

基于 iSCSI 协议的集群光盘服务器的研究与设计^{*}

万继光 詹 玲

(华中科技大学计算机学院 信息存储系统教育部重点实验室 武汉 430074)

摘 要 随着网络上光盘资源的增长,光盘服务器成为光盘网络共享的重要技术。针对传统光盘服务器的缺点,设计并实现了一个基于 iSCSI 协议的高性能集群光盘服务器 CCS(Cluster CD Server)。CCS 系统采用高自治、自维护的集群策略,提高了系统的并行性和可扩展性;采用逻辑的名字空间管理,既实现了单一名字空间,又避免了单点失效;采用虚拟技术并对 iSCSI 协议进行了精简和优化,减少了网络小包的传输数量,提高了网络带宽的利用率。在详细介绍了 CCS 系统的设计方案的同时,进行了相应的试验测试和性能分析。

关键词 光盘服务器, iSCSI, 集群, SAMBA, 单一名字空间

Research and Design of an iSCSI-based Cluster CD Server

WAN Ji-Guang ZHAN Ling

(Key Laboratory of Data Storage System, Ministry of Education, Huazhong University of Science and Technology, Wuhan 430074)

Abstract With the increasing of CD data in internet, CD Server has become the new important technology of network share. Considering the disadvantages of the traditional CD server, we presented a novel iSCSI-based High Performance Cluster CD server which adopted high autonomy and maintenance by itself to improve the parallelism and the expandability of the system; the CCS adopted logic namespace management to realize a simple namespace and avoid Single Point of Failure; and it simplified and optimized iSCSI protocol by using virtual technology to reduce the quantity of transmission of the network small bag and improve the utilization ratio of the bandwidth of the network. The paper introduced the design of the CCS in detail, and performed an intensively performance evaluation study on the CCS.

Keywords CD Server, iSCSI, Cluster, SAMBA, Single namespace

1 前言

数据光盘因为其价格低廉、容量大、使用方便等多种优势已成为目前重要的信息载体。但随着光盘数量急剧增加使得其在管理、使用及数据共享等方面困扰广大用户,因此光盘服务器作为一种新的技术应运而生,在很大程度上解决了这一问题,使客户机能以硬盘的访问速度来共享光盘上的信息资源,消除了 CD-ROM 驱动器瓶颈问题,极大地改善了光盘网络共享的性能。

现有的光盘服务器在光盘共享方面,一般采用 SMB/CIFS 协议, SMB/CIFS 协议功能强大是它的优点,但是,它的性能比较低,特别是刚刚开始访问时,响应时间特别长,让用户难以接受。另一方面, SMB/CIFS 协议只能够在局域网内使用,它不支持在 Internet 远程访问,这大大限制了光盘资源的共享,在 Internet 网飞速发展的今天,不支持 Internet 远程访问是用户难以接受的。

随着计算机和网络技术的飞速发展,数字化信息呈爆炸性增长的趋势,单一的光盘服务器已经不能够满足用户的需要,而且用户对存储系统的可用性和可扩展性提出了更高的要求,这些都导致了一种新的光盘服务器(集群光盘服务器)的出现。

在国家自然科学基金和 973 项目的资助下,针对 SMB/CIFS 协议的缺点和光盘共享服务的特点,采用 iSCSI 协议和

服务器集群技术,我们设计实现了一种基于 iSCSI 协议的高性能集群光盘服务器 CCS(Cluster CD Server)。本文第 2 节详细描述了 CCS 系统采用高自治、自维护的集群解决方案;第 3 节介绍了 CCS 的软件架构;第 4 节通过实例给出了逻辑名字空间管理方法;第 5 节在分析 iSCSI 协议的基础上,提出了一种 iSCSI 协议的性能优化技术;最后在第 6 节对 CCS 系统进行了详细的性能测试和分析。

2 自治式集群方案

光盘服务器的特点是只提供对光盘数据的只读网络共享,不需要对存储的数据进行修改,因此 CCS 采用了图 1 的集群结构,它借助一个客户端代理来实现集群的功能。避免了集中管理的性能瓶颈,采用全新的设计思想,改变传统的方式把所有用户请求都发给管理服务器方式,而是让用户请求分散到各个服务器,这样大大增强各个服务器的并行度,实现了负载均衡,从而提高了整个集群的性能。

采用高自治、自维护的集群策略,各个服务器能够独立地维护自己的存储资源和光盘的原数据及数据本身,并能够独立提供光盘服务,从而提高系统的可扩展性。采用数据放置策略同数据迁移相结合的技术,维护各服务器的负载均衡。

CCS 系统用户请求处理流程如图 1,用户通过光盘客户端可以向任何一台光盘服务器发请求(如步骤(1)),如果请求的数据存放在此光盘服务器上,就返回请求光盘的数据,否

^{*} 本课题得到国家自然科学基金(60303031)和国家“九七三”重点基础研究发展计划(2004CB318203)资助。万继光 博士研究生,主要研究领域为网络存储系统;詹玲 博士研究生,主要研究领域为网络存储系统。

则,就返回数据存放的光盘服务器 IP(如步骤(2)),然后光盘客户端再自动向相应的光盘服务器发请求(如步骤(3)),最后光盘服务器返回请求的光盘数据(如步骤(4))。

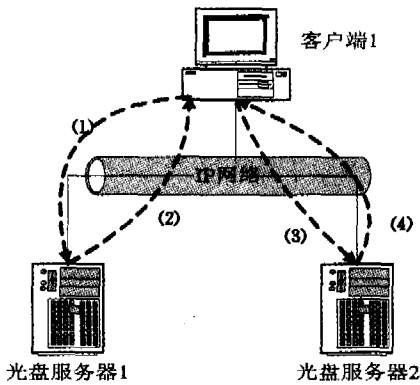


图1 CCS系统用户请求处理流程

3 CCS 的软件架构

CCS 系统采用虚拟光驱技术同镜像技术相结合,实现了光盘的数据块级共享,能够保证光盘数据的完整性,从而解决了单机版光盘的网络共享问题;采用成熟的 iSCSI 协议,并进行了精简和优化,不仅减少了用户响应时间,而且提高了连续访问的数传率,而且 iSCSI 协议是基于 IP 协议的,能够支持 Internet 上的远程访问;为了进一步优化系统性能,针对光盘访问特点,设计了一种两级 Cache 技术(客户端 Cache 加服务器 Cache),从而大大提高了系统的性能;根据用户的需求,还设计了针对光盘共享的灵活计费 and 用户权限管理。

CCS 的系统结构如图 2,按照功能划分,CCS 可以分为九个模块,客户端包含客户端 Cache 模块,客户端 SN(Single Namespace)模块和 iSCSI Initiator 模块,服务器包含全局管理模块、iSCSI Target 模块、访问控制模块、服务器 Cache 模块、服务器 SN 模块和数据备份模块。

其中,iSCSI Initiator 和 iSCSI Target 模块实现了 iSCSI 协议的功能。客户端 SN 和服务器 SN 模块一起实现了 SN 功能。数据备份模块实现了不同光盘服务器之间的数据备份和迁移,从而实现系统的高可靠性和负载平衡。

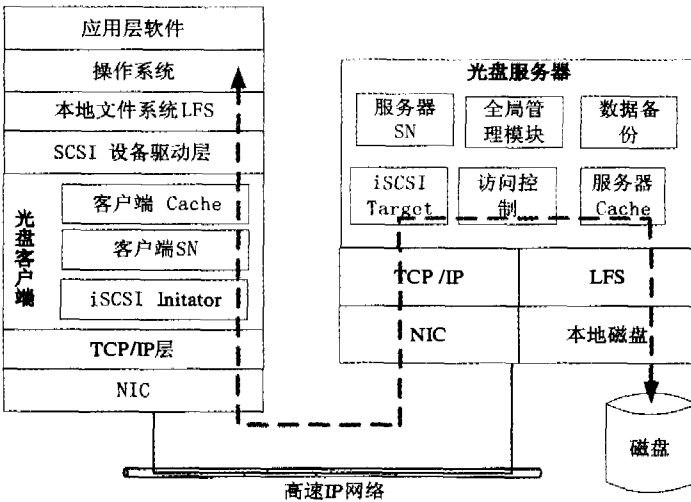


图2 CCS系统结构图

4 逻辑名字空间管理

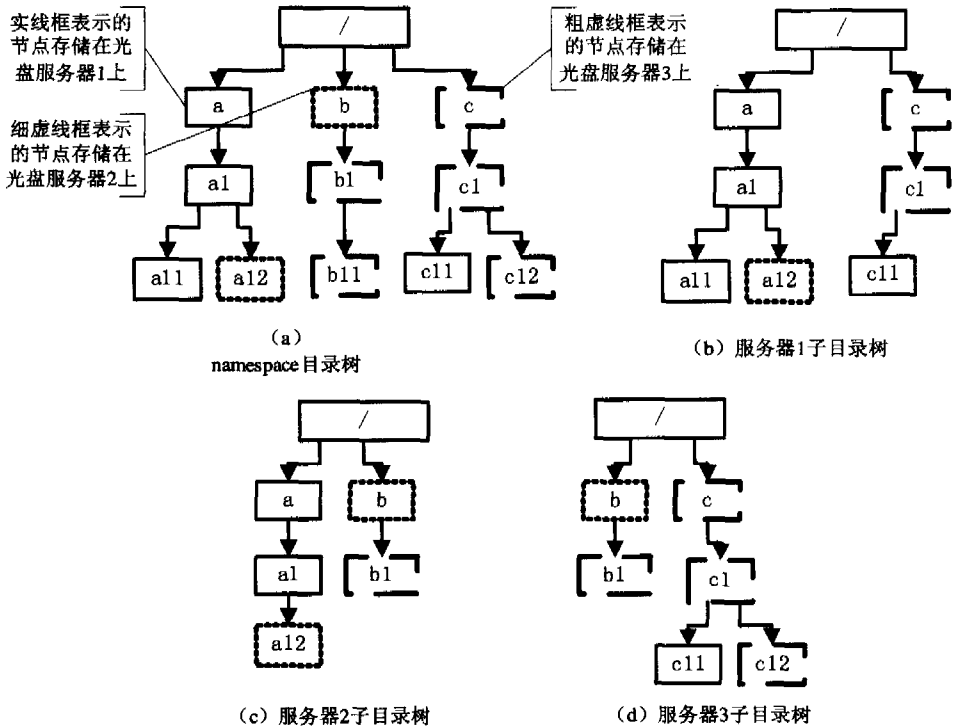


图3 namespace 目录树和各个服务器的子目录树。

采用逻辑上的单一名字空间服务器 SNS(Single Namespace Server)来实现整个系统的 SN 管理,SNS 由各个服务器选举产生(例如,采用 IP 地址最小的服务器担任),其管理的名字空间原数据由各个服务器汇总产生。一旦 SNS

失效,系统就会自动产生新的 SNS,因此不存在单点失效。SNS 只保存 namespace 的一级目录和同父目录不在同一个服务器的目录或文件信息,这样 SNS 保存的全局目录信息很少,且相对静态。

按照上面的数据放置策略,每个服务器上的原数据以一个目录子树的形式存在,各个服务器将它的目录子树提交给 SNS,由 SNS 生成一个全局目录树,也就是 namespace 目录树。图 3 是 namespace 目录树的一个实例。

为了提高性能,SNS 生成 namespace 目录树后,将它发给所有的服务器和客户端,客户端代理有这个 namespace 目录树,就能够直接映射请求到不同的服务器,大大减少了访问 SNS 的次数,从而减轻 SNS 的负载。

为了提高各服务器的自治性,各个服务器不仅保存它上面的文件和目录,也备份跟这些文件和目录相关的文件和目录,组成一个 namespace 目录树中的子目录树,如图 3。需要备份的文件和目录信息分以下三类:

- (1)任何一个服务器上的目录,它在 namespace 目录树所有其它服务器上的祖先目录都在此服务器上有备份,只备份目录结构,不包含文件。
- (2)当一个目录同它的子目录不在同一个服务器上时,它的子目录在它所在的服务器上有备份。
- (3)原则上一个目录下的文件都在同一个服务器,但是当服务器没有存储空间时,也需要将目录下的部分文件移动或新建到其它服务器上,这种情况下,此目录下保留文件的影子文件,这个影子文件只包含文件的原数据,不包含文件的数据本身。

5 iSCSI 协议的分析及针对光盘共享的性能优化

在网络数据传输上,CCS 采用了 iSCSI 协议,并对 iSCSI 协议进行了性能优化。iSCSI 协议最先由 IBM、Cisco 和 HP 等公司提出,现在已经被 IETF 采纳并准备将其标准化。iSCSI 协议定义的是 SCSI 到 TCP/IP 的映射,即将主机的 SCSI 命令封装成 IP 数据包,在 IP 网络上传输,到达目的节点后,再恢复成封装前的 SCSI 命令,从而实现 SCSI 命令在 IP 网络上的直接、透明传输。由于 TCP/IP 协议的飞速发展,iSCSI 协议已经成为构建 IP-SAN 的重要协议。

iSCSI 协议实质是在一个 IP 隧道上跑 SCSI 命令,是一种 SCSI 远程过程调用模型,而 Initiator 端对所有的 SCSI 命令不进行任何处理。这样,因为 SCSI 命令中有很多控制命令,这些命令的包都很小,而在网络上传输小包效率很低。采用 IoMeter 测试工具对 iSCSI 协议测试得到的 IP 包长度分布的情况如图 4,可以看出将近一半的网络包小于 127 字节,而以以太网数据包最大长度为 1518 字节,这导致网络物理带宽的严重浪费。

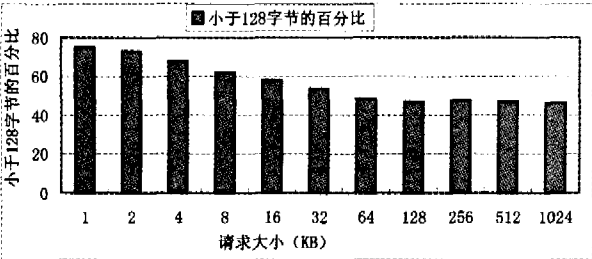


图 4 iSCSI 协议小于 128 字节的包百分比

针对光盘共享的特点,为了提高网络传输的性能,采用简化 iSCSI 协议的思想,设计了一个精简 iSCSI 协议。主要设计思想是通过采用虚拟技术,在 Initiator 虚拟 SCSI 光盘驱动器,并将接收到的大部分能够处理的 SCSI 命令在 Initiator 直接处理,只有读光盘数据等少数几个 Initiator 不能够处理的

SCSI 命令才通过 iSCSI 协议传输到 Target 端。这样,能够大大减少网络传输过程中的小包。

其请求处理过程如图 5,用户程序通过 Windows API 接口向操作系统发送请求,如步骤(1);操作系统将接收到的请求转换为 SCSI 命令发给虚拟光盘驱动,如步骤(2);虚拟光盘驱动采用虚拟技术将收到的命令能够处理的直接虚拟一个结果,然后将其返回操作系统,如步骤(7);只有 READ_CD 等少数几个不能够虚拟处理的命令转发给网络 TCP 层,如步骤(3);网络 TCP 层再将 READ_CD 等 SCSI 命令发给光盘服务器,如步骤(4);光盘服务器处理 READ_CD 等 SCSI 命令,并将读取的数据通过网络发给客户端的网络 TCP 层,如步骤(5);然后,数据经过虚拟光盘驱动和操作系统,最终返回给用户程序,如步骤(6)(7)(8);至此,用户的一个请求过程完成。

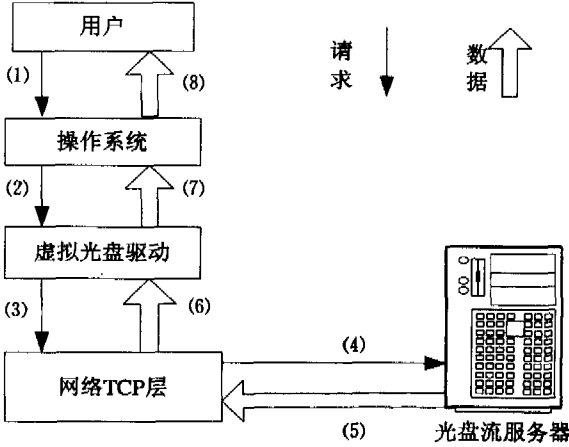


图 5 精简 iSCSI 协议用户请求处理过程图

通过对 iSCSI 协议进行优化,CCS 系统性能有明显的提高,图 6 是对 CCS 测试得到的 IP 包长度分布的情况,其小包的比例明显比 iSCSI 少很多。

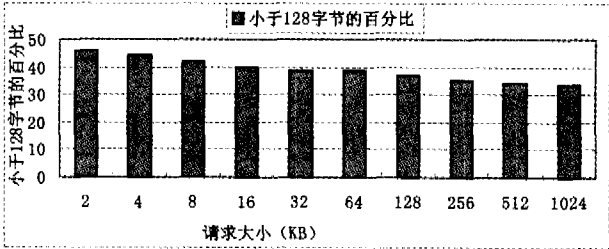


图 6 精简 iSCSI 协议小于 128 字节的包百分比

6 性能测试与结果分析

存储系统性能评价领域内有许多性能指标,本文采用的性能指标为吞吐率;吞吐率是指存储系统在单位时间内能处理的 I/O 请求数量,有两种形式:I/O 速率,即每秒钟 I/O 访问次数;数据传输率,简称数传率,即每秒钟传输的数据量。在事务处理应用中追求高的 I/O 速率,在光盘应用中则要求较高的数传率,本文的测试采用数传率。

因为 CCS 系统采用自治式集群方案,且光盘服务器对外只提供读服务,光盘服务器上的数据很少修改,从而光盘服务器之间的交互很少,因此各个光盘服务器是独立对外提供服务的。整个 CCS 系统的性能是各个服务器的性能之和,当各个客户端的请求平衡分布到各个服务器上时,整个 CCS 系统

能够获得最好的性能,当各个客户端的请求集中到一个服务器上时,系统性能降到最低,因此用户的请求分布情况对 CCS 系统性能影响很大,动态负载平衡是 CCS 系统下一步要解决的问题,这里通过分析单个服务器的性能来说明 CCS 系统的性能。

6.1 性能测试平台

为了验证 CCS 服务器的性能,本文分别在 100M 以太网和 1000M 以太网上对其进行了性能测试。测试是在两台通过 IP 交换机连接起来的 PC 机上进行的,其中一台是客户机,另一台是服务器。两机的配置大致相同: Intel Celeron 1.8GHZ CPU; 256M RAM; 硬盘: Maxtor 60GB ATA Disk, 最大数传率 40MB/s; 100Mb 和 1000Mb 网卡; 操作系统: 服务器为 RedHat 9.0, 客户端为 Windows 2003。为进行对比,在此平台上分别测试了 CCS 服务器、SAMBA 服务器和 IDE 磁盘的性能。SAMBA 服务器和 IDE 磁盘的测试采用 Iometer 2003.05.10 测试工具,因为 Iometer 2003.05.10 测试工具不能够测试光盘,CCS 服务器采用自己开发的 CCS 测试工具。

6.2 测试数据及分析

图 7,8 分别为 100Mb 以太网环境中,CCS 和 SAMBA 服务器的随机读数传率曲线和顺序读数传率曲线。随着 I/O 请求块大小的增加,二种服务器的数传率也都随之增大。总体来说,随机读 CCS 服务器性能略低于 SAMBA 服务器,但是顺序读 CCS 服务器性能要明显高于 SAMBA 服务器,在请求数据块大小为 1024kB 时,顺序读 CCS 服务器的性能达到 9MB/s,基本达到 100Ms 以太网的有效带宽,这时网络是影响性能的瓶颈。

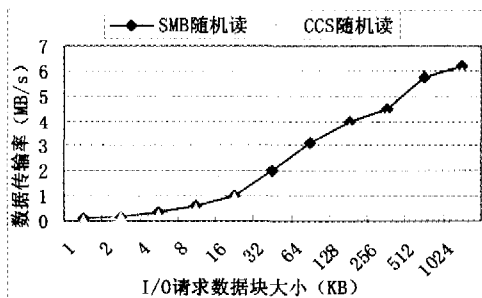


图7 随机读数传率曲线(100M 以太网)

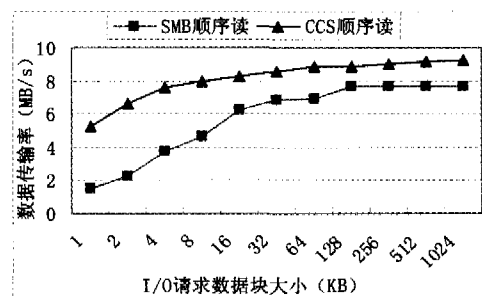


图8 顺序读数传率曲线(100M 以太网)

图 9,10 分别为 1000Mb 以太网环境中,CCS,SAMBA 服务器和 IDE 磁盘的随机读数传率曲线和顺序读数传率曲线。同样,随机读 CCS 服务器性能略低于 SAMBA 服务器,但是顺序读 CCS 服务器性能要明显高于 SAMBA 服务器。因为是在 1000Mb 以太网环境,网络不再是性能的瓶颈,在请求数据块大小为 1024kB 时,顺序读 CCS 服务器的性能达到 40MB/s,接近 IDE 磁盘的有效带宽,因此,这时磁盘是影响系

统性能的瓶颈。

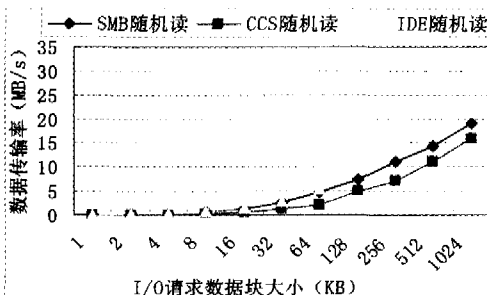


图9 随机读数传率曲线(1000M 以太网)

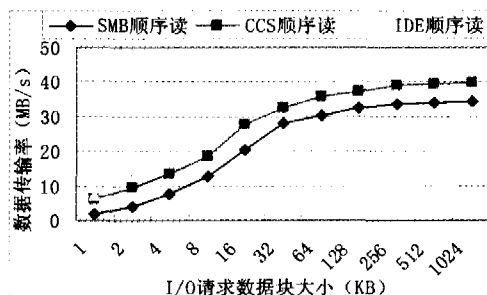


图10 顺序读数传率曲线(1000M 以太网)

总体来说,因为光盘的应用主要是顺序读,CCS 系统对顺序读进行了针对性的优化,因此 CCS 服务器顺序读性能比较高。在 100Mb 以太网环境中,网络是系统的瓶颈,为了充分利用网络带宽,CCS 系统对 iSCSI 协议进行精简,在数据传输过程中,尽量减少控制命令的传输,充分利用了网络的物理带宽。在 1000Mb 以太网环境中,IDE 磁盘是系统的瓶颈,为了充分利用磁盘带宽,CCS 系统采用 Cache 技术,对顺序读进行了磁盘读取优化,从而使 CCS 系统的顺序读对磁盘的带宽利用率比较高。

结束语 光盘服务器主要是针对光盘网络共享的应用来设计的,采用虚拟光驱技术同镜像技术相结合,实现了光盘的数据级共享,既能够保证光盘数据的完整性,又大大提高了光盘共享的性能,且能够支持远程共享。对于负载平衡问题,CCS 系统通过服务器之间数据迁移解决了静态负载平衡,尽可能保证各个服务器上的数据量平衡,但是没有解决动态负载平衡,也就是热点数据的性能问题,因此动态负载平衡是 CCS 系统下一步要解决的重点问题。

参考文献

- Gibson G A. Network attached storage architecture. Communications of the ACM, 2000, 43(11): 37~45
- Sachs M, Leff A, Seigny D. LAN and I/O Convergence: A Survey of the Issues. IEEE Computer, 1994, 27(12): 24~33
- Lin-wen lee, Scheuermann P. File Assignment in Parallel I/O Systems with Minimal Variance of Service Time. IEEE computer, 2000, 49(2): 127~140
- Boden N, et al. Myrinet: A Gigabit-per-Second Local Area Network. IEEE Micro, Feb. 1995
- Gibson G. Cost-effective high-bandwidth storage architecture. In: Proceedings of ACM ASPLOS, Oct. 1998
- Gabger R, Patt N. Using System-Level Models to Evaluate I/O Subsystem Designs. IEEE computer, 1998, 47(6): 667~678
- Anderson T F, et al. The Case for NOW(Networks of Workstations). IEEE Micro, Feb. 1995
- Strunk D, Goodson R. Self-Securing Storage: Protecting Data in Compromised Systems. In: Proc. of the 4th Symposium on Operating Systems Design and Implementation, San Diego, CA. Oct. 2000