第10章　网络文件系统

网络文件系统(NFS)是目前最典型的采用了RPC的网络服务。它允许用户完全像访问自己的本地文件一样，访问远程主机上的文件。这是通过综合客户机端上的内核功能（使用远程文件系统）和服务器端上的NFS服务器（提供文件数据）来实现的。对客户机来说，这种文件访问完全是透明的，而且是通过不同的服务器和主机架构来实现的。

NFS的优点如下：

■ 供所有用户访问的数据可保存在一台中央主机上，并在启动时随客户装载这个目录。举个例子来说，你可将所有用户的账号信息保存在一台主机上，令网络上的所有主机都从它那里装入/home。依靠已安装了的NIS，用户就可以登录到任何系统，采用的仍是同一个文件集。

■ 占据大量磁盘空间的数据也可保存在一台单一的主机上。比如，所有与　LaTeX和METAFONT相关的文件和程序可集中在同一个地方保存和维护。

■ 管理性数据也可保存在一台单一的主机上。要在 20台机器上安装同一个文件时，不再需要rpc。

NFS中的大部分都是Rick Sladkey（邮件地址jrs@world.std.com）编写的，他还编写了NFS内核代码的Linux实施方案和大部分NFS服务器。NFS服务器源于unfsd用户空间NFS服务器和hnfs Harris NFS服务器，前者最初是Mark Shand编写的，后者则Donald Becker编写的。

先来看看NFS的工作原理：客户机可以像安装物理设备一样，要求在本地目录上安装远程主机内的目录。但是用于指定远程目录的句法有所不同。比如，为了将 vlager上的/home目录安装到vale主机上的/users，管理员将在vale上执行下面的命令（注意，可省略 -t nfs参数，因为mount能够从冒号得知它指定了一个NFS卷）：

```
# mount vlager:/home /users
```

然后，mount将试着通过RPC，链接到vlager主机上的mountd安装程序。服务器便查看vale是否被许可安装这个目录，如果可以，就为 vale返回一个文件句柄。后续发出的请求/users目录下所有文件时，都将采用这个句柄。

当有人通过NFS访问文件时，内核会任命一个指向服务器主机上的 nfsd（NFSdaemon）的RPC调用。这个调用将采用前面的文件句柄、准备访问的文件名和用户的用户 ID以及组ID作为自己的参数。采用这些参数的目地是判断用户是否被许可访问指定文件。为了防止未授权用户读取和修改文件，两台主机上的用户ID和组ID必须是一样的。

在多数实施中，客户机和服务器的 NFS机器都是作为内核级程序来实施的，这种内核级程序在系统启动时，便从用户空间开始运行。它们就是服务器主机上的 NFS daemon（nfsd）和客户机上运行的BLOCK I/O Daemon（biod）。为了改进网络流通量， boid执行了先读后写方式的异步I/O数据块；并且，若干个nfsd程序可同步运行。

NFS实施和客户机代码紧密集成在内核的虚拟文件系统（ VFS）层相比，稍有不同，它不需要通过boid进行额外的控制。另一方面，服务器代码完全在用户空间内运行，所以同时运

行多个服务器备份将由于涉及到同步问题的缘故，几乎是不可能的（请注意， NFS的内核版正在开发过程中）。目前，NFS还缺乏先读后写机制，但有幸的是， Rick Sladkey正在筹划此事（后写这个问题是指内核缓冲区按 device/inode对制作索引，因此，不能用于 NFS安装文件系统）。

关于NFS代码，最大的问题是内核版本 1分配的内存块不得大于4K；不这样的话，将导致连网代码不能对有些大型数据报进行处理，这些数据报在减掉报头等之后，字节数仍然超过3500。也就是说，对运行于这些系统（它们采用的是默认设置大型 UDP数据报，例如，SunOS系统采用的是8K）上的NFS daemon来说，要对它们进行投递，需要将它们人工地拆成小的数据包。某些情况下，这样做可能导致性能急剧降低（ Alan Cox是这样解释的：NFS规范要求在服务器返回应答之前，将每次写操作填入（ flush）磁盘。由于 BSD内核只能接受[4K]的写操作，所以将四个 1K大小包写入基于BSD的NFS服务器，将产生四个 4K大小的写操作）。这个问题在后来的内核 1.1版本中得以解决，而且还对客户机代码进行了修改，使之可以利用这一特性。

10.1 NFS的准备工作

在将NFS用作服务器或客户机之前，必须确保自己的内核能支持 NFS。新版本的内核中，proc文件系统（/proc/filesystems）上有一个简单的接口，就是为此专门设计的，利用 cat便可显示它：

```
$ cat /proc/filesystems
minix
ext2
msdos
nodev   proc
nodev   nfs
```

如果上面的显示中没有出现 nfs，就必须利用已启用的 NFS编译你自己的内核。关于配置内核网络选项的详情，参见第 3章。

对于1.1版本之前的内核，要找出其中是否已启用 NFS支持，最简单的方法是安装一个NFS文件系统。对此，可在 /tmp目录下创建一个目录，试着在它上面安装一个本地目录，如下所示：

```
# mkdir /tmp/test
# mount localhost:/etc /tmp/test
```

如果安装失败，并出现这样的错误消息：“ fs type nfs no supported by kernel ”（内核不支持文件系统类型的 NFS），你就必须利用已启用的 NFS，编译一个新内核。其他的错误消息则无伤大雅，因为你还没有开始动手在自己的主机上配置 NFS deamon呢！

10.2 NFS卷的安装

NFS卷（不是指文件系统，因为它们并不是真正的文件系统）的安装方式和普通文件系统的安装方式非常相似。利用下面的语法，调用 mount：

```
# mount -t nfs nfs volume local dir options
```

nfs_volume被指定为remote_host:remote_dir。由于这种表示法是NFS文件系统独有的，所以可省去-t nfs选项。

另外还有许多选项，供安装NFS之后的mount所用。它们要么跟在命令行的 -o开关后，要么位于NFS卷的/etc/fstab条目的options字段内。两种情况下，不同的选项间用句点隔开。命令行上指定的选项始终优先于fstab文件内所给的选项。

/etc/fstab内的示例条目如下：

```
# volume mount point type   options
news./usr/spool/news   /usr/spool/news nfs time-14，intr
```

然后，用下面的语句，安装这个NFS卷：

```
# mount news:/user/spool/news
```

在没有fstab条目的情况下， NFS安装调用看起来更为难看。比方说，假设你从一台名为Moonshot的主机安装用户的根目录，该主机采用的数据块是默认的 4K大小。为了适应Linux数据报的要求，可能需要执行下面的命令，将原先的数据块减少为 2K：

```
# mount moonshot:/home /home -o rsize=2048, wsize=2048
```

其所有有效的选项列表与NFS能够识别的mount工具（由Rick Sladkey编写）一起，在nfs手册中都对它们有着详细的说明。大家可在 Rik Faith的util-linux包内，找到上文的mount工具。表10-1列举了一部分常用的选项。

表10-1 部分安装选项

选　　项	说　　明
rsize=n，wsize=n	选项指定NFS客户机读取请求的数据报大小。由于上面提到的UDP数据报大小的限制，所以当前的默认设置是1024个字节
time0=n	设置NFS客户机等待请求完成所需的时间（以十分之一秒计）。默认设置是0.7秒
Hard	显式标记该NFS卷为硬安装。这是默认设置
Soft	软安装驱动程序（与硬安装相反）
Intr	允许发出信号，中断NFS调用。该选项用于服务器没有作出应答时

如果服务器暂时不可访问，除了 rsize和wsize之外，其他所有选项都可用于客户机行为。它们以下面的方式进行合作：只要客户机向 NFS服务器发出请求，便希望在 timeout（超时）选项中指定的时间间隔之后完成操作。如果在指定时间内，没有收到确认，就会产生副超时，将在下一个时间段内重试这一操作。如果超过了 60秒，就会产生一个主超时。

默认情况下，主超时将导致客户机在控制台打印一条消息并重新再试，这一次采用的超时值是前一次的两倍。这种情况可能永无休止地重复下去。固执地重试某次操作，直到有服务器作出应答的卷叫做硬安装卷。与之对应的是软安装卷，只要一发生主超时，软安装卷就会为调用进程生成一个I/O（输入／输出）错误。由于缓冲区引入了后写机制，在调用进程下一次调用write(2)函数之前，这个错误条件是不会传回调用进程本身的，所以，程序根本无法确定自己是否已成功将数据写入软安装卷。

不管你安装卷时采用的是硬安装还是软安装，必须考虑到你想从该卷中访问何种类型的信息。例如，如果你用NFS安装自己的程序，肯定不想自己的 X会话变得一团糟，因为有人同时启用了7个xv备份或以太网插件，使你的网络凝住了。如果对这些程序采用硬安装的话，必须确保你的计算机能够在与自己的 NFS服务器重新建立联系之前，一直处于等待状态。另一

方面，对诸如NFS安装新闻分区或FTP规档之类的非关键数据来说，它们可采用软安装，这样的话，在远程主机暂时不可抵达或已经关机时，不会将你的会话挂断。如果到服务器的网络连接比较脆弱，或通过一个已载入的路由器中转，只好要么利用 timeo选项增大初始超时值，要么硬安装NFS卷，但允许发出信号中断NFS调用，如此一来，仍然可以中断任何一个已经挂断的文件访问。

通常，mountd程序将以特定的方式监视哪些目录是由哪台主机安装的。这些信息可利用showmount程序显示出来，showmount程序也包含在NFS服务器包内。但mountd还没有这个功能。

10.3 NFS Daemon

如果想为其他主机提供NFS服务，必须在自己的机器上运行 nfsd和mountd程序。作为基于RPC的程序，它们不是由inetd管理，而是在系统启动时，就开始运行并注册为端口映射器的。因此，只须在运行 rpc.portmap之后，保证启用它们就行了。通常，应该把下面的内容包含到你的rc.inet2脚本中：

```
if [ -x /usr/sbin/rpc.mountd ]; then
        /usr/sbin/rpc.mountd; echo -n " mountd"
fi
if [ -x /usr/sbin/rpc.nfsd ]; then
        /usr/sbin/rpc.nfsd; echo -n " nfsd"
fi
```

对NFS daemon为其客户机提供的文件来说，它们的拥有者信息通常只包含在数字化的用户和组ID内。如果客户机和服务器两者关联的用户和组名和这些数字化 ID内的一致，就可以说它们共享同一个uid/gid空间。例如，利用NIS将passwd信息分发到局域网内的所有主机就属于这种情况。

但有时，也有彼此不符的情况。除了更新客户机的 uid和gid，使之与服务器的相符之外，还可用ugidd映射程序来处理这种情况。利用下面介绍的 map_daemon选项，便可借助于客户机上的ugidd，要求nfsd将服务器的uid/gid空间映射为客户机的uid/gid空间。

ugidd是一个基于RPC的服务器，与nfsd和mountd一样，是通过rc.inet2开始启用的。

```
if [ -x /usr/sbin/rpc.ugidd ]; then
/usr/sbin/rpc.ugidd; echo -n “ ugidd ”
fi
```

注意 当我发现包括这一信息很有用时，某些人可能把“参照rc.inet看作是较低级的启动方法。取而代之的是，可在etc/rc.d/*层次中找到你希望运行的各项服务的脚本。

10.4 导出文件

上面选项适用于客户机的 NFS配置，而服务器应该采用哪些选项呢？服务器的配置选项应该在/etc/exports文件内设置。

默认情况下，mountd不允许任何人从本地主机安装目录，这是非常明智的一种措施。为允许一台或多台主机安装目录，必须在 exports（导出）文件中指定它。导出文件示例如下：

```
# exports file for vlager
/home              vale(rw) vstout(rw) vlight(rw)
/usr/X386          vale(ro) vstout(ro) vlight(ro)
/usr/TeX           vale(ro) vstout(ro) vlight(ro)
/                  vale(rw,no root squash)
/home/ftp          (ro)
```

每一行都定义了一个目录和允许装入该目录的主机。主机名通常采用完整资格域名，也可以增加 *和?通配符，表示采用的是 Bourne外壳。举个例子来说， lab*.foo.com对应lab01.foo.com和laber.foo.com。如果不指定主机名，就像上面示例中的 /home/ftp目录一样，是不允许任何一台主机安装所定义的目录的。

在导出文件中查看客户机主机时， mountd将使用gethostbyaddr (2)调用，查找客户机的主机名。利用DNS之后，这个函数调用就会返回客户机的规范名，所以你必须确保不要在导出文件内采用主机别名。如果不采用 DNS，函数调用返回的就是它在主机文件内找到的第一个与客户机地址相符的主机名。

主机名后面是一个可选项，它是一个用括号括起来的标记清单，各标记用句点隔开。表10-2展示了一部分标记及值。

表10-2 部分标记及值

标 记	值
nsecure	允许从该机起，进行未经验证的访问
unix-rpc	需要从该机起，进行Unix域的RPC身份验证。这个标记只要求请求从一个保留 Internet端口（也就是端口号必须小于1024）发起。默认情况下，这个标记是打开的
secure-rpc	要求从该机起，进行安全 RPC身份验证。这个标记尚未实施。个中原由，参见 Sun的Secure RPC文档
kerberos	要求从该机起，对访问实行 Kerberos身份验证。这个标记也没有实施。参见 MIT的Kerberos身份验证系统文档
root squash	这是一个安全特性，它在客户机上的 uid 0到服务器上的uid65534(-2)之间，实行请求映射，从而否决超级用户对特定主机的特殊访问权。后一个 uid应该和用户nobody（无人）对应
no root squash	不从uid 0映射请求。默认设置是开
ro	安装文件结构，只读。默认设置是开
rw	安装文件结构，只写。默认设置是开
link relative	通过在../ ' s上加必要的号码，将绝对象征性链接（链接内容以斜杠开头）转换为相对链接，从而从包含指向服务器上根目录的目录中，取得自己需要的链接。这个标记只适用于一台主机的整个文件系统都已安装时，有些链接可能无处可指，或更糟的是，指向它们根本不打算拥有的链接。默认设置是开
link absolute	让所有的象征性链接保持原样（对 Sun提供的NFS服务器而言，则是普通行为）
map daemon	该标记要求NFS假定这样一个前提：客户机和服务器不共享同一个 uid/gid空间。然后，nfsd通过查询客户机的ugidd程序，建立一个映射客户机和服务器 ID关系的列表

只要开始运行 nfsd或mountd，导出文件错误分析将以“注意”形式，报告给 syslogd的daemon设备。

注意，主机名是根据客户机的 IP地址逆向映射而来的，所以必须保证你的解析器配置无误。如果采用的是 BIND，而且非常注重安全问题，则应该启用电子欺骗，检查自己的host.conf文件。

10.5 自动安装器

有时，安装所有的NFS卷相当费事，一来因为卷数较多，二来因为耗时。因此，一个所谓的“自动安装器”应运而生。它其实是一个daemon，会自动而透明地将安装需要的NFS卷，并在没有使用它们达一段时间后，取消安装。自动安装器的过人之处在于它能够从备用地方安装特定的卷。比如，你将自己的X程序和支持文件备份保存在两台或三台主机上，通过NFS，令其他主机安装它们。如果利用自动安装器，可指定这三台主机将被安装在 /usr/X386上，然后，自动安装器便试着进行安装，直到安装尝试成功为止。

常随Linux一起使用的自动安装器叫作amd。这个程序起初是Jan Simon Pendry编写的，已经被Rick Sladkey移植到Linux系统中。目前版本是amd-5.3。

关于amd的详情，不在本章讨论之列；要想找到一本好的参考手册，不防参阅其源代码；其中包含一个内容丰富翔实的texinfo文件。