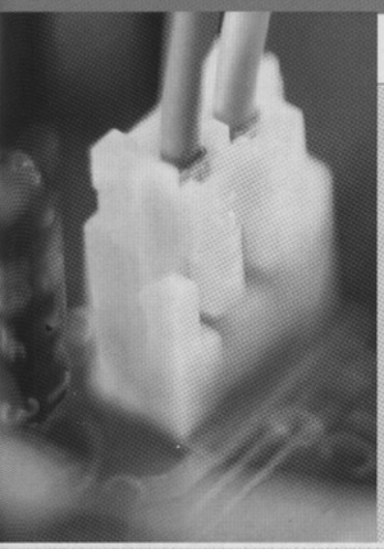

一种用于流式内容分发的分布式文件系统

High-Available File System Used in Stream Content Distributing

摘　要：DistFS是一种用来提供文件夹元数据的同步和按需提供数据流的文件系统，主要应用于Linux以及Solaris,HP-UX等UNIX操作系统中，它存在于操作系统的虚文件系统和物理文件系统的中间，是相对独立的一层。通过在Internet上的代理系统配置多媒体文件的缓存，该文件系统的流式传输特性能够最大限度的发挥其在服务器访问上所具备的许多优势。

关键词：DistFS　无缝重整　同步复制

陈胜利　罗晓沛（合肥中国科学技术大学计算机系　230026）

1 前言

Web和文件服务器对大多数单位而言，具有战略上的重要意义。在www和其他服务器之间进行数据复制，以及在不同的数据中心之间同步数据，可以使地理上分散的数据中心，通过服务器复制避免主干线上的交通阻塞，从而为各种应用提供高品质的服务。

在我们的项目（“大型数据中心存储系统”）进行当中、设计了一种叫做DistFS的分布式文件系统，它提供几个重要特性以提高数据复制分发和存储备份性能：

- **复制：**为服务器上的文件夹集合产生镜像。
- **无连接下操作：**在服务器或网络连接失败期间，仍能保证文件夹的有效更新。
- **无缝重整：**一旦文件系统检测到可恢复的网络或服务器，就自动同步恢复系统。

在Linux以及Solaris,HP-UX等UNIX操作系统中，DistFS嵌入在宿主操作系统的虚文件系统和物理文件系统的中间，是相对独立的一层。它并不会占用宿主操作系统很多的资源，并且对于用户的各种应用来说，这一层是完全透明的。

在正常操作中，DistFS 是一个基于客户机/服务器的文件系统，可以在多个服务器上保留文件系统的副本。当服务器或网络连接失败，DistFS 会在日志文件中记录下文件系统的更新过程；当服务器或网络恢复正常后，系统会利用这些日志对所有服务器进行数据更新。

使用DistFS，即便是在一台分布式服务器失效的情况下，网络上其它各节点也能保持同步化。

2 DistFS的特点

DistFS具有以下特点：

当文件和文件夹发生改变时，文件系统核心层会修改日志并进行自动更新。在连接正常时，日志的更新机制会立即在服务器上作重整。一旦客户登陆网络，该机制便能传送给系统复制的某些文件夹集合。当网络或服务器失效时，或移动客户计算机正常中断连接时，DistFS会自动支持无连接操作。

为确保可扩展性，DistFS引入了日志文件系统的恢复机制，不论数据备份还是服务器上的缓存内容都会通过日志文件的传输来解决。

3 DistFS的机制

DistFS是一种客户机/服务器文件系统，它基于版本和文件夹的更新，而与网络连接状况无关。有别于其他的网络文件系统，DistFS专门有一个核心模块来封装现有的日志文件系统，它负责截获对文件夹或文件的更新，并对已进行的操作和被更新的文件夹和文件版本作详细的日志记录。必要时，为获取当前数据，DistFS可以捕获各应用对数据的访问。

无论在服务器或客户端，DistFS都将文件数据存储在当前磁盘文件系统中。在客户端，这个文件系统在缓存中包含当前或最近使用的文件的副本。在服务器端，这个磁盘文件系统在文件夹集合中包含所有文件的有效副本。

DistFS采用的大块数据传输机制是非常灵活的。利用高性能服务器

和网络，它可以检验安全连接并使用复杂的压缩技术（例如：rsync 协议）进行性能优化。

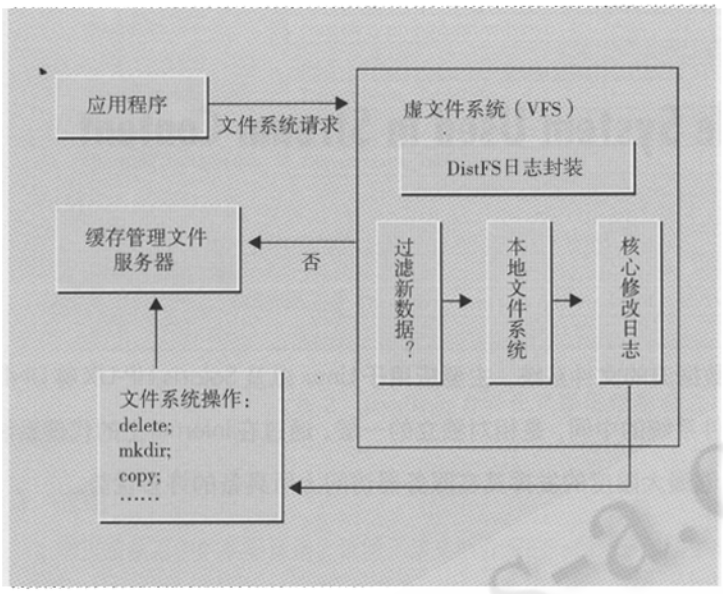

图 1　DistFS 过滤过程

DistFS服务器能够灵活的进行配置，可以通过转发所有更新的副本，来保持存储在另一DistFS机器上文件夹集合中数据备份的完整性和一致性。这一特性保证服务器能在很大的集群应用中得到同步化。

DistFS文件系统的更新借助于写令牌。要更改文件系统中的数据，系统必须首先获得来自于服务器对该文件夹集合唯一的写令牌。一旦完成修改，先对服务器作更新，然后更新所有客户端的备份，从而使DistFS的客户和服务器的文件夹集合在时间上实现同步。

4 无连接操作和重整

当DistFS客户端无法与服务器相连，或当服务器无法接收合法客户的更新时，DistFS自动转换成无连接操作。事实上,这些情况会经常发生，比如当服务器或网络的硬件及软件出现故障时。在无连接状态下，DistFS允许访问已存储的文件并通过日志记录进行更新,从而提供了高可用性的数据访问。

当连接恢复时，DistFS能够同步化存储在客户端和服务器端日志文件中的变更。首先，DistFS将启动更新传递机制来转发已存储在服务器上的变更（但这种情况对再次连接的客户是无效的）。其次，DistFS将在服务器上重整客户端的更新。当更新传递和重整完成时，文件夹集合也相应地被同步化了。

在更新传递过程中，有时候会出现更新冲突。这种情况多发生在以下情况，当相同的文件夹或文件在无连接的客户端被修改，或在客户端已完成对服务器的更新重整时，一旦检测出冲突存在，DistFS即刻中断更新传递，并启动冲突解决方案（这是一个用以解决更新冲突的软件模块）。冲突解决当中，只有再连接的客户受影响。典型的冲突可以通过强制文件夹的布局加以避免。在具体应用设计中，按照简单的CGI程序设计和WEB的向导提示执行，可以避免在复杂系统中的冲突。

5 DistFS 用于流式内容分发

很多单位和部门已经应用Web服务器来存储海量的多媒体文件，诸如视频和音频数据。为了将这些形式的内容传送给终端用户，最为重要的是如何从Internet主干连接上下载这些内容。如果在靠近Internet数据源的边缘处配置多媒体文件的缓存，DistFS的流式传输协议就能够发挥许多优势。比如说：

- 实时对所有系统作更新内容通知
- 智能化的处理连接中断
- 高性能的文件转移功能
- 利用小容量的缓存提供全面的服务器映像

图2表示在接近Internet边缘的两个Web服务器所起到的缓存作用，用以处理来自大型数据中心的文件。如图所示：服务器上的数据变更是如何迅速提交到缓存，然后，缓存客户是如何以响应式来获取数据的。

DistFS缓存可被配置成在新文件出现时，发出通知立即获取数据；或延迟数据流直到第一次访问发生。在前者情况下，数据的完整副本会马上建立；后者则建立元数据的副本以及为文件数据的响应流作初始化。

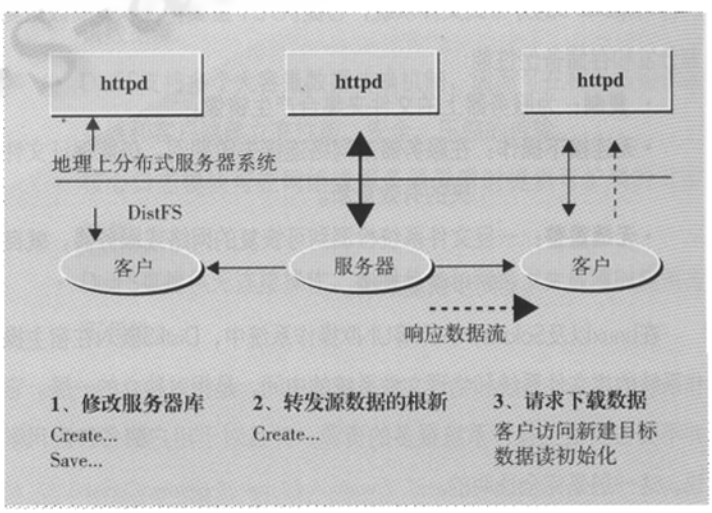

图 2　DistFS 媒体分发

当系统通知缓存存在新文件时，它会将文件作为稀疏对象安装在文件系统中。这意味着文件的属性是正确无误的，但数据块还不存在。当一应用程序，例如web或视频服务器，第一次尝试阅读文件，

下转第 30 页 >>

日志封装模块会检查数据是否存在，并可以从客户端到服务器查看初始化文件的数据流是否正确的初始化了。

在写入缓存时，DistFS的缓存管理可判断到哪些文件是很少使用的，并将它们挪走。在磁盘空间紧张时，DistFS将挪走最初的文件。

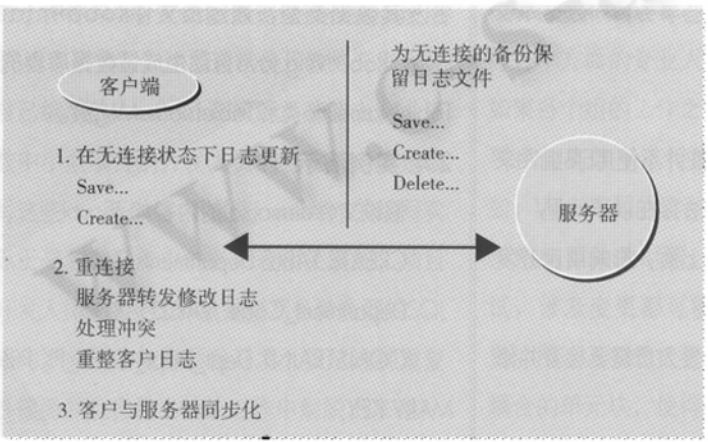

图 3　DistFS 再连接顺序

现有成型的技术中，Akamai 和Xosoft是基于集成在WWW服务器的网络协议，提出媒体分发的解决方案。DistFS可以与它们组合，以完善其功能。Xosoft QuickMirror 使用与DistFS提供的相似的缓存，但Xosoft相对于DistFS文件系统缺乏易实现性，规模性和可恢复性。

总的说来,在流式内容分发的应用中,DistFS具有以下优势：

- 用来描述更新重整和转发变化的日志文件都可自动更新，文件数据按需获取。
- 当文件第一次被读取，DistFS系统将从服务器到客户端初始化该数据流。
- DistFS支持多个大容量数据传输机制，因而可在相互备份的服务器和客户端建立高性能、安全、压缩的传输。
- 核心日志模块可在网络或服务器异常时提供可扩展的日志恢复。
- DistFS可提供无连接操作，在服务器和网络发生故障时提供高可用性。

参考文献

1 《Modern Operating System II》Andrew S. Tanenbaum ,Prentice Hall 出版，2001年2月。
2 《Advanced Programming in the UNIX Environment》, W.Richard Stevens, Addison-Wesley Pub Co出版，1992年6月。
3 《TCP/IP详解 卷3: TCP事务协议、HTTP、NNTP和UNIX域协议》, W.Richard Stevens，机械工业出版社，2002年1月。