

一种新的专用附网存储系统的研究与实现^{*}

Study and Implement of a New Special Network Attached Storage System

韩德志^{1,2} 谢长生²

(河南信阳师范学院)¹(华中理工大学计算机学院外存储国家专业实验室 武汉430074)²

Abstract This article designs and implements a kind of network server system. This storage system has very good I/O performance by experiment. If it is used as a special network storage server, it will greatly mitigate the I/O bottleneck of network storage.

Keywords Network attached storage, Request reduce mechanism, Direct access mechanism, Multi-stage striping

1 引言

随着 Internet/Intranet 技术的发展,网络上的信息资源呈爆炸型增长趋势,通过网络进行传输的信息量不断膨胀,大量信息需要进行处理,并通过网络传输,这对信息存储系统提出了空前的要求。传统的存储体系都是存储设备通过诸如 IDE/SCSI 等 I/O 总线与服务器相连。客户机的数据访问必须通过服务器,然后经过 I/O 总线访问相应的存储设备,服务器实际上起到一种存储转发的作用。连接的客户数量增多时,I/O 总线将会成为一个潜在的瓶颈,并且会影响到服务器本身功能,严重情况下甚至会导致系统的崩溃。这种附属于网络服务器的存储方式已不能适应来自应用的越来越高的要求。因此,探索新的存储体系结构就显得非常必要。

目前,出现了三种新的存储体系结构:附网存储、存储局域网和 Jmi 技术。存储局域网虽然性能好,但其价格贵,成本高,一般的中小企业和单位根本无力建构这种昂贵的网络;Jmi 技术虽然前景比较乐观,但还在研制和试用阶段;附网存储技术由于其价格低廉、安装简易、具有高速等特性,具有广泛的应用前景。

2 新的专用网络存储服务器系统的整体结构

本文在国家自然科学基金项目以及华中理工大学外存储国家专业实验室和美国 Rhode Island 大学合作项目的资助下,为提高网络存储速度的要求,设计并实现了一种全新的专用网络存储服务器系统。该系统重

点对软件进行了优化,通过实际测试表明,该存储系统与传统的依附于服务器的存储方式相比,读写速度有大幅度提高。

专用的网络存储服务器对不同的网络提供不同的访问方法。对于内部网络中的用户,可以通过提供 NFS、FTP 和 CIFS 协议来存取服务器上的数据,用户还可以通过 Web 浏览器读取服务器上的文件资源;对于 Internet 上的或者是公用网络的用户由于不能使用 NFS、CIFS 这些服务,用户可以通过 Web 方式浏览服务器上的数据,或通过 FTP 方式访问服务器上的数据。另外由于专用网络存储服务器是一个黑匣子的结构,对于其管理我们可以通过 Telnet 协议或者是远程的 Web 管理的方法,有权管理权限的用户只需要利用 Telnet 或者是 Web 浏览器就可以很容易地对服务器上的数据、用户的存储权限等参数进行相应的管理。由于特别是对软件进行优化,这种专用的网络存储服务器具有极高的读写响应速度。本文我们将重点介绍专用网络存储服务的软件组成及优化。

2.1 硬件组成

专用网络存储服务器强调的是网络和存储的功能,所以在构造硬件结构中可以只考虑网络和存储两方面的问题。系统除了 CPU 单元以外,最主要的也就是存储模块和网络功能模块,存储模块主要是提供对 IDE/EIDE、SCSI 总线技术的支持,提供工业标准 EIDE 控制器、SCSI 控制器、阵列控制器,使得系统可以任意挂接各种存储设备,如光盘塔、磁盘阵列等。网络控制模块,实际上是实现网络适配器的功能,用于进行数据帧的生成、识别与传输,数据编译,地址译码,数据

^{*} 本文受国家自然科学基金项目资助。韩德志 硕士研究生,研究方向:计算机体系结构及附网存储技术。谢长生 教授,博士生导师,研究方向:计算机海量存储结构系统及高性能 I/O。

传输的出错检测和硬件故障的检测等等,它最终提供一个普通网络接口和高速光纤通道接口,提供100MB或更高的速率,使得系统可以方便地与以太网相连或挂接在高速光纤通道连成的存储局域网上。CACHE/ROM/RAM模块为系统提供CACHE, RAM, ROM, FLASH空间,系统核心操作系统和相关系统软件都可以固化在ROM或者FLASH RAM上,系统启动时可以直接引导ROM/FLASH中固化的程序。

2.2 软件组成

整个系统必须有一个核心操作系统的支持,见图1,在核心操作系统之上是对应的设备(如网络设备,存储设备)的设备驱动模块。作为网络存储服务器我们更多的要考虑网络和存储方面的功能,所以在设备驱动上一层应该是对一些基本网络协议的支持(如TCP/IP, UDP/IP, SPX/IPX),然后是对专用网络存储服务器进行网络数据访问的一些文件共享协议(如NFS, SMB, CIFS, NCP),在网络协议之上才是专用网络存储服务器提供的一些网络应用,如远程网络管理等。在核心操作系统之上,设备驱动模块之下是I/O优化模块,主要对读写性能优化。

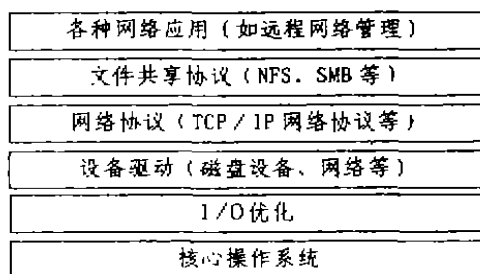


图1 专用的网络存储服务器软件模块图

3 专用网络存储服务器核心操作系统的选择及Web管理

3.1 核心操作系统的选择

由于存储服务器专门用作存储的服务,它强调数据访问的性能,为了满足大量客户端的同时的数据访问,它要求操作系统能够对I/O请求进行迅速及时的处理,以便数据能及时返回给客户端,所以存储服务器对实时性要求比较高。

在我们的开发工作中先后使用了DOS, SCO-UNIX作为服务器的控制软件的运行平台,从试验取得的数据来看,这两种操作系统存在以下不足:(1)在任务间切换性能方面,DOS操作系统本身在API一级并不支持多任务,对于需要多个任务之间转换的实现比

较复杂且效率低下;SCO-UNIX操作系统虽然支持多任务,但是它本身存在任务间切换时延较长的缺陷,不适合于附网存储服务器这种对实时性要求比较高的场合。(2)为了改善附网存储服务器的总体性能,我们在系统软件中添加了缓存管理部分来加快附网存储服务器的响应速度,提高数据传输率。由于DOS操作系统的API不直接支持对大块内存的使用而使得缓存管理无法申请大块连续内存,最终导致DOS系统中的缓存管理效率较低;在UNIX核心中,一次申请的最大内存块仅为64kB,虽然我们努力找到了解决办法,但是这样做使得软件的可移植性受到影响。

由于存在以上这些问题,我们选择了一种适合于附网存储服务器特点的操作系统Linux。Linux具有下列特点:Linux是一个具备POSIX所有特性的操作系统;是一个开放式的操作系统,任何用户都可以为它开发相关的驱动程序和应用程序;共享内存页面,在Linux下,多个进程可以使用同一块内存页面(每块大小为4k);优秀的磁盘缓冲调度功能;软件移植性好;强大的网络功能。Linux操作系统本身就是在Internet上成长起来的,所以它提供了全面的网络支持。另外,最为关键的是Linux是一个免费的操作系统。

3.2 网络存储服务器Web管理

由于网络存储服务器是一个黑匣子的结构,对于其管理我们只能通过网络进行远程管理,对服务器进行网络管理有如下两种方式:(1)管理员远程Telnet到服务器上,然后利用相应的命令对资源进行管理,(2)远程的Web管理,有权限的用户只需要Web浏览器就可以很容易地对服务器上的数据、用户的存储权限等参数进行相应的管理。这种方法,拥有较好的GUI界面,用户管理方便,适用于远程管理,是目前最为流行的一种方法,它使得网络管理员能够使用Web来监视和控制网络资源。网络上的任何用户,只要拥有适当的管理权限和一个Web浏览器,就可以在任何位置、任何操作系统平台上对网络存储服务器进行管理,由于Web浏览器对于所有的平台几乎都是一个标准配置,因此这种管理方式对于用户来说,就不用安装任何的客户端软件了,用户需要的仅仅是一个Web浏览器。

4 读写性能的优化

专用的网络存储服务器除了网络功能外,主要是存取功能,为了优化专用存取服务器的读写性能,我们主要采用如下四种方法:

(1)通过I/O磁盘和用户内存的DMA传输(而不

是通过服务器 cache 内存缓冲)来提高 I/O 读写速度:

(2)通过开发一个具有多级分流的设备驱动程序使我们的专用的网络存储服务器能同时挂多个磁盘阵列,使多个磁盘阵列组成一个具有一个高速的大磁盘阵列。

(3)为了减少请求数,实现 I/O 盘和用户内存的 DMA 传输,我们也开发了一个磁盘碎片整理程序,在系统空闲时,对整个磁盘系统进行整理,把每一个文件尽量放在相邻的盘块。

4.1 直接的访问机制

为了避免对象磁盘阵列、光盘塔这样的设备读写

都要经过服务器 cache 缓冲然后再到用户内存,使读写直接在用户目的缓冲区内和存储磁盘之间完成,应在逻辑文件系统(简称 LFS)和设备驱动程序中建立新的命令流和数据流,设计直接访问设备驱动程序的新功能,它们包括起始扇区号、扇区数和在虚拟内存中用户缓冲区的地址。我们称这种经过流的命令请求和数流传输机制叫做直访问机制。图2显示了传统的和我们设计的栈传输过程图,2(a)显示了传统的读信息传输过程,它分为两步:第一步把磁盘信息输入到服务器核心高速缓存 cache;第二从服务器核心 cache 传入用户内存。

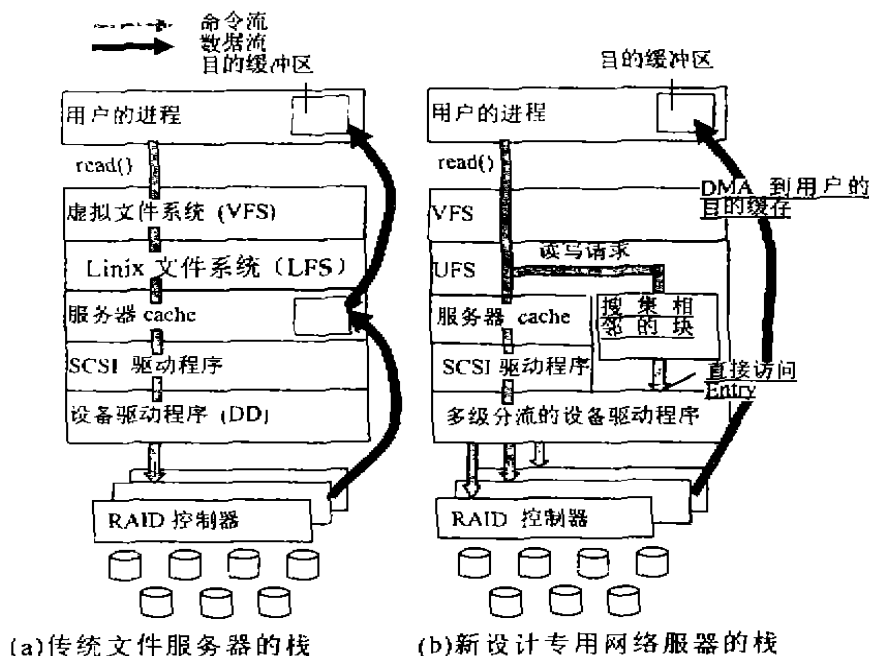


图2 传统的文件服务器和新设计的专用服务器处理栈

图2(b)是我们设计的传输栈,在获得文件的读写请求时,LFS首先了解文件的分块情况,在磁盘上搜索相邻的块,多级分流的设备驱动程序把相邻的块视为一个大文件块(因为搜索相邻块进程传给多级分流驱动程序的直接访问入口,该入口包括相邻的多个块的第一个块号和相邻块数及对应的虚拟内存的地址);接着,在磁盘阵列控制器和用户进程缓冲区之间直接实现DMA传输,而对其它象建立目录或访问单个磁盘信息则取道服务器cache缓存。

4.2 多级分流的设备驱动程序

在我们设计的附网存储服务器系统中,我们专门为RAID控制器PCI卡开发了一套多级分流设备驱动程序,该程序具有两个功能,一是多级分流功能,能同时访问两个或更多个磁盘阵列控制器,并把多个磁盘

阵列构成一个具有高速的大磁盘阵列;第二个功能是提供直接访问机制,每个文件块地址信息包括在用户虚拟内存中的缓冲地址以及本块在磁盘中起始扇区号和扇区数。多级分流的设备驱动程序主要有两个功能,一是把虚拟地址转换成物理地址,二是搜索相邻的块也就是在页表中看哪些页是相邻的(在Linux中一个文件块可能包括若干页,块相邻,组成它们的页也自然是相邻的),然后就每一组相邻的页向RAID控制器发一个有DMA功能的请求,RAID控制器根据多级分流的设备驱动程序提供的地址信息,把请求中给出的所有相邻的块通过DMA方式一次性传入用户缓冲区内,然后再由用户缓冲内存通过网络传到用户的计算机。

4.3 磁盘碎片整理程序

为了更好地提高专用附网存储服务器的性能,在读写一个文件时,尽量少发读写命令,使每发一次读或写命令时,尽量读或写更多的数据块(因为每发一次读写命令,执行读或写命令时都需要一定的时间),在此基础上实现 DMA 传输机制,我们在 Linux 环境下开发了一个碎片整理程序,定期对服务器挂接的磁盘存储系统进行整理,以保证一个文件的数据块能尽量存放在相邻的扇区。

5 试验评估

5.1 服务器连续读速度测试

对我们的系统先进行试验评估代替数据分析,这里 NR 和 ND 分别代表 RAID 控制器数和总的硬盘数,文件读请求块尺寸用 BS 表示。在表1中,试验是对我们所设计的服务器和普通文件服务器进行的,CPU 都是 Pentium 500,内存都为128MB,试验时让它们分别挂接1~3个RAIDO,每个RAIDO 所含有18硬盘,文件的分块尺寸分别从32k到64k,请求文件的大小分别为1MB到10MB,然后分别测试它们的读速度。

表1 在不同的磁盘阵列数和文件分块下的连续读速度测试

		专用网络存储服务器			传统文件服务器		
RAIDO		连续读速度(M字节/秒)			连续读速度(M字节/秒)		
文件分块	请求文件的块大小 BS	RAID 控制器数 NR			RAID 控制器数 NR		
		1	2	3	1	2	3
32KB	1MB	31.1	45.2	58.2	20.2	39.9	50.8
	50kB	33.8	47.4	68.5	23.4	40.8	55.0
64kB	1MB	59.5	67.5	82.9	43.5	59.3	72.6
	50MB	61.8	77.9	92.5	44.1	67.2	80.2

从表中我们可以看到,在同样 CPU 和内存的情况下,专用的网络存储服务器比传统文件服务器读速度有大幅度提高,最大提高幅度63%,最小幅度为13%。

5.2 专用附网存储服务器连续写速度测试

在连续写速度测试时我们用有18个磁盘的 RAID

IO,文件分块大小为32kB~64kB,RAIDO 数从1~3,测试结果如表2所示,从表中我们可以看出专用的附网存储服务器对连续写也有大幅度提高。

表2 文件块大小为64kB和RAIDO数为3连续写速度的测试

		专用网络存储服务器			传统文件服务器		
RAIDO		连续写速度(M字节/秒)			连续写速度(M字节/秒)		
请求块大小 BS	每个RAIDO所含磁盘数 DN	9	12	18	9	12	18
1MB		40.1	48.2	54.2	28.6	38.2	46.1
50MB		59.5	55.3	65.3	48.1	45.3	54.9
100MB		64.3	60.1	76.4	52.8	48.9	63.2

从上表可以看出,在同样的条件下,专用的网络存储服务器连续写速度比普通文件服务器速度有大幅度的提高,最大提高幅度为40.2%,最小幅度为17.6%。

小结 网络存储是近年来国际上比较热门的一个研究方向,而附网存储技术又是网络存储中较新的课题,本文所设计的专用附网存储服务器系统,通过试验验证,读写性能良好,下一步我们将进一步完善其功能,并将其成熟的产品推向市场。

参考文献

- 1 ANSI SCSI-3 Fast-20 Parallel Interface X3T10/1047D Working Group, Revision
- 2 Penner A F. Fibre Channel Gigabit Communications and I/O for Computer Networks. McGraw Hill New York, 1996
- 3 MoDoetal D. A Quantitative Analysis of Cache Policies for Scalable Network File Systems 15th SOSOP Dec. 1995
- 4 Hu Yming, Yang Q. DCD-disk caching disk. A New approach for boosting I/O performance. In: Proc of 13th Int Symposium on Computer Architecture (Philadelphia Pennsylvania), May 1996. 169~178
- 5 Tsujoka T, Onoda T. An ultra high-speed file server with 105 Mbytes/s read performance based on a personal computer. IEICE Trans. Commun. 1998, E81-B(12): 2503~2508
- 6 Deering S, Hinden R. Internet Protocol, Version 6 (IPv6) Specification. RFC2460, Dec. 1998
- 7 Blake S, Black D, Carlson M, et al. An Architecture for Differentiated Services. RFC2475, Dec. 1998
- 8 Nichols K, Blake S. Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Header. RFC 2474, Dec. 1998
- 9 邓志成,王茜,等. 区分服务体系的服务质量机制浅析. 计算机科学, 2000, 27(6)
- 10 林闯,单志广,盛立杰,等. Internet 区分服务及其几个热点问题的研究. 计算机学报, 2000, 23(4)
- 11 Braden R, Zhang L, Berson S, et al. Resource ReSerVation Protocol (RSVP)-Version 1 Function Specification. RFC 2205, Sep. 1997
- 12 Wroclawski. The Use of RSVP with IETF Integrated Services RFC2210, Sep. 1997
- 13 Almqvist P. Type of Service in the Internet Protocol Suite RFC1349, July 1992

(上接第75页)

参考文献