Linux 文件系统精通指南

作者：Sheryl Calish

究竟什么是“文件系统”？Sheryl Calish 介绍了这个概念以及它的实际应用

尽管内核是 Linux 的核心，但文件却是用户与操作系统交互所采用的主要工具。这对 Linux 来说尤其如此，这是因为在 UNIX 传统中，它使用文件 I/O 机制管理硬件设备和数据文件。

遗憾的是，新手通常会混淆介绍 Linux 文件系统概念的术语。术语文件系统 可以在 Linux 文件编制中互换使用，用于指代几个不同但相关的概念。除磁盘分区的具体实例外，文件系统还指代数据结构以及分区中文件的管理方法。

另新手更感困惑的是，该术语还用于指代系统中文件的整体组织形式：目录树。此外，该术语还可以指代目录树中的每个子目录，如在 /home 文件系统中。某些人认为，这些目录和子目录不能称作真正意义上的文件系统，除非它们均驻留在各自的磁盘分区上。然而，其他人却将其称作文件系统，这无疑又增添了困惑。

Linux 老手可以从上下文中理解这些术语的含义。而新手却很难在一时半会儿就辨别出这样的上下文。

本文的主要目标就是提供足够的背景知识，以帮助您辨别此术语的上下文。在阐明文件系统术语的细微差别的过程中，您还将学习如何将某些非常有用的相关工具从理论应用上升到实际应用。

本文主要介绍了 2.4 版 Linux 内核中的 Linux 磁盘分区和文件管理系统特性。此外，还介绍了 2.6 版 Linux 内核中的新特性。

磁盘分区概述

Linux 和 UNIX 中的基本文件存储单元都是磁盘分区，即将一个或多个硬盘的逻辑划分，操作系统将每个逻辑分区视为独立的磁盘。文件和文件管理系统“居住”在磁盘分区中。Linux 将这些磁盘分区作为设备处理，进而通过 /dev 目录中的特殊文件使用文件 I/O 机制。

有两种类型的设备文件：块和字符/原始。两者之间的一个重要差别是，块设备被缓冲，而字符设备因为没有文件管理系统，所以不被缓冲。在 Oracle 集群文件系统 (OCFS) 推出之前，使用原始设备是提高 Oracle 数据文件分区性能的常见方法。（在本文的后续部分，我们将详细介绍原始设备。）

存储在磁盘最开始位置的分区表提供了该磁盘上分区的映射。可以使用 fdisk 命令查看系统的分区表。

```
# fdisk -l

Disk /dev/hda:240 heads, 63 sectors, 1940 cylinders
```

```
Units = cylinders of 15120 * 512 bytes

Device Boot      Start         End      Blocks   Id  System
/dev/hda             1         286     2162128+   c     Win95 FAT32 (LBA)
/dev/hda2   *      288        1940    12496680    5     Extended
/dev/hda5          288         289       15088+  83     Linux
/dev/hda6          290         844     4195768+  83     Linux
/dev/hda7          845         983     1050808+  82     Linux swap
/dev/hda8          984        1816     6297448+  83     Linux
/dev/hda9         1817        1940      937408+  83     Linux
```

分区表中的名称 /dev/hda 至 /dev/hdd 分别代表 IDE 驱动器 1 至 4，其中 hda 代表驱动器 1，hdb 代表驱动器 2，依此类推。驱动器内的分区用数字指代，因此 /dev/hda5 是第一个 IDE 驱动器上的第五个分区。对于 SCSI 驱动器，使用了类似的命名模式：/dev/sda to /dev/sdd。

第一至第四个分区保留给主分区，第五个及随后的分区用于逻辑分区。因此，以上所示的分区表中有一个驱动器 hda，它包含一个主分区 hda1、一个扩展分区 hda2 和五个逻辑分区 /dev/hda5 至 /dev/hda9。以名称 shmfs 列出的文件系统表示根据 Linux 2.4 中的 POSIX 标准挂载为特殊文件系统的共享内存文件系统。

您可能已经注意到，在 fdisk 列表中 LBA 是括在括号中的。LBA 表示逻辑块寻址，它将硬盘的柱面、块和扇区模式转换为线性块编号进行处理。

在 Linux 中，分区分为主分区、扩展分区和逻辑分区。术语主分区 是先前 x86 系统上四个分区限制的遗留产物。与 DOS 和 Windows 不同，Linux 可以从主分区或逻辑分区启动。用作逻辑分区占位符的主分区称作扩展分区。扩展分区本身拥有指向一个或多个逻辑分区（它们只是主分区的子分区）的分区表。在以上的 fdisk 列表中，hda2 就是一个扩展分区。

文件管理系统概述

要使分区后的磁盘可用，必须在其上构建文件系统。这种情况下，通常还将文件系统称作“分区类型”、“基于磁盘的文件系统”和“文件系统类型”。实际上，可以将这些文件系统看作是文件管理系统，这是因为该称呼正体现了它们的功能：它们通过维护文件上的元数据，使系统上的文件保持状态一致。

Linux 项目的特点之一是需要实现与每个可用实用程序的多个样式和首选设置的兼容性，而这种兼容性在可用文件管理系统的选择上体现得最为明显。Linux 内核内部的虚拟文件系统 (VFS) 实现了此选择。VFS 采用了一组可由其他文件管理系统使用的基本数据结构。这些数据结构是超级块、inode、dentry（或目录文件）和数据块。

每个分区都包含一个超级块，用于维护分区中文件系统上的信息，包括一组在每个超级块中唯一编号的 inode、空闲 inode 的数目以及 inode 总数、数据块总数、空闲数据块数和文

件系统的状态。文件系统的状态有两种：干净（当文件处于未更改状态时）和脏（当有未写入磁盘的文件系统更改时）。超级块中的一个 inode 对应着一个文件。

除文件名外，inode 包含了有关文件的所有信息，其中包括：

- 地址
- 类型
- 大小
- 所有者
- 对文件数据所在块的引用
- 文件最后一次修改和访问的时间戳。

可以使用以下命令查看文件的 inode：

```
$ ls -i
```

正如前面已经提到的，inode 只在超级块中唯一编号，且每个分区只有一个超级块，这就是硬链接无法跨越多个分区的原因。

文件名通过 dentry 对象（用户看到的是目录文件）链接到一个 inode 编号。数据块保存实际的文件数据。

Linux 支持任何具备 VFS 定义的基本函数集的文件管理系统。对于像 vfat 这样的文件管理系统，Linux 项目提供了它自己的设备驱动程序。

您可以从以下输出中看到，不同的文件管理系统可以存在于同一系统的不同分区上。

```
df -T
Filesystem     Type         1K Blocks     Used     Available  Use%   Mounted on
/dev/hda6      reiserfs     4195632       2015020  2180612    49%    /
/dev/hda5      ext2         14607         3778     10075       8%    /boot
/dev/hda9      reiserfs     937372        202368   735004     22%    /home
/dev/hda8      reiserfs     6297248       3882504  2414744    62%    /opt
shmfs          shm          256220        0        256220      0%    /dev/shm
/dev/hda1      vfat         2159992       1854192  305800     86%    /windows/C
```

当前，Oracle 用户遇到的最常用的文件管理系统是 ext2/ext3、ReiserFS（不受 Oracle 支持）和 OCFS。以下是非 Oracle 分区主要特性的汇总表。

特性	*ext2*	*ext3*	*ReiserFS3.6（不受 Oracle 支持）*
最大分区大小	4TB	4TB	16TB
最大文件大小	2GB-4GB	2GB-4GB	8TB

块大小	1KB-4KB	1KB-4KB	只有 4KB
日志功能	无	是	有
崩溃后重新启动	慢	快	非常快
用于恢复清除文件的工具	有	有	无
崩溃后数据的状态	良好	非常好	一般
ACL 支持	有	有	无
稳定性	优秀	良好	良好

由于 ext2 和 ReiserFS 均提供了用户级安全性以及更高效的磁盘空间使用等特性，因此尽管至少 ext2 确实提供了碎片整理工具，但几乎不需要这些工具。Ext2 是传统的、事实上的标准 Linux 文件管理系统。它是 Red Hat 版本 Linux 的默认文件管理系统，而 ReiserFS 是 SUSE 的默认文件管理系统。ext2/ext3 的最大文件大小实际上取决于所选择的块大小和硬件体系结构。ext2 的许多特性之一是它允许由磁盘分区决定块大小。ReiserFS 技术允许在磁盘分区中使用可变的文件大小(这是因为它基于平衡树技术而不是基于范围)，因此除日志功能以外，高效的空间使用也是其设计所固有的。

日志文件管理系统（如 ext3 和 ReiserFS）记录对文件系统元数据：inode、空闲块分配映射、inode 映射等的更改。当系统崩溃时，可以通过此方式检查日志以获得最近修改的元数据，从而确保快速恢复文件系统。此功能对大型系统尤其重要。如果没有此功能，则在出现硬件故障后，对于 ext2 等文件系统，需要在重新启动时运行 fsck 工具。对于大型文件系统，此过程可能要花费几个小时。

当然，记录日志需要付出一定的代价，即需要在处理时间和恢复之间寻求一个平衡。对于 ext3，可以选择日志记录模式，这些模式允许在寻求上述平衡时做出某些自主决定。journal 模式（记录所有文件系统数据，包括数据块和元数据）是最安全但也是最慢的模式。默认模式（称作 orderd）只记录元数据，但在写元数据之前先将数据块写入磁盘，从而在快速恢复和快速性能之间取得折衷。最快的模式是 writeback 模式，该模式只记录元数据。在此模式中，可能会丢失文件数据，但文件系统自身的完整性将得到维护。

在编写本文档期间，Reiser4 刚好已经发布。同 ReiserFS3.6 一样，ReiserFS4 只记录元数据。与 ReiserFS3.6 不同的是，它基于新的舞蹈树算法，此算法似乎比平衡树算法更快。它还可以扩展到使用无数个 CPU，而且在磁盘写入时具有内置加密和压缩功能。

OCFS 是 Oracle 真正应用程序集群 (RAC)、配置文件和数据库文件的指定文件管理系统。其他文件（甚至是 Oracle 软件文件）将在 ext2/ext3 或 ReiserFS 上获得更好的性能。

当前，就文件管理系统的选择来讲，共同的见解是，除少数情况外，ext2、ext3 和 ReiserFS 之间的性能基本相当。然而在各种系统的拥护者之间却爆发了激烈的争论。ReiserFS 由于

能够处理可变的文件大小，因此更适用于具有许多小文件的系统。当然，如果您正要或计划在 Linux 上运行 Oracle RAC，则可能需要为 Oracle 数据文件和配置文件安装 OCFS 或使用自动存储管理 (ASM)。

除了最常见的 ext2/ext3 和 ReiserFS 文件系统以外，Linux 还支持其他本地文件系统，包括 IBM 的 jsf 和 SGI 的 xfs。对传统 UNIX 文件系统的支持包括 SYSV、BSD、Solaris、Next 和 Veritas VxFS。在各个级别支持的其他文件系统包括

- Microsoft 的 fat、ntfs、vfat、fat32
- IBM 的 hpfs (OS/2)
- Apple 的 Macintosh hfs
- Amiga 的 affs
- Acorn 磁盘文件系统 adfs

请注意，Oracle 不支持某些文件系统，因此在使用这些文件系统时风险自负。

Linux 内核 2.6 版中最重要的新特性是访问控制列表 (ACL)。ACL 允许为一个或多个用户列表或用户组授予对单个文件的使用权限。其他新特性包括：

- 对 CD-ROM 上使用的 ISO 9660 文件系统的增强支持
- 可以存储在文件系统中的默认挂载选项
- 用于加速文件搜索的索引目录
- 对 Windows 的逻辑磁盘管理器（动态磁盘）的支持
- 能够将 ntfs 挂载为读/写，但写仍处于试验状态
- 对 fat12（旧 DOS 文件系统）的增强支持

处理分区和文件系统的工具

要添加一个新磁盘或调整现有磁盘的大小，您需要使用 fdisk 或 cfdisk。尽管 cfdisk 表面上更易于使用，但 fdisk 已被证实最适用于磁盘分区。以下是有关使用 Linux 版本 fdisk 的几个原则，帮助您了解其可能得到的结果。

首先，以超级用户身份用设备名称调用 fdisk：

```
# fdisk /dev/hda

The number of cylinders for this disk is set to 1940.
There is nothing wrong with that, but this is larger than 1024,
and could in certain setups cause problems with:
1) software that runs at boot time (e.g., old versions of LILO)
2) booting and partitioning software from other OSs
(e.g., DOS FDISK, OS/2 FDISK)

Command (m for help):m
```

可以通过使用 p（或 print）命令显示分区表。使用 n 或 new 命令可创建新分区；使用 w 或 write 命令可把新分区表写入磁盘。输入新命令后，fdisk 需要知道您要创建逻辑分区还是主分区：

```
Command (m for help):n
Command action
l   logical (5 or over)
p   primary partition (1-4)
l
No free sectors available

Command (m for help):
```

您可以看到，如果没有任何空闲空间（如上所示），则您将收到以上消息。但如果有空闲空间，则 fdisk 需要知道您想要的分区号。如果输入“p”（代表主分区），则您将需要做出以下选择。

```
Partition number (1-4):
```

对于逻辑分区，您将需要做出以下选择

```
Partition number (5 or over):
```

然后，您可以输入新分区的起始柱面号。fdisk 将推荐一个默认编号，如下所示：

```
First cylinder (1-1940, default 1):1
```

您可以选择接受此编号。接下来，您需要输入最后一个柱面或分区大小：

```
Last cylinder or +sizeM or +sizeK(1-1940), default 5721:1G
```

此刻,fdisk 将假设这是一个常规 Linux 分区(由分区表“ID”列中的十六进制数字 83 标识）。可以使用 fdisk 中的 t 或 type 命令更改分区类型。可以使用 l 或 list 命令取得 fdisk 的可用分区类型。以下是可用类型的部分列表：

ID	*System*
82	Linux swap
83	Linux
85	Linux extended
8e	Linux LVM

必须注意，在您运行 write 命令之前，您在 fdisk 中执行的任何操作都将是临时的—如果您出于任何原因要离开 fdisk，则这确实很有好处。

重新组织分区和文件管理系统

由于每个分区都包含各自的文件管理系统，因此调整分区大小涉及调整文件管理系统和分区的大小。因此，可用的重新分区工具取决于所用文件管理系统的类型。对于 ext2/ext3 系统，有一些可以选择将 resize2fs 与 fdisk、GNU Parted 或 Partition Magic 结合使用。对于 ReiserFS，只能将 cfdisk 与 resize_resiszerfs 搭配使用，这是因为 GNU Parted 对于 ReiserFS 来说仍需改进。

resize2fs 和 resize_reiserfs 都可以调整文件管理系统的大小，并要求使用单独的分区大小调整程序 — fdisk 或 cfdisk。我本人曾使用过 GNU Parted 对 ext2 分区进行重新分区。这是一个相当容易使用的程序。GNU Parted 对 ReiserFS 的支持将来会变得更稳健。Partition Magic 是一个用于 DOS 和 Windows 的商业程序，但如果从它附带的启动软盘或 CD-ROM 运行，则可以用于 Linux ext2/ext3 分区。

尽管实际的命令取决于您所要更改到的系统，但更改文件管理系统的一般过程涉及

- 备份分区上的文件
- 删除分区中的文件
- 如果使用的是 fdisk，则可能删除一个分区以便为两个更小的分区留出空间
- 使用相应的命令生成新文件系统。例如，要创建 ext2 文件系统，您可以使用

```
$ mke2fs /dev/hda5  15088
```

```
_ ..I
```

可以随意指定块计数，如以上命令中的 15,088。以上事件序列的唯一例外是使用以下命令从 ext2 系统移植到 ext3 系统

```
$tune2fs -j /dev/hda3
```

但仍应进行备份。

挂载分区

仅当具有超级用户权限的用户挂载了分区，分区才在 Linux 中可用。对于 /etc/fstab 文件中列出的 Linux 分区，系统启动时会自动挂载。对于 CD-ROM 和软盘驱动器，通常只需单击相应图标即可。

可与挂载选项结合使用的选项取决于文件管理系统。例如，您可以按如下方式指定 ext3 日志选项：

```
$ mount -t ext3 -o data=journaled /dev/hda9 /home
```

要拆下软盘或 CD-ROM，您需要在拆下它之前使用以下命令将其卸载

```
$ umount /media/floppy
```

在 Linux 2.4 之前，一个文件系统只能挂载一次。而现在，不限制文件系统的挂载次数。

结论

Linux 文件系统是一个多方面的概念。本讨论旨在作为根据您自己的需要对它的有用性和合意性进行进一步研究的基础。

在本文的第 2 部分中，我们将介绍集群文件系统，其中包括 OCFS