 1.2.1-1.fc13 for package: corosync-1.2.1-1.fc13.x86_64
--> Processing Dependency: libquorum.so.4(COROSYNC_QUORUM_1.0)(64bit) for package: corosync-1.2.1-1.fc13.x86_64
```

```
--> Processing Dependency: libvotequorum.so.4(COROSYNC_VOTEQUORUM_1.0)(64bit) for package:
 corosync-1.2.1-1.fc13.x86_64
--> Processing Dependency: libcpg.so.4(COROSYNC_CPG_1.0)(64bit) for package: corosync-1.2.1-1.fc13.x86_64
--> Processing Dependency: libconfdb.so.4(COROSYNC_CONFDB_1.0)(64bit) for package: corosync-1.2.1-1.fc13.x86_64
--> Processing Dependency: libcfg.so.4(COROSYNC_CFG_0.82)(64bit) for package: corosync-1.2.1-1.fc13.x86_64
--> Processing Dependency: libpload.so.4(COROSYNC_PLOAD_1.0)(64bit) for package: corosync-1.2.1-1.fc13.x86_64
--> Processing Dependency: liblogsys.so.4()(64bit) for package: corosync-1.2.1-1.fc13.x86_64
--> Processing Dependency: libconfdb.so.4()(64bit) for package: corosync-1.2.1-1.fc13.x86_64
--> Processing Dependency: libcoroipcc.so.4()(64bit) for package: corosync-1.2.1-1.fc13.x86_64
--> Processing Dependency: libcpg.so.4()(64bit) for package: corosync-1.2.1-1.fc13.x86_64
--> Processing Dependency: libquorum.so.4()(64bit) for package: corosync-1.2.1-1.fc13.x86_64
--> Processing Dependency: libcoroipcs.so.4()(64bit) for package: corosync-1.2.1-1.fc13.x86_64
--> Processing Dependency: libvotequorum.so.4()(64bit) for package: corosync-1.2.1-1.fc13.x86_64
--> Processing Dependency: libcfg.so.4()(64bit) for package: corosync-1.2.1-1.fc13.x86_64
--> Processing Dependency: libtotem_pg.so.4()(64bit) for package: corosync-1.2.1-1.fc13.x86_64
--> Processing Dependency: libpload.so.4()(64bit) for package: corosync-1.2.1-1.fc13.x86_64
---> Package pacemaker.x86_64 0:1.1.5-1.fc13 set to be updated
--> Processing Dependency: heartbeat >= 3.0.0 for package: pacemaker-1.1.5-1.fc13.x86_64
--> Processing Dependency: net-snmp >= 5.4 for package: pacemaker-1.1.5-1.fc13.x86_64
--> Processing Dependency: resource-agents for package: pacemaker-1.1.5-1.fc13.x86_64
--> Processing Dependency: cluster-glue for package: pacemaker-1.1.5-1.fc13.x86_64
--> Processing Dependency: libnetsnmp.so.20()(64bit) for package: pacemaker-1.1.5-1.fc13.x86_64
--> Processing Dependency: libcrmcluster.so.1()(64bit) for package: pacemaker-1.1.5-1.fc13.x86_64
--> Processing Dependency: libpengine.so.3()(64bit) for package: pacemaker-1.1.5-1.fc13.x86_64
--> Processing Dependency: libnetsnmpagent.so.20()(64bit) for package: pacemaker-1.1.5-1.fc13.x86_64
--> Processing Dependency: libesmtp.so.5()(64bit) for package: pacemaker-1.1.5-1.fc13.x86_64
--> Processing Dependency: libstonithd.so.1()(64bit) for package: pacemaker-1.1.5-1.fc13.x86_64
--> Processing Dependency: libhbclient.so.1()(64bit) for package: pacemaker-1.1.5-1.fc13.x86_64
--> Processing Dependency: libpils.so.2()(64bit) for package: pacemaker-1.1.5-1.fc13.x86_64
--> Processing Dependency: libpe_status.so.2()(64bit) for package: pacemaker-1.1.5-1.fc13.x86_64
--> Processing Dependency: libnetsnmpmibs.so.20()(64bit) for package: pacemaker-1.1.5-1.fc13.x86_64
--> Processing Dependency: libnetsnmphelpers.so.20()(64bit) for package: pacemaker-1.1.5-1.fc13.x86_64
--> Processing Dependency: libcib.so.1()(64bit) for package: pacemaker-1.1.5-1.fc13.x86_64
--> Processing Dependency: libccmclient.so.1()(64bit) for package: pacemaker-1.1.5-1.fc13.x86_64
--> Processing Dependency: libstonith.so.1()(64bit) for package: pacemaker-1.1.5-1.fc13.x86_64
--> Processing Dependency: liblrm.so.2()(64bit) for package: pacemaker-1.1.5-1.fc13.x86_64
--> Processing Dependency: libtransitioner.so.1()(64bit) for package: pacemaker-1.1.5-1.fc13.x86_64
--> Processing Dependency: libpe_rules.so.2()(64bit) for package: pacemaker-1.1.5-1.fc13.x86_64
--> Processing Dependency: libcrmcommon.so.2()(64bit) for package: pacemaker-1.1.5-1.fc13.x86_64
--> Processing Dependency: libplumb.so.2()(64bit) for package: pacemaker-1.1.5-1.fc13.x86_64
--> Running transaction check
---> Package cluster-glue.x86_64 0:1.0.2-1.fc13 set to be updated
--> Processing Dependency: perl-TimeDate for package: cluster-glue-1.0.2-1.fc13.x86_64
--> Processing Dependency: libOpenIPMIutils.so.0()(64bit) for package: cluster-glue-1.0.2-1.fc13.x86_64
--> Processing Dependency: libOpenIPMIposix.so.0()(64bit) for package: cluster-glue-1.0.2-1.fc13.x86_64
--> Processing Dependency: libopenhpi.so.2()(64bit) for package: cluster-glue-1.0.2-1.fc13.x86_64
--> Processing Dependency: libOpenIPMI.so.0()(64bit) for package: cluster-glue-1.0.2-1.fc13.x86_64
---> Package cluster-glue-libs.x86_64 0:1.0.2-1.fc13 set to be updated
---> Package corosynclib.x86_64 0:1.2.1-1.fc13 set to be updated
--> Processing Dependency: librdmacm.so.1(RDMACM_1.0)(64bit) for package: corosynclib-1.2.1-1.fc13.x86_64
--> Processing Dependency: libibverbs.so.1(IBVERBS_1.0)(64bit) for package: corosynclib-1.2.1-1.fc13.x86_64
--> Processing Dependency: libibverbs.so.1(IBVERBS_1.1)(64bit) for package: corosynclib-1.2.1-1.fc13.x86_64
--> Processing Dependency: libibverbs.so.1()(64bit) for package: corosynclib-1.2.1-1.fc13.x86_64
--> Processing Dependency: librdmacm.so.1()(64bit) for package: corosynclib-1.2.1-1.fc13.x86_64
---> Package heartbeat.x86_64 0:3.0.0-0.7.0daab7da36a8.hg.fc13 set to be updated
--> Processing Dependency: PyXML for package: heartbeat-3.0.0-0.7.0daab7da36a8.hg.fc13.x86_64
---> Package heartbeat-libs.x86_64 0:3.0.0-0.7.0daab7da36a8.hg.fc13 set to be updated
---> Package libesmtp.x86_64 0:1.0.4-12.fc12 set to be updated
---> Package net-snmp.x86_64 1:5.5-12.fc13 set to be updated
--> Processing Dependency: libsensors.so.4()(64bit) for package: 1:net-snmp-5.5-12.fc13.x86_64
---> Package net-snmp-libs.x86_64 1:5.5-12.fc13 set to be updated
---> Package pacemaker-libs.x86_64 0:1.1.5-1.fc13 set to be updated
---> Package resource-agents.x86_64 0:3.0.10-1.fc13 set to be updated
--> Processing Dependency: libnet.so.1()(64bit) for package: resource-agents-3.0.10-1.fc13.x86_64
--> Running transaction check
---> Package OpenIPMI-libs.x86_64 0:2.0.16-8.fc13 set to be updated
---> Package PyXML.x86_64 0:0.8.4-17.fc13 set to be updated
```

```
---> Package libibverbs.x86_64 0:1.1.3-4.fc13 set to be updated
--> Processing Dependency: libibverbs-driver for package: libibverbs-1.1.3-4.fc13.x86_64
---> Package libnet.x86_64 0:1.1.4-3.fc12 set to be updated
---> Package librdmacm.x86_64 0:1.0.10-2.fc13 set to be updated
---> Package lm_sensors-libs.x86_64 0:3.1.2-2.fc13 set to be updated
---> Package openhpi-libs.x86_64 0:2.14.1-3.fc13 set to be updated
---> Package perl-TimeDate.noarch 1:1.20-1.fc13 set to be updated
--> Running transaction check
---> Package libmlx4.x86_64 0:1.0.1-5.fc13 set to be updated
--> Finished Dependency Resolution

Dependencies Resolved


================================================================================
 Package                Arch     Version                            Repository      Size
================================================================================
Installing:
 corosync               x86_64   1.2.1-1.fc13                       fedora         136 k
 pacemaker              x86_64   1.1.5-1.fc13                       fedora         543 k
Installing for dependencies:
 OpenIPMI-libs          x86_64   2.0.16-8.fc13                      fedora         474 k
 PyXML                  x86_64   0.8.4-17.fc13                      fedora         906 k
 cluster-glue           x86_64   1.0.2-1.fc13                       fedora         230 k
 cluster-glue-libs      x86_64   1.0.2-1.fc13                       fedora         116 k
 corosynclib            x86_64   1.2.1-1.fc13                       fedora         145 k
 heartbeat              x86_64   3.0.0-0.7.0daab7da36a8.hg.fc13     updates        172 k
 heartbeat-libs         x86_64   3.0.0-0.7.0daab7da36a8.hg.fc13     updates        265 k
 libesmtp               x86_64   1.0.4-12.fc12                      fedora          54 k
 libibverbs             x86_64   1.1.3-4.fc13                       fedora          42 k
 libmlx4                x86_64   1.0.1-5.fc13                       fedora          27 k
 libnet                 x86_64   1.1.4-3.fc12                       fedora          49 k
 librdmacm              x86_64   1.0.10-2.fc13                      fedora          22 k
 lm_sensors-libs        x86_64   3.1.2-2.fc13                       fedora          37 k
 net-snmp               x86_64   1:5.5-12.fc13                      fedora         295 k
 net-snmp-libs          x86_64   1:5.5-12.fc13                      fedora         1.5 M
 openhpi-libs           x86_64   2.14.1-3.fc13                      fedora         135 k
 pacemaker-libs         x86_64   1.1.5-1.fc13                       fedora         264 k
 perl-TimeDate          noarch   1:1.20-1.fc13                      fedora          42 k
 resource-agents        x86_64   3.0.10-1.fc13                      fedora         357 k

Transaction Summary
================================================================================
Install      21 Package(s)
Upgrade       0 Package(s)

Total download size: 5.7 M
Installed size: 20 M
Downloading Packages:
Setting up and reading Presto delta metadata
updates-testing/prestodelta                                | 164 kB     00:00
fedora/prestodelta                                         |  150 B     00:00
Processing delta metadata
Package(s) data still to download: 5.7 M
(1/21): OpenIPMI-libs-2.0.16-8.fc13.x86_64.rpm             | 474 kB     00:00
(2/21): PyXML-0.8.4-17.fc13.x86_64.rpm                     | 906 kB     00:01
(3/21): cluster-glue-1.0.2-1.fc13.x86_64.rpm               | 230 kB     00:00
(4/21): cluster-glue-libs-1.0.2-1.fc13.x86_64.rpm          | 116 kB     00:00
(5/21): corosync-1.2.1-1.fc13.x86_64.rpm                   | 136 kB     00:00
(6/21): corosynclib-1.2.1-1.fc13.x86_64.rpm                | 145 kB     00:00
(7/21): heartbeat-3.0.0-0.7.0daab7da36a8.hg.fc13.x86_64.rpm | 172 kB     00:00
(8/21): heartbeat-libs-3.0.0-0.7.0daab7da36a8.hg.fc13.x86_64.rpm | 265 kB     00:00
(9/21): libesmtp-1.0.4-12.fc12.x86_64.rpm                  |  54 kB     00:00
(10/21): libibverbs-1.1.3-4.fc13.x86_64.rpm                |  42 kB     00:00
(11/21): libmlx4-1.0.1-5.fc13.x86_64.rpm                   |  27 kB     00:00
(12/21): libnet-1.1.4-3.fc12.x86_64.rpm                    |  49 kB     00:00
(13/21): librdmacm-1.0.10-2.fc13.x86_64.rpm                |  22 kB     00:00
```

```
(14/21): lm_sensors-libs-3.1.2-2.fc13.x86_64.rpm                        |  37 kB     00:00
(15/21): net-snmp-5.5-12.fc13.x86_64.rpm                                | 295 kB     00:00
(16/21): net-snmp-libs-5.5-12.fc13.x86_64.rpm                           | 1.5 MB     00:01
(17/21): openhpi-libs-2.14.1-3.fc13.x86_64.rpm                          | 135 kB     00:00
(18/21): pacemaker-1.1.5-1.fc13.x86_64.rpm                              | 543 kB     00:00
(19/21): pacemaker-libs-1.1.5-1.fc13.x86_64.rpm                         | 264 kB     00:00
(20/21): perl-TimeDate-1.20-1.fc13.noarch.rpm                           |  42 kB     00:00
(21/21): resource-agents-3.0.10-1.fc13.x86_64.rpm                       | 357 kB     00:00

Total                                                    539 kB/s | 5.7 MB     00:10
warning: rpmts_HdrFromFdno: Header V3 RSA/SHA256 Signature, key ID e8e40fde: NOKEY
fedora/gpgkey                                                           | 3.2 kB     00:00 ...
Importing GPG key 0xE8E40FDE "Fedora (13) <fedora@fedoraproject.org%gt;" from /etc/pki/rpm-gpg/RPM-GPG-KEY-
fedora-x86_64

Running rpm_check_debug
Running Transaction Test
Transaction Test Succeeded
Running Transaction
  Installing     : lm_sensors-libs-3.1.2-2.fc13.x86_64                          1/21
  Installing     : 1:net-snmp-libs-5.5-12.fc13.x86_64                           2/21
  Installing     : 1:net-snmp-5.5-12.fc13.x86_64                                3/21
  Installing     : openhpi-libs-2.14.1-3.fc13.x86_64                            4/21
  Installing     : libibverbs-1.1.3-4.fc13.x86_64                               5/21
  Installing     : libmlx4-1.0.1-5.fc13.x86_64                                  6/21
  Installing     : librdmacm-1.0.10-2.fc13.x86_64                               7/21
  Installing     : corosync-1.2.1-1.fc13.x86_64                                 8/21
  Installing     : corosynclib-1.2.1-1.fc13.x86_64                              9/21
  Installing     : libesmtp-1.0.4-12.fc12.x86_64                               10/21
  Installing     : OpenIPMI-libs-2.0.16-8.fc13.x86_64                          11/21
  Installing     : PyXML-0.8.4-17.fc13.x86_64                                  12/21
  Installing     : libnet-1.1.4-3.fc12.x86_64                                  13/21
  Installing     : 1:perl-TimeDate-1.20-1.fc13.noarch                          14/21
  Installing     : cluster-glue-1.0.2-1.fc13.x86_64                            15/21
  Installing     : cluster-glue-libs-1.0.2-1.fc13.x86_64                       16/21
  Installing     : resource-agents-3.0.10-1.fc13.x86_64                        17/21
  Installing     : heartbeat-libs-3.0.0-0.7.0daab7da36a8.hg.fc13.x86_64        18/21
  Installing     : heartbeat-3.0.0-0.7.0daab7da36a8.hg.fc13.x86_64             19/21
  Installing     : pacemaker-1.1.5-1.fc13.x86_64                               20/21
  Installing     : pacemaker-libs-1.1.5-1.fc13.x86_64                          21/21

Installed:
  corosync.x86_64 0:1.2.1-1.fc13                    pacemaker.x86_64 0:1.1.5-1.fc13

Dependency Installed:
  OpenIPMI-libs.x86_64 0:2.0.16-8.fc13
  PyXML.x86_64 0:0.8.4-17.fc13
  cluster-glue.x86_64 0:1.0.2-1.fc13
  cluster-glue-libs.x86_64 0:1.0.2-1.fc13
  corosynclib.x86_64 0:1.2.1-1.fc13
  heartbeat.x86_64 0:3.0.0-0.7.0daab7da36a8.hg.fc13
  heartbeat-libs.x86_64 0:3.0.0-0.7.0daab7da36a8.hg.fc13
  libesmtp.x86_64 0:1.0.4-12.fc12
  libibverbs.x86_64 0:1.1.3-4.fc13
  libmlx4.x86_64 0:1.0.1-5.fc13
  libnet.x86_64 0:1.1.4-3.fc12
  librdmacm.x86_64 0:1.0.10-2.fc13
  lm_sensors-libs.x86_64 0:3.1.2-2.fc13
  net-snmp.x86_64 1:5.5-12.fc13
  net-snmp-libs.x86_64 1:5.5-12.fc13
  openhpi-libs.x86_64 0:2.14.1-3.fc13
  pacemaker-libs.x86_64 0:1.1.5-1.fc13
  perl-TimeDate.noarch 1:1.20-1.fc13
  resource-agents.x86_64 0:3.0.10-1.fc13

Complete!
```

```
#
```

2.3. 写在开始之前

在另一台Fedora 12机器上面重复以上操作步骤，这样你就有2台安装了集群软件的节点了。

For the purposes of this document, the additional node is called pcmk-2 with address 192.168.122.102.

2.4. 安装

2.4.1. 设定网络

确认这两个新节点能够通讯:

```
# ping -c 3 192.168.122.102
PING 192.168.122.102 (192.168.122.102) 56(84) bytes of data.
64 bytes from 192.168.122.102: icmp_seq=1 ttl=64 time=0.343 ms
64 bytes from 192.168.122.102: icmp_seq=2 ttl=64 time=0.402 ms
64 bytes from 192.168.122.102: icmp_seq=3 ttl=64 time=0.558 ms

--- 192.168.122.102 ping statistics ---3 packets transmitted, 3 received, 0% packet loss, time 2000ms
rtt min/avg/max/mdev = 0.343/0.434/0.558/0.092 ms
```

Figure 2.18. Verify Connectivity by IP address

Now we need to make sure we can communicate with the machines by their name. If you have a DNS server, add additional entries for the two machines. Otherwise, you'll need to add the machines to /etc/hosts . Below are the entries for my cluster nodes:

```
# grep pcmk /etc/hosts
192.168.122.101 pcmk-1.clusterlabs.org pcmk-1
192.168.122.102 pcmk-2.clusterlabs.org pcmk-2
```

Figure 2.19. Set up /etc/hosts entries

现在让我们ping一下:

```
# ping -c 3 pcmk-2
PING pcmk-2.clusterlabs.org (192.168.122.101) 56(84) bytes of data.
64 bytes from pcmk-1.clusterlabs.org (192.168.122.101): icmp_seq=1 ttl=64 time=0.164 ms
64 bytes from pcmk-1.clusterlabs.org (192.168.122.101): icmp_seq=2 ttl=64 time=0.475 ms
64 bytes from pcmk-1.clusterlabs.org (192.168.122.101): icmp_seq=3 ttl=64 time=0.186 ms

--- pcmk-2.clusterlabs.org ping statistics ---3 packets transmitted, 3 received, 0% packet loss, time 2001ms
rtt min/avg/max/mdev = 0.164/0.275/0.475/0.141 ms
```

Figure 2.20. Verify Connectivity by Hostname

2.4.2. 配置SSH

SSH 是一个方便又安全来的用来远程传输文件或运行命令 的工具. 在这个文档中，我们创建ssh key(用 -N “” 选项)来免去登入要输入密码的麻烦。

警告

不推荐在公网的机器上采用未用密码保护的ssh-key

创建一个密钥并允许所有有这个密钥的用户登入

创建并激活一个新的SSH密钥

```
# ssh-keygen -t dsa -f ~/.ssh/id_dsa -N ""
Generating public/private dsa key pair.
Your identification has been saved in /root/.ssh/id_dsa.
Your public key has been saved in /root/.ssh/id_dsa.pub.
The key fingerprint is:
91:09:5c:82:5a:6a:50:08:4e:b2:0c:62:de:cc:74:44 root@pcmk-1.clusterlabs.org

The key's randomart image is:
+--[ DSA 1024]----+
|==.ooEo..        |
|X O + .o o       |
| * A    +        |
|  +      .       |
| .      S        |
|                 |
|                 |
|                 |
|                 |
+-----------------+

# cp .ssh/id_dsa.pub .ssh/authorized_keys
```

在其他节点安装这个密钥并测试你是否可以执行命令而不用输入密码

```
# scp -r .ssh pcmk-2:
The authenticity of host 'pcmk-2 (192.168.122.102)' can't be established.
RSA key fingerprint is b1:2b:55:93:f1:d9:52:2b:0f:f2:8a:4e:ae:c6:7c:9a.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'pcmk-2,192.168.122.102' (RSA) to the list of known hosts.root@pcmk-2's password:
id_dsa.pub                                    100%  616     0.6KB/s   00:00
id_dsa                                        100%  672     0.7KB/s   00:00
known_hosts                                   100%  400     0.4KB/s   00:00
authorized_keys                               100%  616     0.6KB/s   00:00
# ssh pcmk-2 -- uname -npcmk-2
#
```

Figure 2.22. Installing the SSH Key on Another Host

2.4.3. 简化节点名称

During installation, we filled in the machine's fully qualifier domain name (FQDN) which can be rather long when it appears in cluster logs and status output. See for yourself how the machine identifies itself:

```
# uname -n
pcmk-1.clusterlabs.org
```

```
# dnsdomainname clusterlabs.org
```

第二个命令的输出是正常的，但是我们真的不需要这么详细的输出，我们更改/etc/sysconfig/network文件来达到简化的目的。

```
# cat /etc/sysconfig/network
NETWORKING=yes
HOSTNAME=pcmk-1.clusterlabs.org
GATEWAY=192.168.122.1
```

我们要做的只是要把域名后面的部分去掉。

```
 # sed -i.bak 's/\.[a-z].*//g' /etc/sysconfig/network
```

现在cat一下看看更改是否成功了。

```
# cat /etc/sysconfig/network
NETWORKING=yes
HOSTNAME=pcmk-1
GATEWAY=192.168.122.1
```

然而到这里还没结束，机器还没接受新的配置文件，我们强制它生效。

```
# source /etc/sysconfig/network
# hostname $HOSTNAME
```

现在我们看看是否按达到我们预期的效果：

```
# uname -npcmk-1
# dnsdomainname clusterlabs.org
```

现在在pcmk-2上面重复以上操作.

2.4.4. 配置 Corosync

Choose a port number and multi-cast [7] address. [8] Be sure that the values you chose do not conflict with any existing clusters you might have. For advice on choosing a multi-cast address, see *http://www.29west.com/docs/THPM/multicast-address-assignment.html* For this document, I have chosen port 4000 and used 226.94.1.1 as the multi-cast address.

重要

The instructions below only apply for a machine with a single NIC. If you have a more complicated setup, you should edit the configuration manually.

[7] *http://en.wikipedia.org/wiki/Multicast*
[8] *http://en.wikipedia.org/wiki/Multicast_address*

```
# export ais_port=4000
# export ais_mcast=226.94.1.1
```

然后我们用下面的命令自动获得机器的地址。为了让配置文件能够在机器上面的各个机器通用，我们不使用完整的IP地址而使用网络地址。（译者注:corosync配置文件中的监听地址一项可以填写网络地址，corosync会自动匹配应该监听在哪个地址而不是0.0.0.0）

```
# export ais_addr=`ip addr | grep "inet " | tail -n 1 | awk '{print $4}' | sed s/255/0/`
```

显示并检查配置的环境变量是否正确

```
# env | grep ais_ais_mcast=226.94.1.1
ais_port=4000
ais_addr=192.168.122.0
```

确认以上输出没有错误以后，我们用以下命令来配置corosync

```
# cp /etc/corosync/corosync.conf.example /etc/corosync/corosync.conf
# sed -i.bak "s/.*mcastaddr:.*/mcastaddr:\ $ais_mcast/g" /etc/corosync/corosync.conf
# sed -i.bak "s/.*mcastport:.*/mcastport:\ $ais_port/g" /etc/corosync/corosync.conf
# sed -i.bak "s/.*bindnetaddr:.*/bindnetaddr:\ $ais_addr/g" /etc/corosync/corosync.conf
```

Finally, tell Corosync to load the Pacemaker plugin.

```
# cat <<-END >>/etc/corosync/service.d/pcmk
service {
        # Load the Pacemaker Cluster Resource Manager
        name: pacemaker
        ver:  1
}
END
```

The final configuration should look something like the sample in Appendix B, Sample Corosync Configuration.

重要

When run in version 1 mode, the plugin does not start the Pacemaker daemons. Instead it just sets up the quorum and messaging interfaces needed by the rest of the stack. Starting the dameons occurs when the Pacemaker init script is invoked. This resolves two long standing issues:

a. Forking inside a multi-threaded process like Corosync causes all sorts of pain. This has been problematic for Pacemaker as it needs a number of daemons to be spawned.

b. Corosync was never designed for staggered shutdown - something previously needed in order to prevent the cluster from leaving before Pacemaker could stop all active resources.

2.4.5. 传送配置文件

然后我们把配置文件拷贝到其他节点:

```
# for f in /etc/corosync/corosync.conf /etc/corosync/service.d/pcmk /etc/hosts; do scp $f pcmk-2:$f ; done
corosync.conf                                 100% 1528     1.5KB/s   00:00
hosts                                         100%  281     0.3KB/s   00:00
#
```

检验集群的安装

目录

3.1. 检验Corosync的安装

在第一个节点启动Corosync:

```
# /etc/init.d/corosync start
Starting Corosync Cluster Engine (corosync): [ OK ]
```

查看集群是否正确启动并且已经可以与其他节点建立集群关系

```
# grep -e "corosync.*network interface" -e "Corosync Cluster Engine" -e "Successfully read main configuration
 file" /var/log/messages
Aug 27 09:05:34 pcmk-1 corosync[1540]: [MAIN ] Corosync Cluster Engine ('1.1.0'): started and ready to provide
 service.
Aug 27 09:05:34 pcmk-1 corosync[1540]: [MAIN ] Successfully read main configuration file '/etc/corosync/
corosync.conf'.
# grep TOTEM /var/log/messages
Aug 27 09:05:34 pcmk-1 corosync[1540]: [TOTEM ] Initializing transport (UDP/IP).
Aug 27 09:05:34 pcmk-1 corosync[1540]: [TOTEM ] Initializing transmit/receive security: libtomcrypt SOBER128/
SHA1HMAC (mode 0).
Aug 27 09:05:35 pcmk-1 corosync[1540]: [TOTEM ] The network interface [192.168.122.101] is now up.
Aug 27 09:05:35 pcmk-1 corosync[1540]: [TOTEM ] A processor joined or left the membership and a new membership
 was formed.
```

With one node functional, it's now safe to start Corosync on the second node as well.

```
# ssh pcmk-2 -- /etc/init.d/corosync start
Starting Corosync Cluster Engine (corosync): [ OK ]
#
```

检查集群关系有没有正确建立:

```
# grep TOTEM /var/log/messages
Aug 27 09:05:34 pcmk-1 corosync[1540]: [TOTEM ] Initializing transport (UDP/IP).
Aug 27 09:05:34 pcmk-1 corosync[1540]: [TOTEM ] Initializing transmit/receive security: libtomcrypt SOBER128/
SHA1HMAC (mode 0).
Aug 27 09:05:35 pcmk-1 corosync[1540]: [TOTEM ] The network interface [192.168.122.101] is now up.
Aug 27 09:05:35 pcmk-1 corosync[1540]: [TOTEM ] A processor joined or left the membership and a new membership
 was formed.
Aug 27 09:12:11 pcmk-1 corosync[1540]: [TOTEM ] A processor joined or left the membership and a new membership
 was formed.
```

3.2. 检查Pacemaker的安装

现在我们已经确认Corosync正常，我们可以开始检查其他部分是否正常.

```
# grep pcmk_startup /var/log/messages
Aug 27 09:05:35 pcmk-1 corosync[1540]:  [pcmk ] info: pcmk_startup: CRM: InitializedAug 27 09:05:35 pcmk-1
 corosync[1540]:  [pcmk ] Logging: Initialized pcmk_startup
```

```
Aug 27 09:05:35 pcmk-1 corosync[1540]:  [pcmk ] info: pcmk_startup: Maximum core file size is:
 18446744073709551615
Aug 27 09:05:35 pcmk-1 corosync[1540]:  [pcmk ] info: pcmk_startup: Service: 9Aug 27 09:05:35 pcmk-1
 corosync[1540]:  [pcmk ] info: pcmk_startup: Local hostname: pcmk-1
```

Now try starting Pacemaker and check the necessary processes have been started

```
# /etc/init.d/pacemaker start
Starting Pacemaker Cluster Manager: [ OK ]

# grep -e pacemakerd.*get_config_opt -e pacemakerd.*start_child -e "Starting Pacemaker" /var/log/messages
Feb  8 13:31:24 pcmk-1 pacemakerd: [13155]: info: get_config_opt: Found 'pacemaker' for option: name
Feb  8 13:31:24 pcmk-1 pacemakerd: [13155]: info: get_config_opt: Found '1' for option: ver
Feb  8 13:31:24 pcmk-1 pacemakerd: [13155]: info: get_config_opt: Defaulting to 'no' for option: use_logd
Feb  8 13:31:24 pcmk-1 pacemakerd: [13155]: info: get_config_opt: Defaulting to 'no' for option: use_mgmtd
Feb  8 13:31:24 pcmk-1 pacemakerd: [13155]: info: get_config_opt: Found 'on' for option: debug
Feb  8 13:31:24 pcmk-1 pacemakerd: [13155]: info: get_config_opt: Found 'yes' for option: to_logfile
Feb  8 13:31:24 pcmk-1 pacemakerd: [13155]: info: get_config_opt: Found '/var/log/corosync.log' for option:
 logfile
Feb  8 13:31:24 pcmk-1 pacemakerd: [13155]: info: get_config_opt: Found 'yes' for option: to_syslog
Feb  8 13:31:24 pcmk-1 pacemakerd: [13155]: info: get_config_opt: Found 'daemon' for option: syslog_facility
Feb  8 16:50:38 pcmk-1 pacemakerd: [13990]: info: main: Starting Pacemaker 1.1.5 (Build: 31f088949239+):
 docbook-manpages publican ncurses trace-logging cman cs-quorum heartbeat corosync snmp libesmtp
Feb  8 16:50:38 pcmk-1 pacemakerd: [13990]: info: start_child: Forked child 14022 for process stonith-ng
Feb  8 16:50:38 pcmk-1 pacemakerd: [13990]: info: start_child: Forked child 14023 for process cib
Feb  8 16:50:38 pcmk-1 pacemakerd: [13990]: info: start_child: Forked child 14024 for process lrmd
Feb  8 16:50:38 pcmk-1 pacemakerd: [13990]: info: start_child: Forked child 14025 for process attrd
Feb  8 16:50:38 pcmk-1 pacemakerd: [13990]: info: start_child: Forked child 14026 for process pengine
Feb  8 16:50:38 pcmk-1 pacemakerd: [13990]: info: start_child: Forked child 14027 for process crmd

# ps axf PID TTY   STAT  TIME COMMAND
  2 ?    S<   0:00 [kthreadd]
  3 ?    S<   0:00 \_ [migration/0]
... lots of processes ...
13990 ?  S    0:01 pacemakerd
14022 ?  Sa   0:00 \_ /usr/lib64/heartbeat/stonithd
14023 ?  Sa   0:00 \_ /usr/lib64/heartbeat/cib
14024 ?  Sa   0:00 \_ /usr/lib64/heartbeat/lrmd
14025 ?  Sa   0:00 \_ /usr/lib64/heartbeat/attrd
14026 ?  Sa   0:00 \_ /usr/lib64/heartbeat/pengine
14027 ?  Sa   0:00 \_ /usr/lib64/heartbeat/crmd
```

Next, check for any ERRORs during startup - there shouldn't be any.

```
# grep ERROR: /var/log/messages | grep -v unpack_resources
#
```

Repeat on the other node and display the cluster's status.

```
# ssh pcmk-2 -- /etc/init.d/pacemaker start
Starting Pacemaker Cluster Manager: [ OK ]
# crm_mon
============
Last updated: Thu Aug 27 16:54:55 2009Stack: openais
Current DC: pcmk-1 - partition with quorum
Version: 1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f
2 Nodes configured, 2 expected votes
0 Resources configured.
============
Online: [ pcmk-1 pcmk-2 ]
```

Pacemaker Tools

目录

4.1. 使用Pacemaker工具

在万恶的旧社会，配置Pacemaker需要管理员具备读写XML的能力。 根据UNIX精神，也有许多不同的查询和配置集群的命令。

自从Pacemaker 1.0，这一切都改变了，我们有了一个集成的脚本化的集群控制shell,它把麻烦的XML配置隐藏了起来。它甚至允许你一次做出许多修改并自动提交(并检测是否合法)。

让我们花点时间熟悉一下它能做什么。

```
# crm --help
```

```
usage:
    crm [-D display_type]
    crm [-D display_type] args
    crm [-D display_type] [-f file]

    Use crm without arguments for an interactive session.
    Supply one or more arguments for a "single-shot" use.
    Specify with -f a file which contains a script. Use '-' for
    standard input or use pipe/redirection.

    crm displays cli format configurations using a color scheme
    and/or in uppercase. Pick one of "color" or "uppercase", or
    use "-D color,uppercase" if you want colorful uppercase.
    Get plain output by "-D plain". The default may be set in
    user preferences (options).

Examples:

    # crm -f stopapp2.cli
    # crm < stopapp2.cli
    # crm resource stop global_www
    # crm status
```

The primary tool for monitoring the status of the cluster is crm_mon (also available as crm status). It can be run in a variety of modes and has a number of output options. To find out about any of the tools that come with Pacemaker, simply invoke them with the --help option or consult the included man pages. Both sets of output are created from the tool, and so will always be in sync with each other and the tool itself.

Additionally, the Pacemaker version and supported cluster stack(s) are available via the --feature option to pacemakerd.

```
# pacemakerd --features
```

```
Pacemaker 1.1.9-3.fc20.2 (Build: 781a388)
 Supporting v3.0.7:  generated-manpages agent-manpages ncurses libqb-logging libqb-ipc upstart systemd nagios
 corosync-native
```

pacemakerd --help

```
pacemakerd - Start/Stop Pacemaker

Usage: pacemakerd mode [options]
Options:
 -?, --help   This text
 -$, --version   Version information
 -V, --verbose   Increase debug output
 -S, --shutdown   Instruct Pacemaker to shutdown on this machine
 -F, --features   Display the full version and list of features Pacemaker was built with

Additional Options:
 -f, --foreground   (Ignored) Pacemaker always runs in the foreground
 -p, --pid-file=value  (Ignored) Daemon pid file location

Report bugs to pacemaker@oss.clusterlabs.org
```

crm_mon --help

```
crm_mon - Provides a summary of cluster's current state.

Outputs varying levels of detail in a number of different formats.

Usage: crm_mon mode [options]
Options:
 -?, --help   This text
 -$, --version   Version information
 -V, --verbose   Increase debug output
 -Q, --quiet   Display only essential output

Modes:
 -h, --as-html=value Write cluster status to the named html file
 -X, --as-xml   Write cluster status as xml to stdout. This will enable one-shot mode.
 -w, --web-cgi   Web mode with output suitable for cgi
 -s, --simple-status  Display the cluster status once as a simple one line output (suitable for nagios)

Display Options:
 -n, --group-by-node   Group resources by node
 -r, --inactive   Display inactive resources
 -f, --failcounts   Display resource fail counts
 -o, --operations   Display resource operation history
 -t, --timing-details   Display resource operation history with timing details
 -c, --tickets    Display cluster tickets
 -W, --watch-fencing    Listen for fencing events. For use with --external-agent, --mail-to and/or --snmp-traps
 where supported
 -A, --show-node-attributes  Display node attributes

Additional Options:
 -i, --interval=value  Update frequency in seconds
 -1, --one-shot   Display the cluster status once on the console and exit
 -N, --disable-ncurses   Disable the use of ncurses
 -d, --daemonize   Run in the background as a daemon
 -p, --pid-file=value  (Advanced) Daemon pid file location
 -E, --external-agent=value A program to run when resource operations take place.
 -e, --external-recipient=value A recipient for your program (assuming you want the program to send something
 to someone).

Examples:

Display the cluster status on the console with updates as they occur:

 # crm_mon

Display the cluster status on the console just once then exit:
```

```
 # crm_mon -1

Display your cluster status, group resources by node, and include inactive resources in the list:

 # crm_mon --group-by-node --inactive

Start crm_mon as a background daemon and have it write the cluster status to an HTML file:

 # crm_mon --daemonize --as-html /path/to/docroot/filename.html

Start crm_mon and export the current cluster status as xml to stdout, then exit.:

 # crm_mon --as-xml


Report bugs to pacemaker@oss.clusterlabs.org
```

注意

如果SNMP或者email选项没有出现在选项中，说明pacemaker编译的时候没有打开对他们的支持，你需要联系提供这个发行版本的人，或者自己编译。

创建一个主/备集群

目录

5.1. 浏览现有配置

当Pacemaker启动的时候，它会自动记录节点的数量和详细信息，以及基层软件(本文中是corosync)和Pacemaker的版本。

这是初始配置文件的模样：

```
# crm configure show
node pcmk-1
node pcmk-2
property $id="cib-bootstrap-options" \
    dc-version="1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f" \
    cluster-infrastructure="openais" \
    expected-quorum-votes="2"
```

For those that are not of afraid of XML, you can see the raw configuration by appending "xml" to the previous command.

这是本文档最后一次显示XML。(作者怨念很深啊)

```
# crm configure show xml
<?xml version="1.0" ?>
<cib admin_epoch="0" crm_feature_set="3.0.1" dc-uuid="pcmk-1" epoch="13" have-quorum="1" num_updates="7"
 validate-with="pacemaker-1.0">
 <configuration>
  <crm_config>
   <cluster_property_set id="cib-bootstrap-options">
    <nvpair id="cib-bootstrap-options-dc-version" name="dc-version" value="1.1.5-
bdd89e69ba545404d02445be1f3d72e6a203ba2f"/>
    <nvpair id="cib-bootstrap-options-cluster-infrastructure" name="cluster-infrastructure" value="openais"/>
    <nvpair id="cib-bootstrap-options-expected-quorum-votes" name="expected-quorum-votes" value="2"/>
   </cluster_property_set>
  </crm_config>
  <rsc_defaults/>
  <op_defaults/>
  <nodes>
   <node id="pcmk-1" type="normal" uname="pcmk-1"/>
   <node id="pcmk-2" type="normal" uname="pcmk-2"/>
  </nodes>
  <resources/>
  <constraints/>
 </configuration>
</cib>
```

在我们做出任何改变之前，我们最好检查下配置文件。

```
# crm_verify -L
crm_verify[2195]: 2009/08/27_16:57:12 ERROR: unpack_resources: Resource start-up disabled since no STONITH
 resources have been defined
crm_verify[2195]: 2009/08/27_16:57:12 ERROR: unpack_resources: Either configure some or disable STONITH with
 the stonith-enabled option
crm_verify[2195]: 2009/08/27_16:57:12 ERROR: unpack_resources: NOTE: Clusters with shared data need STONITH to
 ensure data integrity
Errors found during check: config not valid -V may provide more details
#
```

就像你看到的，这个工具发现了一些错误。

In order to guarantee the safety of your data [1] , Pacemaker ships with STONITH [2] enabled. However it also knows when no STONITH configuration has been supplied and reports this as a problem (since the cluster would not be able to make progress if a situation requiring node fencing arose).

目前，我们禁用这个特性，然后在 配置STONISH 章节来配置它。这里要指出，使用STONITH是非常有必要的。关闭这个特性就是告诉集群:假装故障的节点已经安全的关机了。一些供应商甚至不允许这个特性被关闭。

我们将 stonith-enabled设置为 false 来关闭STONITH

```
# crm configure property stonith-enabled=false
# crm_verify -L
```

设置完这个选项以后，校验配置文件就正常了。

警告

The use of stonith-enabled=false is completely inappropriate for a production cluster. We use it here to defer the discussion of its configuration which can differ widely from one installation to the next. See *# 9.1 # “What Is STONITH”* for information on why STONITH is important and details on how to configure it.

5.2. 添加一个资源

首先要做的是配置一个IP地址，不管集群服务在哪运行，我们要一个固定的地址来提供服务。在这里我选择192.168.122.101作为浮动IP，给它取一个好记的名字 ClusterIP 并且告诉集群 每30秒检查它一次

重要

选择的IP地址不能被节点所占用

```
# crm configure primitive ClusterIP ocf:heartbeat:IPaddr2 \
```

[1] If the data is corrupt, there is little point in continuing to make it available

[2] A common node fencing mechanism. Used to ensure data integrity by powering off "bad" nodes

```
    params ip=192.168.122.101 cidr_netmask=32 \
    op monitor interval=30s
```

The other important piece of information here is ocf:heartbeat:IPaddr2.

This tells Pacemaker three things about the resource you want to add. The first field, ocf, is the standard to which the resource script conforms to and where to find it. The second field is specific to OCF resources and tells the cluster which namespace to find the resource script in, in this case heartbeat. The last field indicates the name of the resource script.

可以运行下面的命令来获得可用的资源类

```
# crm ra classesheartbeat
lsb ocf / heartbeat pacemakerstonith
```

找到OCF中Pacemaker和Heartbeat提供的资源脚本，运行下面的命令

```
# crm ra list ocf pacemaker
ClusterMon   Dummy     Stateful    SysInfo    SystemHealth  controld
ping      pingd
# crm ra list ocf heartbeat
AoEtarget       AudibleAlarm      ClusterMon       Delay
Dummy          EvmsSCC         Evmsd          Filesystem
ICP           IPaddr          IPaddr2         IPsrcaddr
LVM           LinuxSCSI        MailTo          ManageRAID
ManageVE         Pure-FTPd        Raid1          Route
SAPDatabase       SAPInstance       SendArp         ServeRAID
SphinxSearchDaemon   Squid          Stateful         SysInfo
VIPArip         VirtualDomain     WAS           WAS6
WinPopup         Xen           Xinetd          anything
apache          db2           drbd           eDir88
iSCSILogicalUnit    iSCSITarget      ids           iscsi
ldirectord       mysql          mysql-proxy      nfsserver
oracle          oralsnr         pgsql          pingd
portblock        rsyncd          scsi2reservation    sfex
tomcat          vmware
#
```

现在检查下IP 资源是不是已经添加了，并且看看是否处在可用状态。

```
# crm configure shownode pcmk-1
node pcmk-2primitive ClusterIP ocf:heartbeat:IPaddr2 \
    params ip="192.168.122.101" cidr_netmask="32" \
    op monitor interval="30s"
property $id="cib-bootstrap-options" \
    dc-version="1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f" \
    cluster-infrastructure="openais" \
    expected-quorum-votes="2" \
    stonith-enabled="false" \
# crm_mon
============
Last updated: Fri Aug 28 15:23:48 2009
Stack: openais
Current DC: pcmk-1 - partition with quorum
Version: 1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f
2 Nodes configured, 2 expected votes
1 Resources configured.
============

Online: [ pcmk-1 pcmk-2 ]
ClusterIP (ocf::heartbeat:IPaddr): Started pcmk-1
```

5.3. 做一次失效备援

作为一个高可用的集群，我们在继续本文档之前，我们要需要测试失效备援 。

首先，找到IP资源现在在哪个节点上运行。

```
# crm resource status ClusterIP
resource ClusterIP is running on: pcmk-1
#
```

Shut down Pacemaker and Corosync on that machine.

```
# ssh pcmk-1 -- /etc/init.d/pacemaker stop
Signaling Pacemaker Cluster Manager to terminate: [ OK ]
Waiting for cluster services to unload:. [ OK ]
# ssh pcmk-1 -- /etc/init.d/corosync stop
Stopping Corosync Cluster Engine (corosync): [ OK ]
Waiting for services to unload: [ OK ]
#
```

当Corosync停止运行以后，我们到另外一个节点用crm_mon来检查集群状态.

```
# crm_mon
============
Last updated: Fri Aug 28 15:27:35 2009
Stack: openais
Current DC: pcmk-2 - partition WITHOUT quorum
Version: 1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f
2 Nodes configured, 2 expected votes
1 Resources configured.
============

Online: [ pcmk-2 ]OFFLINE: [ pcmk-1 ]
```

关于集群状态，我们有三个地方需要注意，首先，如我们所料pcmk-1已经下线了，然而我们发现ClusterIP不在任何地方运行！

5.3.1. 法定人数和双节点集群

This is because the cluster no longer has quorum, as can be seen by the text "partition WITHOUT quorum" (emphasised green) in the output above. In order to reduce the possibility of data corruption, Pacemaker's default behavior is to stop all resources if the cluster does not have quorum.

当有半数以上的节点在线时，这个集群就认为自己拥有法定人数了，是“合法”的，换而言之就是下面的公式:

```
total_nodes < 2 * active_nodes
```

Therefore a two-node cluster only has quorum when both nodes are running, which is no longer the case for our cluster. This would normally make the creation of a two-node cluster pointless [3] , however it is possible to control how Pacemaker behaves when quorum is lost. In particular, we can tell the cluster to simply ignore quorum altogether.

[3] Actually some would argue that two-node clusters are always pointless, but that is an argument for another time

```
# crm configure property no-quorum-policy=ignore
# crm configure show
node pcmk-1
node pcmk-2
primitive ClusterIP ocf:heartbeat:IPaddr2 \
    params ip="192.168.122.101" cidr_netmask="32" \
    op monitor interval="30s"
property $id="cib-bootstrap-options" \
    dc-version="1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f" \
    cluster-infrastructure="openais" \
    expected-quorum-votes="2" \
    stonith-enabled="false" \
    no-quorum-policy="ignore"
```

过了一会，集群会在剩下的那个节点上启动这个IP。请注意集群现在依然没有达到法定人数。

```
# crm_mon
============
Last updated: Fri Aug 28 15:30:18 2009
Stack: openais
Current DC: pcmk-2 - partition WITHOUT quorum
Version: 1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f
2 Nodes configured, 2 expected votes
1 Resources configured.
============
Online: [ pcmk-2 ]
OFFLINE: [ pcmk-1 ]
ClusterIP (ocf::heartbeat:IPaddr): Started pcmk-2
```

现在模拟节点恢复，我们启动 pcmk-1 上面的Corosync服务，然后检查集群状态。

```
# /etc/init.d/corosync start
Starting Corosync Cluster Engine (corosync): [ OK ]
# /etc/init.d/pacemaker start
Starting Pacemaker Cluster Manager: [ OK ]# crm_mon
============
Last updated: Fri Aug 28 15:32:13 2009
Stack: openais
Current DC: pcmk-2 - partition with quorum
Version: 1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f
2 Nodes configured, 2 expected votes
1 Resources configured.
============
Online: [ pcmk-1 pcmk-2 ]

ClusterIP    (ocf::heartbeat:IPaddr):    Started pcmk-1
```

现在我们可以看到让某些人惊奇的事情，IP资源回到原来那个节点(pcmk-1)上去了。

5.3.2. 防止资源在节点恢复后移动

In some circumstances, it is highly desirable to prevent healthy resources from being moved around the cluster. Moving resources almost always requires a period of downtime. For complex services like Oracle databases, this period can be quite long.

To address this, Pacemaker has the concept of resource stickiness which controls how much a service prefers to stay running where it is. You may like to think of it as the "cost" of any downtime. By default, Pacemaker assumes there is zero cost associated with

moving resources and will do so to achieve "optimal" [4] resource placement. We can specify a different stickiness for every resource, but it is often sufficient to change the default.

```
# crm configure rsc_defaults resource-stickiness=100
# crm configure show
node pcmk-1
node pcmk-2
primitive ClusterIP ocf:heartbeat:IPaddr2 \
    params ip="192.168.122.101" cidr_netmask="32" \
    op monitor interval="30s"
property $id="cib-bootstrap-options" \
    dc-version="1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f" \
    cluster-infrastructure="openais" \
    expected-quorum-votes="2" \
    stonith-enabled="false" \
    no-quorum-policy="ignore"rsc_defaults $id="rsc-options" \
    resource-stickiness="100"
```

现在我们重新尝试失效援备测试，我们可以看到，正如我们所料，当pcmk-1不在线的时候ClusterIP还是移动到了pcmk-2

```
# ssh pcmk-1 -- /etc/init.d/pacemaker stop
Signaling Pacemaker Cluster Manager to terminate:          [  OK  ]
Waiting for cluster services to unload:.                   [  OK  ]
# ssh pcmk-1 -- /etc/init.d/corosync stop
Stopping Corosync Cluster Engine (corosync):        [ OK ]
Waiting for services to unload:              [ OK ]
# ssh pcmk-2 -- crm_mon -1
============
Last updated: Fri Aug 28 15:39:38 2009
Stack: openais
Current DC: pcmk-2 - partition WITHOUT quorum
Version: 1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f
2 Nodes configured, 2 expected votes
1 Resources configured.
============

Online: [ pcmk-2 ]
OFFLINE: [ pcmk-1 ]
ClusterIP    (ocf::heartbeat:IPaddr):    Started pcmk-2
```

但是当我们把pcmk-1恢复在线后，ClusterIP现在还是跑在pcmk-2上面。

```
# /etc/init.d/corosync start
Starting Corosync Cluster Engine (corosync): [ OK ]
# /etc/init.d/pacemaker start
Starting Pacemaker Cluster Manager: [ OK ]
# crm_mon
============
Last updated: Fri Aug 28 15:41:23 2009
Stack: openais
Current DC: pcmk-2 - partition with quorum
Version: 1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f
2 Nodes configured, 2 expected votes
1 Resources configured.
============
```

[4] It should be noted that Pacemaker's definition of optimal may not always agree with that of a human's. The order in which Pacemaker processes lists of resources and nodes creates implicit preferences in situations where the administrator has not explicitly specified them

```
Online: [ pcmk-1 pcmk-2 ]

ClusterIP    (ocf::heartbeat:IPaddr):    Started pcmk-2
```

Apache - 添加更多的服务

目录

6.1. Forward

现在我们有了一个基本的但是功能齐全的双节点集群,我们已经可以往里面加些真的服务了。我们准备启动一个Apache服务，因为它是许多集群的主角，并且相对来说比较容易配置。

6.2. 安装Apache

Before continuing, we need to make sure Apache is installed on both hosts.

```
# yum install -y httpdSetting up Install Process
Resolving Dependencies
--> Running transaction check
---> Package httpd.x86_64 0:2.2.13-2.fc12 set to be updated
--> Processing Dependency: httpd-tools = 2.2.13-2.fc12 for package: httpd-2.2.13-2.fc12.x86_64
--> Processing Dependency: apr-util-ldap for package: httpd-2.2.13-2.fc12.x86_64
--> Processing Dependency: /etc/mime.types for package: httpd-2.2.13-2.fc12.x86_64
--> Processing Dependency: libaprutil-1.so.0()(64bit) for package: httpd-2.2.13-2.fc12.x86_64
--> Processing Dependency: libapr-1.so.0()(64bit) for package: httpd-2.2.13-2.fc12.x86_64
--> Running transaction check
---> Package apr.x86_64 0:1.3.9-2.fc12 set to be updated
---> Package apr-util.x86_64 0:1.3.9-2.fc12 set to be updated
---> Package apr-util-ldap.x86_64 0:1.3.9-2.fc12 set to be updated
---> Package httpd-tools.x86_64 0:2.2.13-2.fc12 set to be updated
---> Package mailcap.noarch 0:2.1.30-1.fc12 set to be updated
--> Finished Dependency Resolution

Dependencies Resolved

================================================================================
Package        Arch       Version        Repository     Size
================================================================================
Installing:
httpd        x86_64      2.2.13-2.fc12      rawhide      735 k
Installing for dependencies:
apr          x86_64      1.3.9-2.fc12       rawhide      117 k
apr-util     x86_64      1.3.9-2.fc12       rawhide       84 k
apr-util-ldap    x86_64      1.3.9-2.fc12       rawhide       15 k
httpd-tools      x86_64      2.2.13-2.fc12      rawhide       63 k
mailcap      noarch      2.1.30-1.fc12      rawhide       25 k

Transaction Summary
```

```
================================================================================
Install     6 Package(s)
Upgrade     0 Package(s)

Total download size: 1.0 M
Downloading Packages:
(1/6): apr-1.3.9-2.fc12.x86_64.rpm                 | 117 kB     00:00
(2/6): apr-util-1.3.9-2.fc12.x86_64.rpm            |  84 kB     00:00
(3/6): apr-util-ldap-1.3.9-2.fc12.x86_64.rpm       |  15 kB     00:00
(4/6): httpd-2.2.13-2.fc12.x86_64.rpm              | 735 kB     00:00
(5/6): httpd-tools-2.2.13-2.fc12.x86_64.rpm        |  63 kB     00:00
(6/6): mailcap-2.1.30-1.fc12.noarch.rpm            |  25 kB     00:00
--------------------------------------------------------------------------------
Total                           875 kB/s | 1.0 MB     00:01
Running rpm_check_debug
Running Transaction Test
Finished Transaction Test
Transaction Test Succeeded
Running Transaction
  Installing     : apr-1.3.9-2.fc12.x86_64                      1/6
  Installing     : apr-util-1.3.9-2.fc12.x86_64                 2/6
  Installing     : apr-util-ldap-1.3.9-2.fc12.x86_64            3/6
  Installing     : httpd-tools-2.2.13-2.fc12.x86_64             4/6
  Installing     : mailcap-2.1.30-1.fc12.noarch                 5/6
  Installing     : httpd-2.2.13-2.fc12.x86_64                   6/6

Installed:
  httpd.x86_64 0:2.2.13-2.fc12

Dependency Installed:
  apr.x86_64 0:1.3.9-2.fc12       apr-util.x86_64 0:1.3.9-2.fc12
  apr-util-ldap.x86_64 0:1.3.9-2.fc12 httpd-tools.x86_64 0:2.2.13-2.fc12
  mailcap.noarch 0:2.1.30-1.fc12

Complete!
```

同样的，为了检测Apache服务器，我们要安装wget这个工具。

```
# yum install -y wgetSetting up Install Process
Resolving Dependencies
--> Running transaction check
---> Package wget.x86_64 0:1.11.4-5.fc12 set to be updated
--> Finished Dependency Resolution

Dependencies Resolved

================================================================================
Package     Arch        Version           Repository        Size
================================================================================
Installing:
wget        x86_64      1.11.4-5.fc12         rawhide        393 k

Transaction Summary
================================================================================
Install     1 Package(s)
Upgrade     0 Package(s)

Total download size: 393 k
Downloading Packages:
wget-1.11.4-5.fc12.x86_64.rpm                      | 393 kB     00:00
Running rpm_check_debug
Running Transaction Test
Finished Transaction Test
Transaction Test Succeeded
Running Transaction
  Installing     : wget-1.11.4-5.fc12.x86_64                     1/1
```

```
Installed:
 wget.x86_64 0:1.11.4-5.fc12

Complete!
```

6.3. 准备工作

首先我们为Apache创建一个主页。在Fedora上面默认的Apache docroot是/var/www/html，所以我们在这个目录下面建立一个主页。

```
[root@pcmk-1 ~]# cat <<-END >/var/www/html/index.html <html>
 <body>My Test Site - pcmk-1</body>
 </html>
 END
```

为了方便，我们简化所用的页面并人工地在两个节点直接同步数据。所以在pcmk-2上面运行这个命令。

```
[root@pcmk-2 ~]# cat <<-END >/var/www/html/index.html <html>
 <body>My Test Site - pcmk-2</body>
 </html>
 END
```

6.4. 开启 Apache status URL

为了监控Apache实例的健康状态,并在它挂掉的时候恢复Apache服务，资源agent会假设 server-status URL是可用的。查看/etc/httpd/conf/httpd.conf并确保下面的选项没有被禁用或注释掉。

```
<Location /server-status>
   SetHandler server-status
   Order deny,allow
   Deny from all
   Allow from 127.0.0.1
</Location>
```

6.5. 更新配置文件

At this point, Apache is ready to go, all that needs to be done is to add it to the cluster. Lets call the resource WebSite. We need to use an OCF script called apache in the heartbeat namespace [1] , the only required parameter is the path to the main Apache configuration file and we'll tell the cluster to check once a minute that apache is still running.

```
# crm configure primitive WebSite ocf:heartbeat:apache params configfile=/etc/httpd/conf/httpd.conf op monitor
 interval=1min
# crm configure show
node pcmk-1
node pcmk-2primitive WebSite ocf:heartbeat:apache \ params configfile="/etc/httpd/conf/httpd.conf" \ op
 monitor interval="1min"primitive ClusterIP ocf:heartbeat:IPaddr2 \
    params ip="192.168.122.101" cidr_netmask="32" \
    op monitor interval="30s"
```

[1] Compare the key used here ocf:heartbeat:apache with the one we used earlier for the IP address: ocf:heartbeat:IPaddr2

```
property $id="cib-bootstrap-options" \
    dc-version="1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f" \
    cluster-infrastructure="openais" \
    expected-quorum-votes="2" \
    stonith-enabled="false" \
    no-quorum-policy="ignore"
rsc_defaults $id="rsc-options" \
    resource-stickiness="100"
```

过了一会，我们可以看到集群把apache启动起来了。

```
# crm_mon
============
Last updated: Fri Aug 28 16:12:49 2009
Stack: openais
Current DC: pcmk-2 - partition with quorum
Version: 1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f
2 Nodes configured, 2 expected votes
2 Resources configured.
============

Online: [ pcmk-1 pcmk-2 ]

ClusterIP    (ocf::heartbeat:IPaddr):    Started pcmk-2
WebSite    (ocf::heartbeat:apache):    Started pcmk-1
```

等等！WebSite这个资源跟IP没有跑在同一个节点上面！

6.6. 确保资源在同一个节点运行

To reduce the load on any one machine, Pacemaker will generally try to spread the configured resources across the cluster nodes. However we can tell the cluster that two resources are related and need to run on the same host (or not at all). Here we instruct the cluster that WebSite can only run on the host that ClusterIP is active on.

For the constraint, we need a name (choose something descriptive like website-with-ip), indicate that its mandatory (so that if ClusterIP is not active anywhere, WebSite will not be permitted to run anywhere either) by specifying a score of INFINITY and finally list the two resources.

注意

If ClusterIP is not active anywhere, WebSite will not be permitted to run anywhere.

重要

Colocation constraints are "directional", in that they imply certain things about the order in which the two resources will have a location chosen. In this case we're saying WebSite needs to be placed on the same machine as ClusterIP, this implies that we must know the location of ClusterIP before choosing a location for WebSite.

```
# crm configure colocation website-with-ip INFINITY: WebSite ClusterIP
# crm configure show
node pcmk-1
node pcmk-2
primitive WebSite ocf:heartbeat:apache \
    params configfile="/etc/httpd/conf/httpd.conf" \
    op monitor interval="1min"
primitive ClusterIP ocf:heartbeat:IPaddr2 \
    params ip="192.168.122.101" cidr_netmask="32" \
    op monitor interval="30s"colocation website-with-ip inf: WebSite ClusterIPproperty $id="cib-bootstrap-
options" \
    dc-version="1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f" \
    cluster-infrastructure="openais" \
    expected-quorum-votes="2" \
    stonith-enabled="false" \
    no-quorum-policy="ignore"
rsc_defaults $id="rsc-options" \
    resource-stickiness="100"
# crm_mon
============
Last updated: Fri Aug 28 16:14:34 2009
Stack: openais
Current DC: pcmk-2 - partition with quorum
Version: 1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f
2 Nodes configured, 2 expected votes
2 Resources configured.
============

Online: [ pcmk-1 pcmk-2 ]

ClusterIP    (ocf::heartbeat:IPaddr):    Started pcmk-2
WebSite    (ocf::heartbeat:apache):    Started pcmk-2
```

6.7. 控制资源的启动停止顺序

When Apache starts, it binds to the available IP addresses. It doesn't know about any addresses we add afterwards, so not only do they need to run on the same node, but we need to make sure ClusterIP is already active before we start WebSite. We do this by adding an ordering constraint. We need to give it a name (choose something descriptive like apache-after-ip), indicate that its mandatory (so that any recovery for ClusterIP will also trigger recovery of WebSite) and list the two resources in the order we need them to start.

```
# crm configure order apache-after-ip mandatory: ClusterIP WebSite
# crm configure show
node pcmk-1
node pcmk-2
primitive WebSite ocf:heartbeat:apache \
    params configfile="/etc/httpd/conf/httpd.conf" \
    op monitor interval="1min"
primitive ClusterIP ocf:heartbeat:IPaddr2 \
    params ip="192.168.122.101" cidr_netmask="32" \
    op monitor interval="30s"
colocation website-with-ip inf: WebSite ClusterIPorder apache-after-ip inf: ClusterIP WebSiteproperty $id="cib-
bootstrap-options" \
    dc-version="1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f" \
    cluster-infrastructure="openais" \
    expected-quorum-votes="2" \
    stonith-enabled="false" \
    no-quorum-policy="ignore"
rsc_defaults $id="rsc-options" \
    resource-stickiness="100"
```

6.8. 指定优先的 Location

Pacemaker 并不要求你机器的硬件配置是相同的，可能某些机器比另外的机器配置要好。这种状况下我们会希望设置:当某个节点可用时，资源就要跑在上面之类的规则。为了达到这个效果我们创建location约束。同样的，我们给他取一个描述性的名字(prefer-pcmk-1)，指明我们想在上面跑WebSite这个服务，多想在上面跑(我们现在指定分值为50，但是在双节点的集群状态下，任何大于0的值都可以达到想要的效果)，以及目标节点的名字:

```
# crm configure location prefer-pcmk-1 WebSite 50: pcmk-1
# crm configure show
node pcmk-1
node pcmk-2
primitive WebSite ocf:heartbeat:apache \
    params configfile="/etc/httpd/conf/httpd.conf" \
    op monitor interval="1min"
primitive ClusterIP ocf:heartbeat:IPaddr2 \
    params ip="192.168.122.101" cidr_netmask="32" \
    op monitor interval="30s"location prefer-pcmk-1 WebSite 50: pcmk-1colocation website-with-ip inf: WebSite
 ClusterIP
property $id="cib-bootstrap-options" \
    dc-version="1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f" \
    cluster-infrastructure="openais" \
    expected-quorum-votes="2" \
    stonith-enabled="false" \
    no-quorum-policy="ignore"
rsc_defaults $id="rsc-options" \
    resource-stickiness="100"
# crm_mon
============
Last updated: Fri Aug 28 16:17:35 2009
Stack: openais
Current DC: pcmk-2 - partition with quorum
Version: 1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f
2 Nodes configured, 2 expected votes
2 Resources configured.
============

Online: [ pcmk-1 pcmk-2 ]

ClusterIP    (ocf::heartbeat:IPaddr):    Started pcmk-2WebSite    (ocf::heartbeat:apache):    Started pcmk-2
```

等等，资源还是在pcmk-2上面跑的!

即使我们更希望资源在pcmk-1上面运行，但是 这个优先值还是比资源黏性值要小。

如果要看现在的分值，可以用ptest这个命令

```
ptest -sL
```

注意

Include output There is a way to force them to move though…

6.9. 在集群中手工地移动资源

经常性的会有管理员想要无视集群然后强制把资源移动到指定的地方。 底层的操作就像我们上面创建的location约束一样。只要提供资源和目标地址，我们会补全剩余部分。

```
# crm resource move WebSite pcmk-1
# crm_mon
============
Last updated: Fri Aug 28 16:19:24 2009
Stack: openais
Current DC: pcmk-2 - partition with quorum
Version: 1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f
2 Nodes configured, 2 expected votes
2 Resources configured.
============

Online: [ pcmk-1 pcmk-2 ]

ClusterIP    (ocf::heartbeat:IPaddr):    Started pcmk-1
WebSite    (ocf::heartbeat:apache):    Started pcmk-1
```

Notice how the colocation rule we created has ensured that ClusterIP was also moved to pcmk-1. For the curious, we can see the effect of this command by examining the configuration

```
# crm configure show
node pcmk-1
node pcmk-2
primitive WebSite ocf:heartbeat:apache \
    params configfile="/etc/httpd/conf/httpd.conf" \
    op monitor interval="1min"
primitive ClusterIP ocf:heartbeat:IPaddr2 \
    params ip="192.168.122.101" cidr_netmask="32" \
    op monitor interval="30s"
location cli-prefer-WebSite WebSite \
    rule $id="cli-prefer-rule-WebSite" inf: #uname eq pcmk-1
location prefer-pcmk-1 WebSite 50: pcmk-1
colocation website-with-ip inf: WebSite ClusterIP
property $id="cib-bootstrap-options" \
    dc-version="1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f" \
    cluster-infrastructure="openais" \
    expected-quorum-votes="2" \
    stonith-enabled="false" \
    no-quorum-policy="ignore"
rsc_defaults $id="rsc-options" \
    resource-stickiness="100"
```

斜体部分是用来移动资源到pcmk-1约束，它是自动生成的。

6.9.1. 把控制权交还给集群

当我们完成那些要求要资源移动到pcmk-1的操作——在我们的例子里面啥都没干 ——我们可以用unmove命令把集群恢复到强制移动前的状态。因为我们之前配置了默认的资源黏性值，恢复了以后资源还是会在pcmk-1上面。

```
# crm resource unmove WebSite
# crm configure show
node pcmk-1
node pcmk-2
primitive WebSite ocf:heartbeat:apache \
    params configfile="/etc/httpd/conf/httpd.conf" \
    op monitor interval="1min"
primitive ClusterIP ocf:heartbeat:IPaddr2 \
    params ip="192.168.122.101" cidr_netmask="32" \
    op monitor interval="30s"
location prefer-pcmk-1 WebSite 50: pcmk-1
colocation website-with-ip inf: WebSite ClusterIP
```

```
property $id="cib-bootstrap-options" \
    dc-version="1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f" \
    cluster-infrastructure="openais" \
    expected-quorum-votes="2" \
    stonith-enabled="false" \
    no-quorum-policy="ignore"
rsc_defaults $id="rsc-options" \
    resource-stickiness="100"
```

可以看到自动生成的约束已经没有了。如果我们查看集群的状态，我们也可以看到就如我们所预期的，资源还是在pcmk-1上面跑

```
# crm_mon
============
Last updated: Fri Aug 28 16:20:53 2009
Stack: openais
Current DC: pcmk-2 - partition with quorum
Version: 1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f
2 Nodes configured, 2 expected votes
2 Resources configured.
============

Online: [ pcmk-1 pcmk-2 ]

 ClusterIP    (ocf::heartbeat:IPaddr):    Started pcmk-1
 WebSite    (ocf::heartbeat:apache):    Started pcmk-1
```

用DRBD同步存储

目录

7.1. Background

Even if you're serving up static websites, having to manually synchronize the contents of that website to all the machines in the cluster is not ideal. For dynamic websites, such as a wiki, it's not even an option. Not everyone care afford network-attached storage but somehow the data needs to be kept in sync. Enter DRBD which can be thought of as network based RAID-1. See *http://www.drbd.org/* for more details.

7.2. 安装DRBD软件包

Since its inclusion in the upstream 2.6.33 kernel, everything needed to use DRBD ships with Fedora 13. All you need to do is install it:

```
# yum install -y drbd-pacemaker drbd-udev
Loaded plugins: presto, refresh-packagekit
Setting up Install Process
Resolving Dependencies
--> Running transaction check
---> Package drbd-pacemaker.x86_64 0:8.3.7-2.fc13 set to be updated
--> Processing Dependency: drbd-utils = 8.3.7-2.fc13 for package: drbd-pacemaker-8.3.7-2.fc13.x86_64
--> Running transaction check
---> Package drbd-utils.x86_64 0:8.3.7-2.fc13 set to be updated
--> Finished Dependency Resolution

Dependencies Resolved

================================================================================
 Package                Arch           Version              Repository      Size
================================================================================
Installing:
 drbd-pacemaker         x86_64         8.3.7-2.fc13         fedora          19 k
Installing for dependencies:
 drbd-utils             x86_64         8.3.7-2.fc13         fedora         165 k

Transaction Summary
================================================================================
Install       2 Package(s)
Upgrade       0 Package(s)

Total download size: 184 k
Installed size: 427 k
```

```
Downloading Packages:
Setting up and reading Presto delta metadata
fedora/prestodelta                                       | 1.7 kB     00:00
Processing delta metadata
Package(s) data still to download: 184 k
(1/2): drbd-pacemaker-8.3.7-2.fc13.x86_64.rpm            |  19 kB     00:01
(2/2): drbd-utils-8.3.7-2.fc13.x86_64.rpm                | 165 kB     00:02
 ---------------------------------------------------------------------------
Total                                           45 kB/s | 184 kB     00:04
Running rpm_check_debug
Running Transaction Test
Transaction Test Succeeded
Running Transaction
  Installing     : drbd-utils-8.3.7-2.fc13.x86_64                          1/2
  Installing     : drbd-pacemaker-8.3.7-2.fc13.x86_64                      2/2

Installed:
  drbd-pacemaker.x86_64 0:8.3.7-2.fc13

Dependency Installed:
  drbd-utils.x86_64 0:8.3.7-2.fc13

Complete!
```

7.3. 配置DRBD

在我们设置之前，我们要创建一些空的磁盘分区给它。

7.3.1. 为DRBD创建一个分区

如果你有1Gb以上的空间，就用那么多吧， 在这个指南中根本用不到这么多空间。

```
# lvcreate -n drbd-demo -L 1G VolGroup
Logical volume "drbd-demo" created
# lvs
LV    VG    Attr  LSize  Origin Snap% Move Log Copy% Convert
drbd-demo VolGroup -wi-a- 1.00G
lv_root  VolGroup -wi-ao  7.30G
lv_swap  VolGroup -wi-ao 500.00M
```

在另外一个节点上面执行相同的操作，请确保使用了相同大小的分区。

```
# ssh pcmk-2 -- lvs
LV   VG    Attr  LSize  Origin Snap% Move Log Copy% Convert
lv_root VolGroup -wi-ao  7.30G
lv_swap VolGroup -wi-ao 500.00M
# ssh pcmk-2 -- lvcreate -n drbd-demo -L 1G VolGroup
Logical volume "drbd-demo" created
# ssh pcmk-2 -- lvs
LV    VG    Attr  LSize  Origin Snap% Move Log Copy% Convert
drbd-demo VolGroup -wi-a- 1.00G
lv_root  VolGroup -wi-ao  7.30G
lv_swap  VolGroup -wi-ao 500.00M
```

7.3.2. 配置DRBD

There is no series of commands for building a DRBD configuration, so simply copy the configuration below to /etc/drbd.conf

想知道配置文件的详细信息，请访问 *http://www.drbd.org/users-guide/ch-configure.html*

警告

Be sure to use the names and addresses of your nodes if they differ from the ones used in this guide.

```
global {
 usage-count yes;
}
common {
 protocol C;
}
resource wwwdata {
 meta-disk internal;
 device  /dev/drbd1;
 syncer {
  verify-alg sha1;
 }
 net {
  allow-two-primaries;
 }
 on pcmk-1 {
  disk   /dev/mapper/VolGroup-drbd--demo;
  address  192.168.122.101:7789;
 }
 on pcmk-2 {
  disk   /dev/mapper/VolGroup-drbd--demo;
  address  192.168.122.102:7789;
 }
}
```

注意

TODO: Explain the reason for the allow-two-primaries option

7.3.3. 初始化并载入DRBD

配置完成以后，我们可以来执行初始化了

```
# drbdadm create-md wwwdata
md_offset 12578816
al_offset 12546048
bm_offset 12541952

Found some data
==> This might destroy existing data! <==

Do you want to proceed?
[need to type 'yes' to confirm] yes
Writing meta data...
initializing activity log
NOT initialized bitmap
New drbd meta data block successfully created.
success
```

现在讲DRBD的模块载入内核并检测是不是都正常

```
# modprobe drbd# drbdadm up wwwdata# cat /proc/drbdversion: 8.3.6 (api:88/proto:86-90)
GIT-hash: f3606c47cc6fcf6b3f086e425cb34af8b7a81bbf build by root@pcmk-1, 2009-12-08 11:22:57
 1: cs:WFConnection ro:Secondary/Unknown ds:Inconsistent/DUnknown C r----
  ns:0 nr:0 dw:0 dr:0 a1:0 bm:0 lo:0 pe:0 ua:0 ap:0 ep:1 wo:b oos:12248
```

Repeat on the second node

```
# ssh pcmk-2 -- drbdadm --force create-md wwwdata
Writing meta data...
initializing activity log
NOT initialized bitmap
New drbd meta data block successfully created.
success
# ssh pcmk-2 -- modprobe drbd
WARNING: Deprecated config file /etc/modprobe.conf, all config files belong into /etc/modprobe.d/.
# ssh pcmk-2 -- drbdadm up wwwdata
# ssh pcmk-2 -- cat /proc/drbd
version: 8.3.6 (api:88/proto:86-90)
GIT-hash: f3606c47cc6fcf6b3f086e425cb34af8b7a81bbf build by root@pcmk-1, 2009-12-08 11:22:57
 1: cs:Connected ro:Secondary/Secondary ds:Inconsistent/Inconsistent C r----
    ns:0 nr:0 dw:0 dr:0 a1:0 bm:0 lo:0 pe:0 ua:0 ap:0 ep:1 wo:b oos:12248
```

现在我们要告诉DRBD要用那个数据(那个节点作为主)。因为两边都有一些废数据，我们要在pcmk-1上面执行一下命令。

```
# drbdadm -- --overwrite-data-of-peer primary wwwdata
# cat /proc/drbd
version: 8.3.6 (api:88/proto:86-90)
GIT-hash: f3606c47cc6fcf6b3f086e425cb34af8b7a81bbf build by root@pcmk-1, 2009-12-08 11:22:57
1: cs:SyncSource ro:Primary/Secondary ds:UpToDate/Inconsistent C r----
   ns:2184 nr:0 dw:0 dr:2472 a1:0 bm:0 lo:0 pe:0 ua:0 ap:0 ep:1 wo:b oos:10064
    [=====>.............] sync'ed: 33.4% (10064/12248)K
    finish: 0:00:37 speed: 240 (240) K/sec
# cat /proc/drbd
version: 8.3.6 (api:88/proto:86-90)
GIT-hash: f3606c47cc6fcf6b3f086e425cb34af8b7a81bbf build by root@pcmk-1, 2009-12-08 11:22:57
1: cs:Connected ro:Primary/Secondary ds:UpToDate/UpToDate C r----
   ns:12248 nr:0 dw:0 dr:12536 a1:0 bm:1 lo:0 pe:0 ua:0 ap:0 ep:1 wo:b oos:0
```

pcmk-1 is now in the Primary state which allows it to be written to. Which means it's a good point at which to create a filesystem and populate it with some data to serve up via our WebSite resource.

7.3.4. 向DRBD中添加数据

```
# mkfs.ext4 /dev/drbd1
mke2fs 1.41.4 (27-Jan-2009)
Filesystem label=
OS type: Linux
Block size=1024 (log=0)
Fragment size=1024 (log=0)
3072 inodes, 12248 blocks
612 blocks (5.00%) reserved for the super user
First data block=1
Maximum filesystem blocks=12582912
2 block groups
8192 blocks per group, 8192 fragments per group
1536 inodes per group
```

```
Superblock backups stored on blocks:
    8193

Writing inode tables: done
Creating journal (1024 blocks): done
Writing superblocks and filesystem accounting information: done

This filesystem will be automatically checked every 26 mounts or
180 days, whichever comes first. Use tune2fs -c or -i to override.
```

Now mount the newly created filesystem so we can create our index file

```
# mount /dev/drbd1 /mnt/
# cat <<-END >/mnt/index.html
 <html>
  <body>My Test Site - drbd</body>
 </html>
 END
# umount /dev/drbd1
```

7.4. 在集群中配置DRBD

crm shell一个便捷的特性是可以工作在交互模式下并自动的变更配置中的相关部分。

首先我们打开shell。提示会指出你现在是在交互模式下。

```
# crm cib
crm(live) #
```

Next we must create a working copy of the current configuration. This is where all our changes will go. The cluster will not see any of them until we say it's ok. Notice again how the prompt changes, this time to indicate that we're no longer looking at the live cluster.

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

现在我们可以创建DRBD clone,然后看看修改过后的配置文件。

```
crm(drbd) # configure primitive WebData ocf:linbit:drbd params drbd_resource=wwwdata \
    op monitor interval=60s
crm(drbd) # configure ms WebDataClone WebData meta master-max=1 master-node-max=1 \
    clone-max=2 clone-node-max=1 notify=truecrm(drbd) # configure shownode pcmk-1
node pcmk-2primitive WebData ocf:linbit:drbd \
 params drbd_resource="wwwdata" \
 op monitor interval="60s"primitive WebSite ocf:heartbeat:apache \
    params configfile="/etc/httpd/conf/httpd.conf" \
    op monitor interval="1min"
primitive ClusterIP ocf:heartbeat:IPaddr2 \
    params ip="192.168.122.101" cidr_netmask="32" \
    op monitor interval="30s"ms WebDataClone WebData \
 meta master-max="1" master-node-max="1" clone-max="2" clone-node-max="1" notify="true"
location prefer-pcmk-1 WebSite 50: pcmk-1
colocation website-with-ip inf: WebSite ClusterIP
order apache-after-ip inf: ClusterIP WebSite
property $id="cib-bootstrap-options" \
    dc-version="1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f" \
    cluster-infrastructure="openais" \
    expected-quorum-votes="2" \
```

```
    stonith-enabled="false" \
    no-quorum-policy="ignore"
rsc_defaults $id="rsc-options" \
    resource-stickiness="100"
```

一旦你确认这些修改没问题，我们就提交这个副本，然后用crm_mon来看看修改是否生效了。

```
crm(drbd) # cib commit drbdINFO: commited 'drbd' shadow CIB to the cluster
crm(drbd) # quitbye
# crm_mon
============
Last updated: Tue Sep 1 09:37:13 2009
Stack: openais
Current DC: pcmk-1 - partition with quorum
Version: 1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f
2 Nodes configured, 2 expected votes
3 Resources configured.
============

Online: [ pcmk-1 pcmk-2 ]

ClusterIP    (ocf::heartbeat:IPaddr):    Started pcmk-1
WebSite (ocf::heartbeat:apache):    Started pcmk-1Master/Slave Set: WebDataClone Masters: [ pcmk-2 ] Slaves: [
 pcmk-1 ]
```

注意

Include details on adding a second DRBD resource

现在DRBD已经工作了，我们可以配置一个Filesystem资源来使用它。 此外，对于这个文件系统的定义，同样的我们告诉集群这个文件系统能在哪运行(主DRBD运行的节点)以及什么时候可以启动(在主DRBD启动以后)。

我们再一次的使用交互模式的crm shell

```
# crm
crm(live) # cib new fs
INFO: fs shadow CIB created
crm(fs) # configure primitive WebFS ocf:heartbeat:Filesystem \
    params device="/dev/drbd/by-res/wwwdata" directory="/var/www/html" fstype="ext4"
crm(fs) # configure colocation fs_on_drbd inf: WebFS WebDataClone:Master
crm(fs) # configure order WebFS-after-WebData inf: WebDataClone:promote WebFS:start

We also need to tell the cluster that Apache needs to run on the same
machine as the filesystem and that it must be active before Apache can
start.

crm(fs) # configure colocation WebSite-with-WebFS inf: WebSite WebFS
crm(fs) # configure order WebSite-after-WebFS inf: WebFS WebSite
```

审视一下你的配置：

```
crm(fs) # crm configure show
node pcmk-1
node pcmk-2
primitive WebData ocf:linbit:drbd \
    params drbd_resource="wwwdata" \
```

```
    op monitor interval="60s"
primitive WebFS ocf:heartbeat:Filesystem \
    params device="/dev/drbd/by-res/wwwdata" directory="/var/www/html" fstype="ext4"
primitive WebSite ocf:heartbeat:apache \
    params configfile="/etc/httpd/conf/httpd.conf" \
    op monitor interval="1min"
primitive ClusterIP ocf:heartbeat:IPaddr2 \
    params ip="192.168.122.101" cidr_netmask="32" \
    op monitor interval="30s"
ms WebDataClone WebData \
    meta master-max="1" master-node-max="1" clone-max="2" clone-node-max="1" notify="true"
location prefer-pcmk-1 WebSite 50: pcmk-1
colocation WebSite-with-WebFS inf: WebSite WebFS
colocation fs_on_drbd inf: WebFS WebDataClone:Master
colocation website-with-ip inf: WebSite ClusterIP
order WebFS-after-WebData inf: WebDataClone:promote WebFS:start
order WebSite-after-WebFS inf: WebFS WebSite
order apache-after-ip inf: ClusterIP WebSite
property Sid="cib-bootstrap-options" \
    dc-version="1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f" \
    cluster-infrastructure="openais" \
    expected-quorum-votes="2" \
    stonith-enabled="false" \
    no-quorum-policy="ignore"
rsc_defaults Sid="rsc-options" \
    resource-stickiness="100"
```

看完以后，我们提交它并看看有没有生效。

```
crm(fs) # cib commit fs
INFO: commited 'fs' shadow CIB to the cluster
crm(fs) # quit
bye
# crm_mon
============
Last updated: Tue Sep 1 10:08:44 2009
Stack: openais
Current DC: pcmk-1 - partition with quorum
Version: 1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f
2 Nodes configured, 2 expected votes
4 Resources configured.
============

Online: [ pcmk-1 pcmk-2 ]

ClusterIP    (ocf::heartbeat:IPaddr):    Started pcmk-1
WebSite (ocf::heartbeat:apache): Started pcmk-1
Master/Slave Set: WebDataClone
    Masters: [ pcmk-1 ]
    Slaves: [ pcmk-2 ]
WebFS (ocf::heartbeat:Filesystem): Started pcmk-1
```

7.4.1. 迁移测试

We could shut down the active node again, but another way to safely simulate recovery is to put the node into what is called "standby mode". Nodes in this state tell the cluster that they are not allowed to run resources. Any resources found active there will be moved elsewhere. This feature can be particularly useful when updating the resources' packages.

把一个本地节点设置为standby模式并观察集群把所有资源移动到另外一个节点了。并且注意节点的状态改变为不能运行任何的资源。

```
# crm node standby
```

```
# crm_mon
============
Last updated: Tue Sep 1 10:09:57 2009
Stack: openais
Current DC: pcmk-1 - partition with quorum
Version: 1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f
2 Nodes configured, 2 expected votes
4 Resources configured.
============
Node pcmk-1: standbyOnline: [ pcmk-2 ]

ClusterIP    (ocf::heartbeat:IPaddr):    Started pcmk-2
WebSite (ocf::heartbeat:apache):    Started pcmk-2
Master/Slave Set: WebDataClone
    Masters: [ pcmk-2 ]    Stopped: [ WebData:1 ]
WebFS  (ocf::heartbeat:Filesystem):  Started pcmk-2
```

当我在pcmk-1上面操作完了——本例中没有任何操作，我们只是想让资源移动移动——我们可以让节点变回正常的集群成员。

```
# crm node online
# crm_mon
============
Last updated: Tue Sep 1 10:13:25 2009
Stack: openais
Current DC: pcmk-1 - partition with quorum
Version: 1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f
2 Nodes configured, 2 expected votes
4 Resources configured.
============
Online: [ pcmk-1 pcmk-2 ]
ClusterIP    (ocf::heartbeat:IPaddr):    Started pcmk-2
WebSite (ocf::heartbeat:apache):    Started pcmk-2
Master/Slave Set: WebDataClone
    Masters: [ pcmk-2 ]
    Slaves: [ pcmk-1 ]
WebFS  (ocf::heartbeat:Filesystem):  Started pcmk-2
```

注意我们设置的资源黏性值阻止了资源迁移回pcmk-1

转变为Active/Active

目录

8.1. 需求

The primary requirement for an Active/Active cluster is that the data required for your services is available, simultaneously, on both machines. Pacemaker makes no requirement on how this is achieved, you could use a SAN if you had one available, however since DRBD supports multiple Primaries, we can also use that.

The only hitch is that we need to use a cluster-aware filesystem. The one we used earlier with DRBD, ext4, is not one of those. Both OCFS2 and GFS2 are supported, however here we will use GFS2 which comes with Fedora.

We'll also need to use CMAN for Cluster Membership and Quorum instead of our Corosync plugin.

8.2. Adding CMAN Support

CMAN v3[1] is a Corosync plugin that monitors the names and number of active cluster nodes in order to deliver membership and quorum information to clients (such as the Pacemaker daemons).

In a traditional Corosync-Pacemaker cluster, a Pacemaker plugin is loaded to provide membership and quorum information. The motivation for wanting to use CMAN for this instead, is to ensure all elements of the cluster stack are making decisions based on the same membership and quorum data. [2]

In the case of GFS2, the key pieces are the dlm_controld and gfs_controld helpers which act as the glue between the filesystem and the cluster software. Supporting CMAN enables

[1] http://docs.redhat.com/docs/en-US/Red_Hat_Enterprise_Linux/6/html-single/Cluster_Suite_Overview/index.html#s2-clumembership-overview-CSO

[2] A failure to do this can lead to what is called *internal split-brain* - a situation where different parts of the stack disagree about whether some nodes are alive or dead - which quickly leads to unnecessary down-time and/or data corruption.

us to use the versions already being shipped by most distributions (since CMAN has been around longer than Pacemaker and is part of the Red Hat cluster stack).

警告

Ensure Corosync and Pacemaker are stopped on all nodes before continuing

警告

Be sure to disable the Pacemaker plugin before continuing with this section. In most cases, this can be achieved by removing /etc/corosync/service.d/pcmk and stopping Corosync.

8.2.1. Installing the required Software

```
# yum install -y cman gfs2-utils gfs2-cluster
Loaded plugins: auto-update-debuginfo
Setting up Install Process
Resolving Dependencies
--> Running transaction check
---> Package cman.x86_64 0:3.1.7-1.fc15 will be installed
--> Processing Dependency: modcluster >= 0.18.1-1 for package: cman-3.1.7-1.fc15.x86_64
--> Processing Dependency: fence-agents >= 3.1.5-1 for package: cman-3.1.7-1.fc15.x86_64
--> Processing Dependency: openais >= 1.1.4-1 for package: cman-3.1.7-1.fc15.x86_64
--> Processing Dependency: ricci >= 0.18.1-1 for package: cman-3.1.7-1.fc15.x86_64
--> Processing Dependency: libSaCkpt.so.3(OPENAIS_CKPT_B.01.01)(64bit) for package: cman-3.1.7-1.fc15.x86_64
--> Processing Dependency: libSaCkpt.so.3()(64bit) for package: cman-3.1.7-1.fc15.x86_64
---> Package gfs2-cluster.x86_64 0:3.1.1-2.fc15 will be installed
---> Package gfs2-utils.x86_64 0:3.1.1-2.fc15 will be installed
--> Running transaction check
---> Package fence-agents.x86_64 0:3.1.5-1.fc15 will be installed
--> Processing Dependency: /usr/bin/virsh for package: fence-agents-3.1.5-1.fc15.x86_64
--> Processing Dependency: net-snmp-utils for package: fence-agents-3.1.5-1.fc15.x86_64
--> Processing Dependency: sg3_utils for package: fence-agents-3.1.5-1.fc15.x86_64
--> Processing Dependency: perl(Net::Telnet) for package: fence-agents-3.1.5-1.fc15.x86_64
--> Processing Dependency: /usr/bin/ipmitool for package: fence-agents-3.1.5-1.fc15.x86_64
--> Processing Dependency: perl-Net-Telnet for package: fence-agents-3.1.5-1.fc15.x86_64
--> Processing Dependency: pexpect for package: fence-agents-3.1.5-1.fc15.x86_64
--> Processing Dependency: pyOpenSSL for package: fence-agents-3.1.5-1.fc15.x86_64
--> Processing Dependency: python-suds for package: fence-agents-3.1.5-1.fc15.x86_64
---> Package modcluster.x86_64 0:0.18.7-1.fc15 will be installed
--> Processing Dependency: oddjob for package: modcluster-0.18.7-1.fc15.x86_64
---> Package openais.x86_64 0:1.1.4-2.fc15 will be installed
---> Package openaislib.x86_64 0:1.1.4-2.fc15 will be installed
---> Package ricci.x86_64 0:0.18.7-1.fc15 will be installed
--> Processing Dependency: parted for package: ricci-0.18.7-1.fc15.x86_64
--> Processing Dependency: nss-tools for package: ricci-0.18.7-1.fc15.x86_64
--> Running transaction check
---> Package ipmitool.x86_64 0:1.8.11-6.fc15 will be installed
---> Package libvirt-client.x86_64 0:0.8.8-7.fc15 will be installed
--> Processing Dependency: libnetcf.so.1(NETCF_1.3.0)(64bit) for package: libvirt-client-0.8.8-7.fc15.x86_64
--> Processing Dependency: cyrus-sasl-md5 for package: libvirt-client-0.8.8-7.fc15.x86_64
--> Processing Dependency: gettext for package: libvirt-client-0.8.8-7.fc15.x86_64
--> Processing Dependency: nc for package: libvirt-client-0.8.8-7.fc15.x86_64
--> Processing Dependency: libnuma.so.1(libnuma_1.1)(64bit) for package: libvirt-client-0.8.8-7.fc15.x86_64
```

```
--> Processing Dependency: libnuma.so.1(libnuma_1.2)(64bit) for package: libvirt-client-0.8.8-7.fc15.x86_64
--> Processing Dependency: libnetcf.so.1(NETCF_1.2.0)(64bit) for package: libvirt-client-0.8.8-7.fc15.x86_64
--> Processing Dependency: gnutls-utils for package: libvirt-client-0.8.8-7.fc15.x86_64
--> Processing Dependency: libnetcf.so.1(NETCF_1.0.0)(64bit) for package: libvirt-client-0.8.8-7.fc15.x86_64
--> Processing Dependency: libxenstore.so.3.0()(64bit) for package: libvirt-client-0.8.8-7.fc15.x86_64
--> Processing Dependency: libyajl.so.1()(64bit) for package: libvirt-client-0.8.8-7.fc15.x86_64
--> Processing Dependency: libnl.so.1()(64bit) for package: libvirt-client-0.8.8-7.fc15.x86_64
--> Processing Dependency: libnuma.so.1()(64bit) for package: libvirt-client-0.8.8-7.fc15.x86_64
--> Processing Dependency: libaugeas.so.0()(64bit) for package: libvirt-client-0.8.8-7.fc15.x86_64
--> Processing Dependency: libnetcf.so.1()(64bit) for package: libvirt-client-0.8.8-7.fc15.x86_64
---> Package net-snmp-utils.x86_64 1:5.6.1-7.fc15 will be installed
---> Package nss-tools.x86_64 0:3.12.10-6.fc15 will be installed
---> Package oddjob.x86_64 0:0.31-2.fc15 will be installed
---> Package parted.x86_64 0:2.3-10.fc15 will be installed
---> Package perl-Net-Telnet.noarch 0:3.03-12.fc15 will be installed
---> Package pexpect.noarch 0:2.3-6.fc15 will be installed
---> Package pyOpenSSL.x86_64 0:0.10-3.fc15 will be installed
---> Package python-suds.noarch 0:0.3.9-3.fc15 will be installed
---> Package sg3_utils.x86_64 0:1.29-3.fc15 will be installed
--> Processing Dependency: sg3_utils-libs = 1.29-3.fc15 for package: sg3_utils-1.29-3.fc15.x86_64
--> Processing Dependency: libsgutils2.so.2()(64bit) for package: sg3_utils-1.29-3.fc15.x86_64
--> Running transaction check
---> Package augeas-libs.x86_64 0:0.9.0-1.fc15 will be installed
---> Package cyrus-sasl-md5.x86_64 0:2.1.23-18.fc15 will be installed
---> Package gettext.x86_64 0:0.18.1.1-7.fc15 will be installed
--> Processing Dependency: libgomp.so.1(GOMP_1.0)(64bit) for package: gettext-0.18.1.1-7.fc15.x86_64
--> Processing Dependency: libgettextlib-0.18.1.so()(64bit) for package: gettext-0.18.1.1-7.fc15.x86_64
--> Processing Dependency: libgettextsrc-0.18.1.so()(64bit) for package: gettext-0.18.1.1-7.fc15.x86_64
--> Processing Dependency: libgomp.so.1()(64bit) for package: gettext-0.18.1.1-7.fc15.x86_64
---> Package gnutls-utils.x86_64 0:2.10.5-1.fc15 will be installed
---> Package libnl.x86_64 0:1.1-14.fc15 will be installed
---> Package nc.x86_64 0:1.100-3.fc15 will be installed
--> Processing Dependency: libbsd.so.0(LIBBSD_0.0)(64bit) for package: nc-1.100-3.fc15.x86_64
--> Processing Dependency: libbsd.so.0(LIBBSD_0.2)(64bit) for package: nc-1.100-3.fc15.x86_64
--> Processing Dependency: libbsd.so.0()(64bit) for package: nc-1.100-3.fc15.x86_64
---> Package netcf-libs.x86_64 0:0.1.9-1.fc15 will be installed
---> Package numactl.x86_64 0:2.0.7-1.fc15 will be installed
---> Package sg3_utils-libs.x86_64 0:1.29-3.fc15 will be installed
---> Package xen-libs.x86_64 0:4.1.1-3.fc15 will be installed
--> Processing Dependency: xen-licenses for package: xen-libs-4.1.1-3.fc15.x86_64
---> Package yajl.x86_64 0:1.0.11-1.fc15 will be installed
--> Running transaction check
---> Package gettext-libs.x86_64 0:0.18.1.1-7.fc15 will be installed
---> Package libbsd.x86_64 0:0.2.0-4.fc15 will be installed
---> Package libgomp.x86_64 0:4.6.1-9.fc15 will be installed
---> Package xen-licenses.x86_64 0:4.1.1-3.fc15 will be installed
--> Finished Dependency Resolution

Dependencies Resolved

================================================================================
 Package              Arch        Version                 Repository     Size
================================================================================
Installing:
 cman                 x86_64      3.1.7-1.fc15            updates       366 k
 gfs2-cluster         x86_64      3.1.1-2.fc15            fedora         69 k
 gfs2-utils           x86_64      3.1.1-2.fc15            fedora        222 k
Installing for dependencies:
 augeas-libs          x86_64      0.9.0-1.fc15            updates       311 k
 cyrus-sasl-md5       x86_64      2.1.23-18.fc15          updates        46 k
 fence-agents         x86_64      3.1.5-1.fc15            updates       186 k
 gettext              x86_64      0.18.1.1-7.fc15         fedora        1.0 M
 gettext-libs         x86_64      0.18.1.1-7.fc15         fedora        610 k
 gnutls-utils         x86_64      2.10.5-1.fc15           fedora        101 k
 ipmitool             x86_64      1.8.11-6.fc15           fedora        273 k
 libbsd               x86_64      0.2.0-4.fc15            fedora         37 k
 libgomp              x86_64      4.6.1-9.fc15            updates        95 k
```

```
 libn1                x86_64      1.1-14.fc15            fedora       118 k
 libvirt-client       x86_64      0.8.8-7.fc15           updates      2.4 M
 modcluster           x86_64      0.18.7-1.fc15          fedora       187 k
 nc                   x86_64      1.100-3.fc15           updates       24 k
 net-snmp-utils       x86_64      1:5.6.1-7.fc15         fedora       180 k
 netcf-libs           x86_64      0.1.9-1.fc15           updates       50 k
 nss-tools            x86_64      3.12.10-6.fc15         updates      723 k
 numactl              x86_64      2.0.7-1.fc15           updates       54 k
 oddjob               x86_64      0.31-2.fc15            fedora        61 k
 openais              x86_64      1.1.4-2.fc15           fedora       190 k
 openaislib           x86_64      1.1.4-2.fc15           fedora        88 k
 parted               x86_64      2.3-10.fc15            updates      618 k
 perl-Net-Telnet      noarch      3.03-12.fc15           fedora        55 k
 pexpect              noarch      2.3-6.fc15             fedora       141 k
 pyOpenSSL            x86_64      0.10-3.fc15            fedora       198 k
 python-suds          noarch      0.3.9-3.fc15           fedora       195 k
 ricci                x86_64      0.18.7-1.fc15          fedora       584 k
 sg3_utils            x86_64      1.29-3.fc15            fedora       465 k
 sg3_utils-libs       x86_64      1.29-3.fc15            fedora        54 k
 xen-libs             x86_64      4.1.1-3.fc15           updates      310 k
 xen-licenses         x86_64      4.1.1-3.fc15           updates       64 k
 yajl                 x86_64      1.0.11-1.fc15          fedora        27 k

Transaction Summary
================================================================================
Install      34 Package(s)

Total download size: 10 M
Installed size: 38 M
Downloading Packages:
(1/34): augeas-libs-0.9.0-1.fc15.x86_64.rpm           | 311 kB     00:00
(2/34): cman-3.1.7-1.fc15.x86_64.rpm                  | 366 kB     00:00
(3/34): cyrus-sasl-md5-2.1.23-18.fc15.x86_64.rpm      |  46 kB     00:00
(4/34): fence-agents-3.1.5-1.fc15.x86_64.rpm          | 186 kB     00:00
(5/34): gettext-0.18.1.1-7.fc15.x86_64.rpm            | 1.0 MB     00:01
(6/34): gettext-libs-0.18.1.1-7.fc15.x86_64.rpm       | 610 kB     00:00
(7/34): gfs2-cluster-3.1.1-2.fc15.x86_64.rpm          |  69 kB     00:00
(8/34): gfs2-utils-3.1.1-2.fc15.x86_64.rpm            | 222 kB     00:00
(9/34): gnutls-utils-2.10.5-1.fc15.x86_64.rpm         | 101 kB     00:00
(10/34): ipmitool-1.8.11-6.fc15.x86_64.rpm            | 273 kB     00:00
(11/34): libbsd-0.2.0-4.fc15.x86_64.rpm               |  37 kB     00:00
(12/34): libgomp-4.6.1-9.fc15.x86_64.rpm              |  95 kB     00:00
(13/34): libn1-1.1-14.fc15.x86_64.rpm                 | 118 kB     00:00
(14/34): libvirt-client-0.8.8-7.fc15.x86_64.rpm       | 2.4 MB     00:01
(15/34): modcluster-0.18.7-1.fc15.x86_64.rpm          | 187 kB     00:00
(16/34): nc-1.100-3.fc15.x86_64.rpm                   |  24 kB     00:00
(17/34): net-snmp-utils-5.6.1-7.fc15.x86_64.rpm       | 180 kB     00:00
(18/34): netcf-libs-0.1.9-1.fc15.x86_64.rpm           |  50 kB     00:00
(19/34): nss-tools-3.12.10-6.fc15.x86_64.rpm          | 723 kB     00:00
(20/34): numactl-2.0.7-1.fc15.x86_64.rpm              |  54 kB     00:00
(21/34): oddjob-0.31-2.fc15.x86_64.rpm                |  61 kB     00:00
(22/34): openais-1.1.4-2.fc15.x86_64.rpm              | 190 kB     00:00
(23/34): openaislib-1.1.4-2.fc15.x86_64.rpm           |  88 kB     00:00
(24/34): parted-2.3-10.fc15.x86_64.rpm                | 618 kB     00:00
(25/34): perl-Net-Telnet-3.03-12.fc15.noarch.rpm      |  55 kB     00:00
(26/34): pexpect-2.3-6.fc15.noarch.rpm                | 141 kB     00:00
(27/34): pyOpenSSL-0.10-3.fc15.x86_64.rpm             | 198 kB     00:00
(28/34): python-suds-0.3.9-3.fc15.noarch.rpm          | 195 kB     00:00
(29/34): ricci-0.18.7-1.fc15.x86_64.rpm               | 584 kB     00:00
(30/34): sg3_utils-1.29-3.fc15.x86_64.rpm             | 465 kB     00:00
(31/34): sg3_utils-libs-1.29-3.fc15.x86_64.rpm        |  54 kB     00:00
(32/34): xen-libs-4.1.1-3.fc15.x86_64.rpm             | 310 kB     00:00
(33/34): xen-licenses-4.1.1-3.fc15.x86_64.rpm         |  64 kB     00:00
(34/34): yajl-1.0.11-1.fc15.x86_64.rpm                |  27 kB     00:00
 --------------------------------------------------------------------------------
Total                                       803 kB/s |  10 MB     00:12
Running rpm_check_debug
```

```
Running Transaction Test
Transaction Test Succeeded
Running Transaction
  Installing : openais-1.1.4-2.fc15.x86_64                                  1/34
  Installing : openaislib-1.1.4-2.fc15.x86_64                               2/34
  Installing : libnl-1.1-14.fc15.x86_64                                     3/34
  Installing : augeas-libs-0.9.0-1.fc15.x86_64                              4/34
  Installing : oddjob-0.31-2.fc15.x86_64                                    5/34
  Installing : modcluster-0.18.7-1.fc15.x86_64                              6/34
  Installing : netcf-libs-0.1.9-1.fc15.x86_64                               7/34
  Installing : 1:net-snmp-utils-5.6.1-7.fc15.x86_64                         8/34
  Installing : sg3_utils-libs-1.29-3.fc15.x86_64                            9/34
  Installing : sg3_utils-1.29-3.fc15.x86_64                                10/34
  Installing : libgomp-4.6.1-9.fc15.x86_64                                 11/34
  Installing : gnutls-utils-2.10.5-1.fc15.x86_64                           12/34
  Installing : pyOpenSSL-0.10-3.fc15.x86_64                                13/34
  Installing : parted-2.3-10.fc15.x86_64                                   14/34
  Installing : cyrus-sasl-md5-2.1.23-18.fc15.x86_64                        15/34
  Installing : python-suds-0.3.9-3.fc15.noarch                             16/34
  Installing : ipmitool-1.8.11-6.fc15.x86_64                               17/34
  Installing : perl-Net-Telnet-3.03-12.fc15.noarch                         18/34
  Installing : numactl-2.0.7-1.fc15.x86_64                                 19/34
  Installing : yajl-1.0.11-1.fc15.x86_64                                   20/34
  Installing : gettext-libs-0.18.1.1-7.fc15.x86_64                         21/34
  Installing : gettext-0.18.1.1-7.fc15.x86_64                              22/34
  Installing : libbsd-0.2.0-4.fc15.x86_64                                  23/34
  Installing : nc-1.100-3.fc15.x86_64                                      24/34
  Installing : xen-licenses-4.1.1-3.fc15.x86_64                            25/34
  Installing : xen-libs-4.1.1-3.fc15.x86_64                                26/34
  Installing : libvirt-client-0.8.8-7.fc15.x86_64                          27/34

Note: This output shows SysV services only and does not include native
      systemd services. SysV configuration data might be overridden by native
      systemd configuration.

  Installing : nss-tools-3.12.10-6.fc15.x86_64                             28/34
  Installing : ricci-0.18.7-1.fc15.x86_64                                  29/34
  Installing : pexpect-2.3-6.fc15.noarch                                   30/34
  Installing : fence-agents-3.1.5-1.fc15.x86_64                            31/34
  Installing : cman-3.1.7-1.fc15.x86_64                                    32/34
  Installing : gfs2-cluster-3.1.1-2.fc15.x86_64                            33/34
  Installing : gfs2-utils-3.1.1-2.fc15.x86_64                              34/34

Installed:
  cman.x86_64 0:3.1.7-1.fc15           gfs2-cluster.x86_64 0:3.1.1-2.fc15
  gfs2-utils.x86_64 0:3.1.1-2.fc15

Dependency Installed:
  augeas-libs.x86_64 0:0.9.0-1.fc15
  cyrus-sasl-md5.x86_64 0:2.1.23-18.fc15
  fence-agents.x86_64 0:3.1.5-1.fc15
  gettext.x86_64 0:0.18.1.1-7.fc15
  gettext-libs.x86_64 0:0.18.1.1-7.fc15
  gnutls-utils.x86_64 0:2.10.5-1.fc15
  ipmitool.x86_64 0:1.8.11-6.fc15
  libbsd.x86_64 0:0.2.0-4.fc15
  libgomp.x86_64 0:4.6.1-9.fc15
  libnl.x86_64 0:1.1-14.fc15
  libvirt-client.x86_64 0:0.8.8-7.fc15
  modcluster.x86_64 0:0.18.7-1.fc15
  nc.x86_64 0:1.100-3.fc15
  net-snmp-utils.x86_64 1:5.6.1-7.fc15
  netcf-libs.x86_64 0:0.1.9-1.fc15
  nss-tools.x86_64 0:3.12.10-6.fc15
  numactl.x86_64 0:2.0.7-1.fc15
  oddjob.x86_64 0:0.31-2.fc15
  openais.x86_64 0:1.1.4-2.fc15
```

```
  openaislib.x86_64 0:1.1.4-2.fc15
  parted.x86_64 0:2.3-10.fc15
  perl-Net-Telnet.noarch 0:3.03-12.fc15
  pexpect.noarch 0:2.3-6.fc15
  pyOpenSSL.x86_64 0:0.10-3.fc15
  python-suds.noarch 0:0.3.9-3.fc15
  ricci.x86_64 0:0.18.7-1.fc15
  sg3_utils.x86_64 0:1.29-3.fc15
  sg3_utils-libs.x86_64 0:1.29-3.fc15
  xen-libs.x86_64 0:4.1.1-3.fc15
  xen-licenses.x86_64 0:4.1.1-3.fc15
  yajl.x86_64 0:1.0.11-1.fc15

Complete!
```

8.2.2. Configuring CMAN

注意

The standard Pacemaker config file will continue to be used for resource management even after we start using CMAN. There is no need to recreate all your resources and constraints to the *cluster.conf* syntax, we simply create a minimal version that lists the nodes.

The first thing we need to do, is tell CMAN complete starting up even without quorum. We can do this by changing the quorum timeout setting:

```
# sed -i.sed "s/.*CMAN_QUORUM_TIMEOUT=.*/CMAN_QUORUM_TIMEOUT=0/g" /etc/sysconfig/cman
```

Next we create a basic configuration file and place it in /etc/cluster/cluster.conf. The name used for each clusternode should correspond to that node's uname -n, just as Pacemaker expects. The nodeid can be any positive mumber but must be unique.

Basic cluster.conf for a two-node cluster

```
<?xml version="1.0"?>
<cluster config_version="1" name="my_cluster_name">
  <logging debug="off"/>
  <clusternodes>
    <clusternode name="pcmk-1" nodeid="1"/>
    <clusternode name="pcmk-2" nodeid="2"/>
  </clusternodes>
</cluster>
```

8.2.3. Redundant Rings

For those wishing to use Corosync's multiple rings feature, simply define an alternate name for each node. For example:

```
    <clusternode name="pcmk-1" nodeid="1"/>
        <altname name="pcmk-1-internal"/>
    </clusternode>
```

8.2.4. Configuring CMAN Fencing

We configure the fence_pcmk agent (supplied with Pacemaker) to redirect any fencing requests from CMAN components (such as dlm_controld) to Pacemaker. Pacemaker's fencing subsystem lets other parts of the stack know that a node has been successfully fenced, thus avoiding the need for it to be fenced again when other subsystems notice the node is gone.

警告

Configuring real fencing devices in CMAN will result in nodes being fenced multiple times as different parts of the stack notice the node is missing or failed.

The definition should be placed in the fencedevices section and contain:

```
<fencedevice name="pcmk" agent="fence_pcmk"/>
```

Each clusternode must be configured to use this device by adding a fence method block that lists the node's name as the port.

```
<fence>
  <method name="pcmk-redirect">
    <device name="pcmk" port="node_name_here"/>
  </method>
</fence>
```

Putting everything together, we have:

cluster.conf for a two-node cluster with fencing

```
<?xml version="1.0"?>
<cluster config_version="1" name="mycluster">
  <logging debug="off"/>
  <clusternodes>
    <clusternode name="pcmk-1" nodeid="1">
      <fence>
        <method name="pcmk-redirect">
          <device name="pcmk" port="pcmk-1"/>
        </method>
      </fence>
    </clusternode>
    <clusternode name="pcmk-2" nodeid="2">
      <fence>
        <method name="pcmk-redirect">
          <device name="pcmk" port="pcmk-2"/>
        </method>
      </fence>
    </clusternode>
  </clusternodes>
  <fencedevices>
    <fencedevice name="pcmk" agent="fence_pcmk"/>
  </fencedevices>
</cluster>
```

8.2.5. Bringing the Cluster Online with CMAN

The first thing to do is check that the configuration is valid

```
# ccs_config_validate
Configuration validates
```

Now start CMAN

```
# service cman start
Starting cluster:
   Checking Network Manager...                             [  OK  ]
   Global setup...                                         [  OK  ]
   Loading kernel modules...                               [  OK  ]
   Mounting configfs...                                    [  OK  ]
   Starting cman...                                        [  OK  ]
   Waiting for quorum...                                   [  OK  ]
   Starting fenced...                                      [  OK  ]
   Starting dlm_controld...                                [  OK  ]
   Starting gfs_controld...                                [  OK  ]
   Unfencing self...                                       [  OK  ]
   Joining fence domain...                                 [  OK  ]
```

Once you have confirmed that the first node is happily online, start the second node.

```
[root@pcmk-2 ~]# service cman start
Starting cluster:
   Checking Network Manager...                             [  OK  ]
   Global setup...                                         [  OK  ]
   Loading kernel modules...                               [  OK  ]
   Mounting configfs...                                    [  OK  ]
   Starting cman...                                        [  OK  ]
   Waiting for quorum...                                   [  OK  ]
   Starting fenced...                                      [  OK  ]
   Starting dlm_controld...                                [  OK  ]
   Starting gfs_controld...                                [  OK  ]
   Unfencing self...                                       [  OK  ]
   Joining fence domain...                                 [  OK  ]
# cman_tool nodes
Node  Sts   Inc   Joined               Name
   1   M    548   2011-09-28 10:52:21  pcmk-1
   2   M    548   2011-09-28 10:52:21  pcmk-2
```

You should now see both nodes online. To begin managing resources, simply start Pacemaker.

```
# service pacemaker start
 Starting Pacemaker Cluster Manager: [  OK  ]
```

and again on the second node, after which point you can use crm_mon as you normally would.

```
[root@pcmk-2 ~]# service pacemaker start
 Starting Pacemaker Cluster Manager: [  OK  ]
# crm_mon -1
```

8.3. 创建一个 GFS2 文件系统

8.3.1. 准备工作

Before we do anything to the existing partition, we need to make sure it is unmounted. We do this by telling the cluster to stop the WebFS resource. This will ensure that other resources (in our case, Apache) using WebFS are not only stopped, but stopped in the correct order.

```
# crm_resource --resource WebFS --set-parameter target-role --meta --parameter-value Stopped
# crm_mon
============
Last updated: Thu Sep 3 15:18:06 2009
Stack: openais
Current DC: pcmk-1 - partition with quorum
Version: 1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f
2 Nodes configured, 2 expected votes
6 Resources configured.
============

Online: [ pcmk-1 pcmk-2 ]

Master/Slave Set: WebDataClone
    Masters: [ pcmk-1 ]
    Slaves: [ pcmk-2 ]
ClusterIP    (ocf::heartbeat:IPaddr):    Started pcmk-1
```

注意

注意 Apache and WebFS 两者都已经停止了。

8.3.2. 创建并迁移数据到 GFS2 分区

现在集群的基层和集成部分都正常运行，我们现在创建一个GFS2分区

警告

这个操作会清除DRBD分区上面的所有数据，请备份重要的数据。

我们要为GFS2分区指定一系列附加的参数。

首先我们要用 -p选项来指定我们用的是内核的DLM，然后我们用-j来表示我们为两个日志保留足够的空间(每个操作文件系统的节点各一个)。

最后，我们用-t来指定lock table的名称。这个字段的格式是 clustername:fsname(集群名称:文件系统名称)。fsname的话，我们只要用一个唯一的并且能描述我们这个集群的名称就好了，我们用默认的pcmk。

To specify an alternate name for the cluster, locate the service section containing name: pacemaker in corosync.conf and insert the following line anywhere inside the block:

```
clustername: myname
```

在每个节点都执行以下命令。

```
# mkfs.gfs2 -p lock_dlm -j 2 -t pcmk:web /dev/drbd1
This will destroy any data on /dev/drbd1.
It appears to contain: data

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

Device:          /dev/drbd1
Blocksize:         4096
Device Size        1.00 GB (131072 blocks)
Filesystem Size:      1.00 GB (131070 blocks)
Journals:         2
Resource Groups:      2
Locking Protocol:     "lock_dlm"
Lock Table:        "pcmk:web"
UUID:           6B776F46-177B-BAF8-2C2B-292C0E078613
```

然后再迁移数据到这个新的文件系统。现在我们创建一个跟上次不一样的主页。

```
# mount /dev/drbd1 /mnt/# cat <<-END >/mnt/index.html
<html>
<body>My Test Site - GFS2</body>
</html>
END
# umount /dev/drbd1
# drbdadm verify wwwdata#
```

8.4. 8.5. 重新为集群配置GFS2

```
# crm
crm(live) # cib new GFS2
INFO: GFS2 shadow CIB created
crm(GFS2) # configure delete WebFS
crm(GFS2) # configure primitive WebFS ocf:heartbeat:Filesystem params device="/dev/drbd/by-res/wwwdata"
 directory="/var/www/html" fstype="gfs2"
```

现在我们重新创建这个资源， 我们也要重建跟这个资源相关的约束条件，因为shell会自动删除跟WebFS相关的约束条件。

```
crm(GFS2) # configure colocation WebSite-with-WebFS inf: WebSite WebFS
crm(GFS2) # configure colocation fs_on_drbd inf: WebFS WebDataClone:Master
crm(GFS2) # configure order WebFS-after-WebData inf: WebDataClone:promote WebFS:start
crm(GFS2) # configure order WebSite-after-WebFS inf: WebFS WebSite
crm(GFS2) # configure show
node pcmk-1
node pcmk-2
primitive WebData ocf:linbit:drbd \
    params drbd_resource="wwwdata" \
    op monitor interval="60s"
primitive WebFS ocf:heartbeat:Filesystem \
    params device="/dev/drbd/by-res/wwwdata" directory="/var/www/html" fstype="gfs2"
primitive WebSite ocf:heartbeat:apache \
    params configfile="/etc/httpd/conf/httpd.conf" \
    op monitor interval="1min"
primitive ClusterIP ocf:heartbeat:IPaddr2 \
    params ip="192.168.122.101" cidr_netmask="32" \
```

```
    op monitor interval="30s"
ms WebDataClone WebData \
    meta master-max="1" master-node-max="1" clone-max="2" clone-node-max="1" notify="true"
colocation WebSite-with-WebFS inf: WebSite WebFS
colocation fs_on_drbd inf: WebFS WebDataClone:Master
colocation website-with-ip inf: WebSite ClusterIP
order WebFS-after-WebData inf: WebDataClone:promote WebFS:start
order WebSite-after-WebFS inf: WebFS WebSite
order apache-after-ip inf: ClusterIP WebSite
property $id="cib-bootstrap-options" \
    dc-version="1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f" \
    cluster-infrastructure="openais" \
    expected-quorum-votes="2" \
    stonith-enabled="false" \
    no-quorum-policy="ignore"
rsc_defaults $id="rsc-options" \
    resource-stickiness="100"
```

看看配置文件有没有错误，然后退出she11看看集群的反应。

```
crm(GFS2) # cib commit GFS2
INFO: commited 'GFS2' shadow CIB to the cluster
crm(GFS2) # quit
bye
# crm_mon
============
Last updated: Thu Sep 3 20:49:54 2009
Stack: openais
Current DC: pcmk-2 - partition with quorum
Version: 1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f
2 Nodes configured, 2 expected votes
6 Resources configured.
============

Online: [ pcmk-1 pcmk-2 ]

WebSite (ocf::heartbeat:apache):    Started pcmk-2
Master/Slave Set: WebDataClone
    Masters: [ pcmk-1 ]
    Slaves: [ pcmk-2 ]
ClusterIP    (ocf::heartbeat:IPaddr):    Started pcmk-2WebFS (ocf::heartbeat:Filesystem): Started pcmk-1
```

8.5. 重新配置 Pacemaker 为 Active/Active

基本上所有的事情都已经准备就绪了。最新的DRBD是支持 Primary/Primary(主/主)模式的，并且我们的文件系统的是针对集群的。所有我们要做的事情就是重新配置我们的集群来使用它们(的先进功能)。

这次操作会改很多东西，所以我们再次使用交互模式

crm # cib new active

There’s no point making the services active on both locations if we can’t reach them, so lets first clone the IP address. Cloned IPaddr2 resources use an iptables rule to ensure that each request only gets processed by one of the two clone instances. The additional meta options tell the cluster how many instances of the clone we want (one "request bucket" for each node) and that if all other nodes fail, then the remaining node should hold all of them. Otherwise the requests would be simply discarded.

```
# configure clone WebIP ClusterIP \
    meta globally-unique="true" clone-max="2" clone-node-max="2"
```

现在我们要告诉集群如何决定请求怎样分配给节点。我们要设置 clusterip_hash这个参数来实现它。

打开ClusterIP的配置

```
# configure edit ClusterIP
```

在参数行添加以下内容:

```
clusterip_hash="sourceip"
```

完整的定义就像下面一样:

```
primitive ClusterIP ocf:heartbeat:IPaddr2 \
    params ip="192.168.122.101" cidr_netmask="32" clusterip_hash="sourceip" \
    op monitor interval="30s"
```

以下是完整的配置

```
# crm crm(live)
# cib new active
INFO: active shadow CIB created
crm(active) # configure clone WebIP ClusterIP \
    meta globally-unique="true" clone-max="2" clone-node-max="2"
crm(active) # configure shownode pcmk-1
node pcmk-2
primitive WebData ocf:linbit:drbd \
    params drbd_resource="wwwdata" \
    op monitor interval="60s"
primitive WebFS ocf:heartbeat:Filesystem \
    params device="/dev/drbd/by-res/wwwdata" directory="/var/www/html" fstype="gfs2"
primitive WebSite ocf:heartbeat:apache \
    params configfile="/etc/httpd/conf/httpd.conf" \
    op monitor interval="1min"
primitive ClusterIP ocf:heartbeat:IPaddr2 \
    params ip="192.168.122.101" cidr_netmask="32" clusterip_hash="sourceip" \
    op monitor interval="30s"
ms WebDataClone WebData \
    meta master-max="1" master-node-max="1" clone-max="2" clone-node-max="1" notify="true"
clone WebIP ClusterIP \
    meta globally-unique="true" clone-max="2" clone-node-max="2"
colocation WebSite-with-WebFS inf: WebSite WebFS
colocation fs_on_drbd inf: WebFS WebDataClone:Master
colocation website-with-ip inf: WebSite WebIPorder WebFS-after-WebData inf: WebDataClone:promote WebFS:start
order WebSite-after-WebFS inf: WebFS WebSiteorder apache-after-ip inf: WebIP WebSite
property $id="cib-bootstrap-options" \
    dc-version="1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f" \
    cluster-infrastructure="openais" \
    expected-quorum-votes="2" \
    stonith-enabled="false" \
    no-quorum-policy="ignore"
rsc_defaults $id="rsc-options" \
    resource-stickiness="100"
```

请注意所有跟ClusterIP相关的限制都已经被更新到与WebIP相关，这是使用crm shell的另一个好处。

然后我们要把文件系统和apache资源变成clones。同样的 crm shell会自动更新相关约束。

```
crm(active) # configure clone WebFSClone WebFS
crm(active) # configure clone WebSiteClone WebSite
```

最后要告诉集群现在允许把两个节点都提升为 Primary(换句话说 Master).

```
crm(active) # configure edit WebDataClone
```

把 master-max 改为 2

```
crm(active) # configure show
node pcmk-1
node pcmk-2
primitive WebData ocf:linbit:drbd \
    params drbd_resource="wwwdata" \
    op monitor interval="60s"
primitive WebFS ocf:heartbeat:Filesystem \
    params device="/dev/drbd/by-res/wwwdata" directory="/var/www/html" fstype="gfs2"
primitive WebSite ocf:heartbeat:apache \
    params configfile="/etc/httpd/conf/httpd.conf" \
    op monitor interval="1min"
primitive ClusterIP ocf:heartbeat:IPaddr2 \
    params ip="192.168.122.101" cidr_netmask="32" clusterip_hash="sourceip" \
    op monitor interval="30s"
ms WebDataClone WebData \
    meta master-max="2" master-node-max="1" clone-max="2" clone-node-max="1" notify="true"
clone WebFSClone WebFSclone WebIP ClusterIP \
    meta globally-unique="true" clone-max="2" clone-node-max="2"
clone WebSiteClone WebSitecolocation WebSite-with-WebFS inf: WebSiteClone WebFSClone
colocation fs_on_drbd inf: WebFSClone WebDataClone:Master
colocation website-with-ip inf: WebSiteClone WebIP
order WebFS-after-WebData inf: WebDataClone:promote WebFSClone:start
order WebSite-after-WebFS inf: WebFSClone WebSiteClone
order apache-after-ip inf: WebIP WebSiteClone
property $id="cib-bootstrap-options" \
    dc-version="1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f" \
    cluster-infrastructure="openais" \
    expected-quorum-votes="2" \
    stonith-enabled="false" \
    no-quorum-policy="ignore"
rsc_defaults $id="rsc-options" \
    resource-stickiness="100"
```

看看配置文件有没有错误，然后退出shell看看集群的反应。

```
crm(active) # cib commit active
INFO: commited 'active' shadow CIB to the cluster
crm(active) # quit
bye
# crm_mon
============
Last updated: Thu Sep 3 21:37:27 2009
Stack: openais
Current DC: pcmk-2 - partition with quorum
Version: 1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f
2 Nodes configured, 2 expected votes
6 Resources configured.
============

Online: [ pcmk-1 pcmk-2 ]

Master/Slave Set: WebDataClone
    Masters: [ pcmk-1 pcmk-2 ]
Clone Set: WebIP Started: [ pcmk-1 pcmk-2 ]
Clone Set: WebFSClone Started: [ pcmk-1 pcmk-2 ]
Clone Set: WebSiteClone Started: [ pcmk-1 pcmk-2 ]
```

8.5.1. 恢复测试

注意

TODO: Put one node into standby to demonstrate failover

配置 STONITH

目录

9.1. What Is STONITH

STONITH is an acronym for Shoot-The-Other-Node-In-The-Head and it protects your data from being corrupted by rogue nodes or concurrent access.

因为如果一个节点没有相应，但并不代表它没有在操作你的数据，100%保证数据安全的做法就是在允许另外一个节点操作数据之前，使用STONITH来保证节点真的下线了。

STONITH另外一个用场是在当集群服务无法停止的时候。这个时候，集群可以用STONITH来强制使节点下线，从而可以安全的得在其他地方启动服务。

9.2. 你该用什么样的STONITH设备。

重要的一点是STONITH设备可以让集群区分节点故障和网络故障。

The biggest mistake people make in choosing a STONITH device is to use remote power switch (such as many on-board IMPI controllers) that shares power with the node it controls. In such cases, the cluster cannot be sure if the node is really offline, or active and suffering from a network fault.

Likewise, any device that relies on the machine being active (such as SSH-based "devices" used during testing) are inappropriate.

9.3. 配置STONITH

1. Find the correct driver: stonith_admin --list-installed

2. Since every device is different, the parameters needed to configure it will vary. To find out the parameters associated with the device, run: stonith_admin --metadata --agent type

   ```
   The output should be XML formatted text containing additional
   parameter descriptions. We will endevor to make the output more
   friendly in a later version.
   ```

3. Enter the shell crm Create an editable copy of the existing configuration cib new stonith Create a fencing resource containing a primitive resource with a class of stonith, a type of type and a parameter for each of the values returned in step 2: configure primitive …

4. If the device does not know how to fence nodes based on their uname, you may also need to set the special pcmk_host_map parameter. See man stonithd for details.

5. If the device does not support the list command, you may also need to set the special pcmk_host_list and/or pcmk_host_check parameters. See man stonithd for details.

6. If the device does not expect the victim to be specified with the port parameter, you may also need to set the special pcmk_host_argument parameter. See man stonithd for details.

7. Upload it into the CIB from the shell: cib commit stonith

8. Once the stonith resource is running, you can test it by executing: stonith_admin --reboot nodename. Although you might want to stop the cluster on that machine first.

9.4. 例子

Assuming we have an chassis containing four nodes and an IPMI device active on 10.0.0.1, then we would chose the fence_ipmilan driver in step 2 and obtain the following list of parameters

Obtaining a list of STONITH Parameters

```
# stonith_admin --metadata -a fence_ipmilan
```

```
<?xml version="1.0" ?>
<resource-agent name="fence_ipmilan" shortdesc="Fence agent for IPMI over LAN">
<longdesc>
fence_ipmilan is an I/O Fencing agent which can be used with machines controlled by IPMI. This agent calls
 support software using ipmitool (http://ipmitool.sf.net/).

To use fence_ipmilan with HP iLO 3 you have to enable lanplus option (lanplus / -P) and increase wait after
 operation to 4 seconds (power_wait=4 / -T 4)</longdesc>
<parameters>
        <parameter name="auth" unique="1">
                <getopt mixed="-A" />
                <content type="string" />
                <shortdesc>IPMI Lan Auth type (md5, password, or none)</shortdesc>
        </parameter>
        <parameter name="ipaddr" unique="1">
                <getopt mixed="-a" />
                <content type="string" />
                <shortdesc>IPMI Lan IP to talk to</shortdesc>
        </parameter>
        <parameter name="passwd" unique="1">
                <getopt mixed="-p" />
                <content type="string" />
                <shortdesc>Password (if required) to control power on IPMI device</shortdesc>
        </parameter>
        <parameter name="passwd_script" unique="1">
                <getopt mixed="-S" />
                <content type="string" />
                <shortdesc>Script to retrieve password (if required)</shortdesc>
        </parameter>
        <parameter name="lanplus" unique="1">
                <getopt mixed="-P" />
                <content type="boolean" />
                <shortdesc>Use Lanplus</shortdesc>
        </parameter>
        <parameter name="login" unique="1">
                <getopt mixed="-l" />
                <content type="string" />
                <shortdesc>Username/Login (if required) to control power on IPMI device</shortdesc>
        </parameter>
```

```
        <parameter name="action" unique="1">
                <getopt mixed="-o" />
                <content type="string" default="reboot"/>
                <shortdesc>Operation to perform. Valid operations: on, off, reboot, status, list, diag, monitor
 or metadata</shortdesc>
        </parameter>
        <parameter name="timeout" unique="1">
                <getopt mixed="-t" />
                <content type="string" />
                <shortdesc>Timeout (sec) for IPMI operation</shortdesc>
        </parameter>
        <parameter name="cipher" unique="1">
                <getopt mixed="-C" />
                <content type="string" />
                <shortdesc>Ciphersuite to use (same as ipmitool -C parameter)</shortdesc>
        </parameter>
        <parameter name="method" unique="1">
                <getopt mixed="-M" />
                <content type="string" default="onoff"/>
                <shortdesc>Method to fence (onoff or cycle)</shortdesc>
        </parameter>
        <parameter name="power_wait" unique="1">
                <getopt mixed="-T" />
                <content type="string" default="2"/>
                <shortdesc>Wait X seconds after on/off operation</shortdesc>
        </parameter>
        <parameter name="delay" unique="1">
                <getopt mixed="-f" />
                <content type="string" />
                <shortdesc>Wait X seconds before fencing is started</shortdesc>
        </parameter>
        <parameter name="verbose" unique="1">
                <getopt mixed="-v" />
                <content type="boolean" />
                <shortdesc>Verbose mode</shortdesc>
        </parameter>
</parameters>
<actions>
        <action name="on" />
        <action name="off" />
        <action name="reboot" />
        <action name="status" />
        <action name="diag" />
        <action name="list" />
        <action name="monitor" />
        <action name="metadata" />
</actions>
</resource-agent>
```

from which we would create a STONITH resource fragment that might look like this

Sample STONITH Resource

```
# crm crm(live)# cib new stonith
INFO: stonith shadow CIB created
crm(stonith)# configure primitive impi-fencing stonith::fence_ipmilan \
 params pcmk_host_list="pcmk-1 pcmk-2" ipaddr=10.0.0.1 login=testuser passwd=abc123 \
 op monitor interval="60s"
```

And finally, since we disabled it earlier, we need to re-enable STONITH. At this point we should have the following configuration.

```
crm(stonith)# configure property stonith-enabled="true"crm(stonith)# configure shownode pcmk-1
```

```
node pcmk-2
primitive WebData ocf:linbit:drbd \
    params drbd_resource="wwwdata" \
    op monitor interval="60s"
primitive WebFS ocf:heartbeat:Filesystem \
    params device="/dev/drbd/by-res/wwwdata" directory="/var/www/html" fstype="gfs2"
primitive WebSite ocf:heartbeat:apache \
    params configfile="/etc/httpd/conf/httpd.conf" \
    op monitor interval="1min"
primitive ClusterIP ocf:heartbeat:IPaddr2 \
    params ip="192.168.122.101" cidr_netmask="32" clusterip_hash="sourceip" \
    op monitor interval="30s"primitive ipmi-fencing stonith::fence_ipmilan \ params pcmk_host_list="pcmk-1
 pcmk-2" ipaddr=10.0.0.1 login=testuser passwd=abc123 \ op monitor interval="60s"ms WebDataClone WebData \
    meta master-max="2" master-node-max="1" clone-max="2" clone-node-max="1" notify="true"
clone WebFSClone WebFS
clone WebIP ClusterIP \
    meta globally-unique="true" clone-max="2" clone-node-max="2"
clone WebSiteClone WebSite
colocation WebSite-with-WebFS inf: WebSiteClone WebFSClone
colocation fs_on_drbd inf: WebFSClone WebDataClone:Master
colocation website-with-ip inf: WebSiteClone WebIP
order WebFS-after-WebData inf: WebDataClone:promote WebFSClone:start
order WebSite-after-WebFS inf: WebFSClone WebSiteClone
order apache-after-ip inf: WebIP WebSiteClone
property $id="cib-bootstrap-options" \
    dc-version="1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f" \
    cluster-infrastructure="openais" \
    expected-quorum-votes="2" \
    stonith-enabled="true" \
    no-quorum-policy="ignore"
rsc_defaults $id="rsc-options" \
    resource-stickiness="100"
crm(stonith)# cib commit stonithINFO: commited 'stonith' shadow CIB to the cluster
crm(stonith)# quit
bye
```

附录 A. 配置扼要重述

目录

A.1. 最终的集群配置文件

```
# crm configure show
node pcmk-1
node pcmk-2
primitive WebData ocf:linbit:drbd \
    params drbd_resource="wwwdata" \
    op monitor interval="60s"
primitive WebFS ocf:heartbeat:Filesystem \
    params device="/dev/drbd/by-res/wwwdata" directory="/var/www/html" fstype="gfs2"
primitive WebSite ocf:heartbeat:apache \
    params configfile="/etc/httpd/conf/httpd.conf" \
    op monitor interval="1min"
primitive ClusterIP ocf:heartbeat:IPaddr2 \
    params ip="192.168.122.101" cidr_netmask="32" clusterip_hash="sourceip" \
    op monitor interval="30s"
primitive ipmi-fencing stonith::fence_ipmilan \
    params pcmk_host_list="pcmk-1 pcmk-2" ipaddr=10.0.0.1 login=testuser passwd=abc123 \
    op monitor interval="60s"
ms WebDataClone WebData \
    meta master-max="2" master-node-max="1" clone-max="2" clone-node-max="1" notify="true"
clone WebFSClone WebFS
clone WebIP ClusterIP \
    meta globally-unique="true" clone-max="2" clone-node-max="2"
clone WebSiteClone WebSite
colocation WebSite-with-WebFS inf: WebSiteClone WebFSClone
colocation fs_on_drbd inf: WebFSClone WebDataClone:Master
colocation website-with-ip inf: WebSiteClone WebIP
order WebFS-after-WebData inf: WebDataClone:promote WebFSClone:start
order WebSite-after-WebFS inf: WebFSClone WebSiteClone
order apache-after-ip inf: WebIP WebSiteClone
property $id="cib-bootstrap-options" \
    dc-version="1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f" \
    cluster-infrastructure="openais" \
    expected-quorum-votes="2" \
    stonith-enabled="true" \
    no-quorum-policy="ignore"
rsc_defaults $id="rsc-options" \
    resource-stickiness="100"
```

A.2. 节点列表

这个列表中的集群节点是集群自动添加的。

```
node pcmk-1
node pcmk-2
```

A.3. 集群选项

这是集群自动存储集群信息的地方

- dc-version - DC使用的Pacemaker的版本(包括源代码的hash)
- 集群-基层 - 集群使用的基层软件 (heartbeat or openais/corosync)
- expected-quorum-votes - 预期的集群最大成员数

以及管理员设置集群操作的方法选项

- stonith-enabled=true - 使用STONITH
- no-quorum-policy=ignore - 忽略达不到法定人数的情况，继续运行资源

```
property $id="cib-bootstrap-options" \
    dc-version="1.1.5-bdd89e69ba545404d02445be1f3d72e6a203ba2f" \
    cluster-infrastructure="openais" \
    expected-quorum-votes="2" \
    stonith-enabled="true" \
    no-quorum-policy="ignore"
```

A.4. 资源

A.4.1. 默认选项

这里我们设置所有资源共用的集群选项

- resource-stickiness - 资源粘稠值

```
rsc_defaults $id="rsc-options" \
    resource-stickiness="100"
```

A.4.2. 隔离

注意

TODO: Add text here

```
primitive ipmi-fencing stonith::fence_ipmilan \
    params pcmk_host_list="pcmk-1 pcmk-2" ipaddr=10.0.0.1 login=testuser passwd=abc123 \
```

```
    op monitor interval="60s"
clone Fencing rsa-fencing
```

A.4.3. 服务地址

Users of the services provided by the cluster require an unchanging address with which to access it. Additionally, we cloned the address so it will be active on both nodes. An iptables rule (created as part of the resource agent) is used to ensure that each request only gets processed by one of the two clone instances. The additional meta options tell the cluster that we want two instances of the clone (one "request bucket" for each node) and that if one node fails, then the remaining node should hold both.

```
primitive ClusterIP ocf:heartbeat:IPaddr2 \
    params ip="192.168.122.101" cidr_netmask="32" clusterip_hash="sourceip" \
    op monitor interval="30s"
clone WebIP ClusterIP
    meta globally-unique="true" clone-max="2" clone-node-max="2"
```

注意

TODO: The RA should check for globally-unique=true when cloned

A.4.4. DRBD - 共享存储

在这里，我们定义了DRBD技术服务，并指定DRBD应该管理的资源（从drbd.conf）。我们让它作为主/从资源，并且为了active/active，用设置master-max=2来允许两者都晋升为master。我们还可以设置通知选项，这样，当时集群的节点的状态发生改变时，该集群将告诉DRBD的agent。

```
primitive WebData ocf:linbit:drbd \
    params drbd_resource="wwwdata" \
    op monitor interval="60s"
ms WebDataClone WebData \
    meta master-max="2" master-node-max="1" clone-max="2" clone-node-max="1" notify="true"
```

A.4.5. 集群文件系统

群集文件系统可确保文件读写正确。我们需要指定我们想挂载并使用GFS2的块设备（由DRBD提供）。这又是一个clone，因为它的目的是在两个节点上都可用。这些额外的限制确保它只在有gfs-control和drbd 实例的节点上运行。

```
primitive WebFS ocf:heartbeat:Filesystem \
    params device="/dev/drbd/by-res/wwwdata" directory="/var/www/html" fstype="gfs2"
clone WebFSClone WebFS
colocation WebFS-with-gfs-control inf: WebFSClone gfs-clone
colocation fs_on_drbd inf: WebFSClone WebDataClone:Master
order WebFS-after-WebData inf: WebDataClone:promote WebFSClone:start
order start-WebFS-after-gfs-control inf: gfs-clone WebFSClone
```

A.4.6. Apache

最后我们有了真正的服务，Apache，我们只需要告诉集群在哪里可以找到它的主配置文件，并限制其只在挂载了文件系统和有可用IP节点上运行

```
primitive WebSite ocf:heartbeat:apache \
    params configfile="/etc/httpd/conf/httpd.conf" \
    op monitor interval="1min"
clone WebSiteClone WebSite
colocation WebSite-with-WebFS inf: WebSiteClone WebFSClone
colocation website-with-ip inf: WebSiteClone WebIP
order apache-after-ip inf: WebIP WebSiteClone
order WebSite-after-WebFS inf: WebFSClone WebSiteClone
```

附录 B. Sample Corosync Configuration

Sample Corosync.conf for a two-node cluster

```
# Please read the Corosync.conf.5 manual page
compatibility: whitetank

totem {
    version: 2

    # How long before declaring a token lost (ms)
    token:     5000

    # How many token retransmits before forming a new configuration
    token_retransmits_before_loss_const: 10

    # How long to wait for join messages in the membership protocol (ms)
    join:      1000

    # How long to wait for consensus to be achieved before starting a new
    # round of membership configuration (ms)
    consensus:   6000

    # Turn off the virtual synchrony filter
    vsftype:    none

    # Number of messages that may be sent by one processor on receipt of the token
    max_messages:  20

    # Stagger sending the node join messages by 1..send_join ms
    send_join: 45

    # Limit generated nodeids to 31-bits (positive signed integers)
    clear_node_high_bit: yes

    # Disable encryption
    secauth:    off

    # How many threads to use for encryption/decryption
    threads:      0

    # Optionally assign a fixed node id (integer)
    # nodeid:     1234

    interface {
        ringnumber: 0

        # The following values need to be set based on your environment
        bindnetaddr: 192.168.122.0
        mcastaddr: 226.94.1.1
        mcastport: 4000
    }
}

logging {
    debug: off
    fileline: off
    to_syslog: yes
    to_stderr: off
    syslog_facility: daemon
    timestamp: on
}

amf {
```

```
    mode: disabled
}
```

附录 C. 延伸阅读

- Project Website *http://www.clusterlabs.org*
- Cluster Commands A comprehensive guide to cluster commands has been written by Novell and can be found at: *http://www.novell.com/documentation/sles11/book_sleha/index.html?page=/documentation/sles11/book_sleha/data/book_sleha.html*
- Corosync *http://www.corosync.org*

附录 D. 修订历史

修订 1-1 Mon May 17 2010 Andrew Beekhof *andrew@beekhof.net*

Import from Pages.app

修订 2-1 Wed Sep 22 2010 Raoul Scarazzini *rasca@miamammausalinux.org*

Italian translation

修订 3-1 Wed Feb 9 2011 Andrew Beekhof *andrew@beekhof.net*

Updated for Fedora 13

修订 4-1 Wed Oct 5 2011 Andrew Beekhof *andrew@beekhof.net*

Update the GFS2 section to use CMAN

修订 5-1 Fri Feb 10 2012 Andrew Beekhof *andrew@beekhof.net*

Generate docbook content from asciidoc sources

索引