分布式网络存储管理系统的研究与实现

张 凌 蒋东兴 刘启新 周 霖 沈培华
(清华大学计算机与信息管理中心 北京100084)

Research and Design on the Distributed Network Storage Management System

ZHANG Ling JIANG Dong-Xing LIU Qi-Xin ZHOU Lin SHEN Pei-Hua
(Computer and Information Management Center of Tsinghua University, Beijing 100084)

Abstract The distributed network storage management system will effectively organize the users' storage spaces which are distributed and heterogeneous to provide the users with a piece of integrated and uniform virtual storage space. This paper summarily introduces the fundamental function and structure of the system, and mainly describes the related protocols and the method for realizing.

Keywords Virtual storage space, NBD, Raid5, Standard interface for file access

1. 前言

随着网络技术及相关应用技术的发展,越来越多的用户需要依赖网络工作,用户的大量信息需要通过网络进行存储和转移。随着这种趋势的不断发展,部分用户向网络提出了对其个人数据和信息进行存储和安全维护的要求。然而,传统的网络应用,如 E-mail、FTP 等服务,由于其异种服务间兼容性差、用户所存储的信息安全性差以及单一服务可容纳的数据量有限等缺点,已经不能充分满足用户大容量、高速度的数据存储和传输的需求。

为了满足用户对网络应用的进一步需求,解决传统的网络应用服务中的不足,需要提供专门针对个人用户的网络存储管理服务,为用户提供大容量、可靠的个人存储空间,帮助用户解决个人信息的存储、转移等问题。本文所研究的分布式网络存储管理系统就是一种可以为用户提供上述服务的系统。

2. 系统功能

分布式网络存储管理系统是真正面向个人用户的网络存储管理系统。该系统的主要功能为:将用户分布在网络上的、异构的存储空间(如 FTP、NFS、SMB 等服务器上的存储空间)统一组织和管理起来,提供给用户一片完整、可靠、透明的"虚拟个人网络存储空间"(以下简称:虚拟个人空间),并提供良好的网络存储管理服务。

2.1 系统主要功能

本系统核心功能是,帮助用户统一管理分布、异构的网络存储空间,最终呈现给用户完整和透明的虚拟个人空间。由于用户的网络存储空间是分散的和异构的,这导致用户使用过程繁琐。通过本系统对用户空间的组织和代理访问,用户只需对虚拟个人空间进行访问就可以完成对各存储空间的操作。

系统还提供必要的数据安全保障机制,以保证用户数据的安全。通过对数据进行校验、加密/解密处理和使用安全传输通道等机制,保证用户数据的传输和存储过程安全;通过数据备份和备份恢复等机制,保证用户数据在遭到破坏的情况下可尽快恢复。

另外,系统还提供单一登录访问接口,用户只需一次性登录到虚拟个人空间就可以完成对文件的访问,不必再对各存储空间分别进行重复性的登录操作。用户将自己权限下的个人存储空间统一交给本系统管理,登录各存储空间所必需的身份认证过程由本系统代理完成。

2.2 为用户提供的服务接口

系统向用户提供访问接口,用户通过该接口直接访问自己权限内的虚拟个人空间。系统提供两种访问方式:

(1)*Web 浏览器访问方式* 用户可以通过系统的 Web 管理页面对自己的虚拟个人空间进行访问和操作。该访问方式连接方便,用户无须在本地操作系统中安装任何客户端软件或插件,使用网络中的任何一台计算机都可以完成访问。但是由于 Web 页面功能和协议上的制约,用户对文件的操作(如文件的移动、内容修改等)比较繁琐。

(2)*映射网络驱动器访问方式* 用户利用 Windows 资源管理器中的"映射网络驱动器"功能,将自己权限内的虚拟个人空间映射为客户端的网络驱动器。该访问方式的优点在于文件操作简单。系统通过 Windows 文件系统对虚拟个人空间上的数据进行操作,因此,所有文件操作都等同于本地文件的操作。完成这种方式的访问,用户需要在所使用的计算机上安装相关客户端插件。

另外,系统为用户提供虚拟个人空间的管理接口。例如,允许用户向系统添加(或从系统删除)分布、异构的网络存储空间,并对这些空间的属性进行配置。

3. 结构设计

3.1 系统模型

本系统采用 Client/Server 结构设计,如图1所示。

从图1中可以看出,"分布式网络存储管理系统"(存储管理服务器)和各存储服务空间构成了 Client/Server 的结构。本系统作为客户端计算机的存储访问代理,经过客户端用户授权后,成为虚拟的客户端实现对存储服务空间的文件访问。

另外,由于本系统完全代理客户端对存储服务空间的访问,用户只需对系统提供的虚拟个人空间进行登录和访问,就能完成对各存储服务空间的文件操作。因此,"客户端计算机"和"分布式网络存储管理系统"(存储管理服务器)构成另一对 Client/Server 结构。

从图1中不难看出,存储管理服务器的主要工作就是接收客户端用户的操作指令,并代理客户端用户完成对存储服务

器的文件操作。

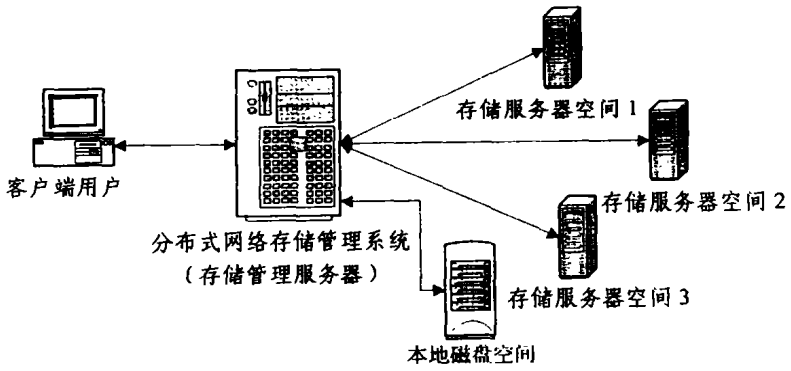

图1　分布式网络存储管理系统的系统模型

3.2　系统结构

图2描述了分布式网络存储管理系统的系统结构。系统的主要模块包括用户接口、网络存储标准访问接口、存储管理、存储访问代理等。

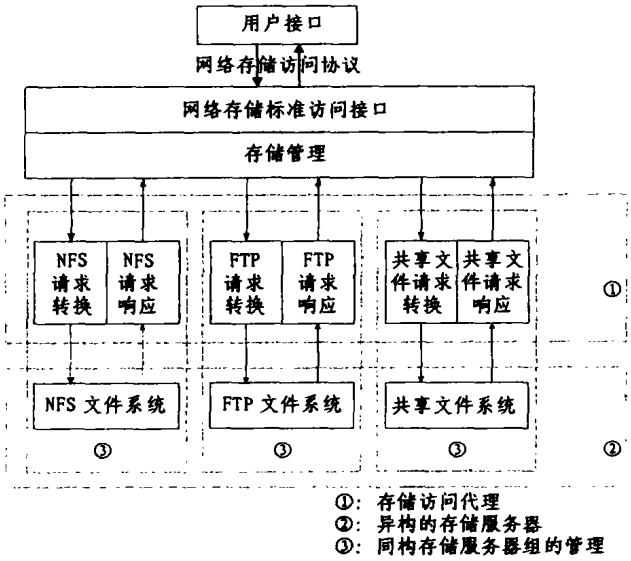

图2　分布式网络存储管理系统的结构

用户接口实现的功能是接收用户向系统发出的请求。用户可以通过访问 Web 页面和在客户端映射网络驱动器两种方式向系统发出操作请求。

由于用户可见的是一片统一、完整的个人存储空间，并且系统内部的存储方式和存储结构对用户透明，因此，用户提出的文件请求只能是面向统一和完整的虚拟个人存储空间的“虚拟”文件请求。这就需要系统能够将这些虚拟的请求解析成对各网络存储空间（或服务器本地磁盘文件系统）的实际数据请求。“解析用户指令”就是标准文件访问接口需要完成的功能。

存储访问代理最终接收发向各存储服务器的操作请求，并作为虚拟的文件服务客户端和各存储服务器发生数据和信息交换，完成作为虚拟客户端实际访问网络存储服务器的功能。

4. 实现方案

4.1　实现技术

“分布式网络存储管理系统”是架构在 Linux 操作系统平台上的服务和管理系统。

系统的实现主要利用了 NFS（Network File System）、SMB（Session Message Block）等文件共享访问协议和 Linux 内核（2.1.101以上版本）中的 NBD（Network Block Device）技术。其中借助 NFS 协议实现系统向网络存储空间请求数据；SMB 协议主要用于向网络存储空间请求数据和为客户端用户提供映射网络驱动器服务；NBD 技术主要用于构建网络块设备，使得系统能够在网络块设备的基础上构建虚拟文件系统。

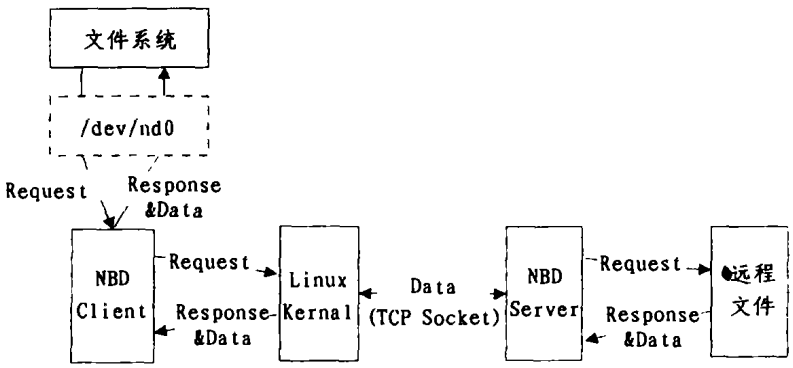

图3　NBD（Network Block Device）工作示意图

另外，系统数据库使用 Oracle，系统程序的编写采用 C 语言，数据库访问部分用 Pro*C 编写。Web 服务器采用 Linux＋Apache，Application Server 采用 Resin（v2.0.5），Web 网页的编写采用 jsp 技术。

4.2 系统协议设计

(1)NBD 技术应用　目前,2.1.101以上版本的 Linux 标准内核均支持 NBD 模块。NBD 技术的应用是本系统得以实现的关键技术。

一般定义的块设备是指用来存储数据并对它的各部分内容提供同等访问权的设备。通过块设备,文件系统可以从硬盘的任何位置获取数据。网络块设备(NBD)技术是在此基础上,利用远程文件作为文件系统的块设备文件,并构建虚拟文件系统。

NBD 技术除内核部分主模块外,主要由两个用户态进程完成对用户态文件的读写操作,这两个功能模块是 NBD-Server 和 NBD-Client。它们负责建立一个基于 TCP 的连接,并将带有文件句柄(File Handle)、数据读写信息的已经连接好的 TCP Socket 传送给内核驱动,并由 NBD-Server 和 Linux 内核共同完成数据交换。当文件系统希望通过块设备(例如/dev/nd0)获取数据时,通过内核模块以及外部用户态进程的共同控制和操作,使文件系统实际读到的是远程的块设备文件所提供的数据。

其中,TCP 连接的数据包格式如下:

```
struct nbd_request {
    u32 magic; /* 数据包标志位 */
    u32 type; /* 数据读写标志:1-Write,0-Read */
    char handle[8]; /* 操作的句柄 */
    u64 from; /* 操作起始位置 */
    u32 len; /* 读写的数据长度 */
};
```

如果此数据包的读写类型为"写",则紧跟在此 TCP 包之后的将是长度为 len 的数据。

```
struct nbd_reply {
    u32 magic; /* 数据包标志位 */
    u32 error; /* 操作错误标志位:0-No error,else
                  -error */
    char handle [8]; /* 操作的句柄,同于 nbd_re-
                        quest 的 handle */
};
```

如果此数据包应答的请求包的读写类型为"读",并且操作无误,则紧跟在此 TCP 包之后的将是长度为 len 的数据。

(2)Raid5协议移植　在 Raid5协议(Redundant Arrays of Inexpensive disks,level5)中,数据被分块存储在不同的磁盘中,并且用来进行纠错的校验信息被分散在组内所有的磁盘上,如图4所示。其中:

$D1 \oplus D2 \oplus \cdots \oplus Dn = P1$

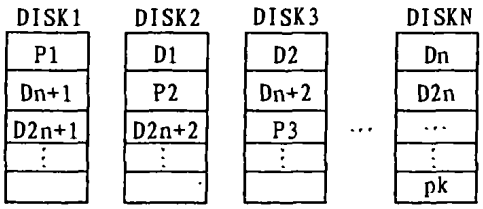

图4　Raid5协议

为了维护用户数据的安全,本系统采用 raid5协议的规则对用户数据进行检错和纠错处理。然而,与 raid5协议原有规则不同,协议原指将 N 块磁盘组成一个阵列,当作单一磁盘使用;本系统的操作对象是异构的文件,因此对协议的实现做出了适应性的调整:

将用户的 M(M>0)个文件编号,并按编号顺序将其连接成一个"虚拟用户文件",并保证:

i. 对所有用户文件的访问即为对"虚拟用户文件"的访问,并且该访问是严格按照单个文件的字节顺序和所有文件连接成"虚拟用户文件"时的编号顺序进行的;

ii. "虚拟用户文件"的第 W 个字节,就是编号为 k 的文件的第 Y 个字节,其中:

$$W = \sum_{i=0}^{k-1} B_i + Y$$

B_i:编号为 i 的文件容量

根据协议需要,这个"完整"的虚拟用户文件被平均分成 N(N>=2)个数据段(Stripe),并将这些数据段虚拟成为 N 块同构的磁盘进行数据操作,如图5所示。所有数据均被分块(D_i)存储在这 N 个数据段中,校验信息(P_i)也被分散地存储在各个数据段中。下文中提及的"Raid5协议"均指适用于本系统的、经过调整的 Raid5协议。

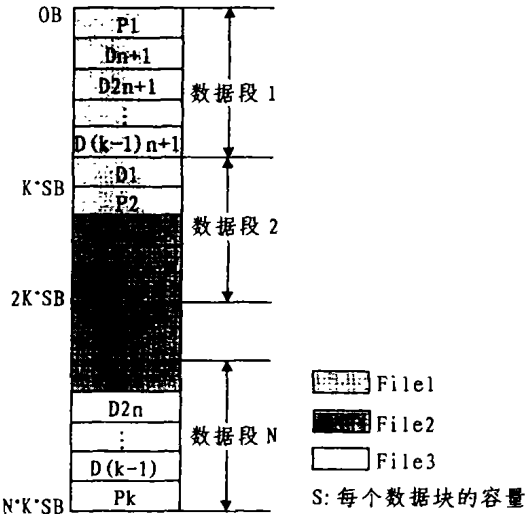

图5　本系统中 raid5协议的实现

但是,在上述存储策略下,仍然存在着不可避免的问题,即在 N*k 个数据块中,只有其中的(N-1)*k 个数据块存储的是有效数据,其余 k 个数据块中存储的是校验信息。因此,对存储容量造成了1/N 的损失。

(3)标准文件访问接口协议设计　在分布式网络存储管理系统中,用户个人的存储空间是分布和异构的,因此,必须存在能够分别访问这些异构的存储文件并返回标准格式的数据的接口。基于此目的,系统设计和实现了标准文件访问接口协议,该协议定义了一组对分布的和异构的文件系统进行访问、并将访问结果通过标准格式返回给上层请求的规则和方法。通过该协议(接口),上层请求不必关心被访问文件的存储结构、格式和位置,并透明地对其进行访问。

标准文件访问接口协议的4个层次分别完成不同的控制功能,如图6所示。其中:

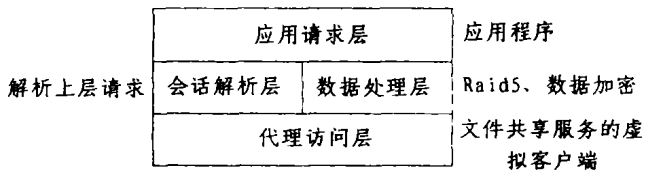

图6　标准文件访问接口层次模型

代理访问层:代理访问层作为文件服务的虚拟客户端,从分布的、异构的文件系统中获得文件数据。这些文件系统包括

本地硬盘文件系统、网络文件共享服务(如 NFS、SMB、FTP 等)和系统重新构建的虚拟文件系统等。代理访问层关心的主要问题是:被访问文件的存储位置、文件系统格式以及文件的访问方式等。通过代理访问层的工作,保证上层协议透明地对文件数据进行处理。

数据处理层:数据处理层主要负责对数据进行 Raid5协议处理、加密/解密等处理,使得交给上层的数据是正确无误、明码的,交给下层的是经过加密和加校验和的存储状态文件。数据处理层主要关心的问题是:数据的加密方法、Raid5协议的数据分段方式以及上层协议请求的数据格式等。

会话解析层:会话解析层解析上层请求,将一个标准的、虚拟的数据请求,解析为面向各异构文件服务空间的实际数据请求。

应用请求层:应用请求层包括发出文件请求的各种应用程序,其中有标准的文件系统请求、其他应用程序的数据调用请求等。

4.3 系统实现

本系统的实现主要依赖两部分功能模块:标准用户文件访问接口和存储访问代理,如图7所示。

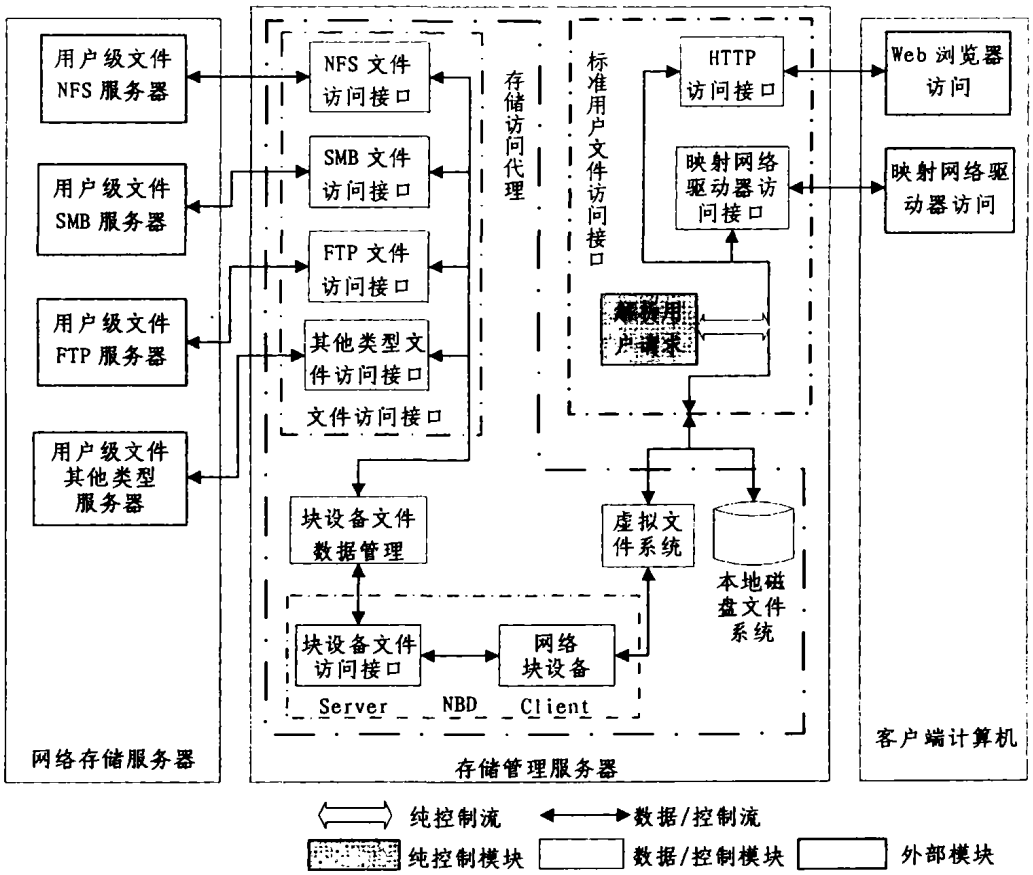

图7 分布式网络存储管理系统实现模型

根据标准文件访问接口协议,"标准用户文件访问接口"模块在"本地磁盘文件系统"和系统构建的"虚拟文件系统"上获取数据,并向客户端访问进程提供标准格式的文件。"HTTP 访问接口"和"映射网络驱动器访问接口"分别接收来自用户的请求,并向系统提出文件系统级请求。"解析用户请求"模块执行"会话解析层"功能,根据被请求文件的绝对路径,判断文件的存储位置,指导文件访问。由于"本地磁盘文件系统"和"虚拟文件系统"的同构性,数据处理层的功能透明。

"存储访问代理"获得"标准用户文件访问接口"发出的数据请求,根据"标准文件访问接口协议",从分布和异构的网络存储空间中获取数据,并应用 NBD 技术构建虚拟文件系统。如上文所述,系统在网络块设备的基础上构建虚拟文件系统。网络块设备的数据来自于 NBD-Server 进程读取的、存储在网络存储空间中的用户数据。NBD-Server 作为"标准文件访问接口协议"的应用请求层的应用程序请求,向网络存储空间请求数据。由于用户请求已经在"标准用户文件访问接口"中被解析,此处会话解析层的功能透明。"文件访问接口" 实现代理访问层功能,其中包含 NFS、SMB、FTP 等协议的访问接口。"块设备文件数据管理"模块对得到的数据进行 Raid5协议处理和加密/解密处理,完成数据处理层功能,并最终将完整、准确的数据交给 NBD-Server 构建网络块设备。

结束语 本文所研究的"分布式网络存储管理系统"从解决移动办公问题的角度出发,提出了区别于一般的网络存储技术的、全新的分布式网络存储概念。系统以用户已有的网络存储空间为基础构建用户的虚拟个人存储空间,并向用户提供比传统网络应用服务更为可靠和安全的数据管理和服务。目前系统已经在提供 NFS 和 SMB 服务的网络存储空间上进行小规模测试,运行效果良好。

参 考 文 献

1 刘启新. 虚拟个人网络空间及其在网络教学系统中的应用:[硕士论文]. 清华大学计算机科学与技术系. 2001.1
2 蒋东兴,刘启新,戚丽,石岩. 虚拟个人网络空间及其关键技术研究. 清华大学计算机与信息管理中心. 2001.10
3 Machek P. Network Block Device. http://atrey.karlin.mff.cuni.cz/~pavel/nbd/nbd.html
4 张凌,蒋东兴,刘启新,周霖,沈培华. 分布式网络存储管理系统的研究. 中国国际存储技术大会. 2002