一种面向下一代互联网的广域网智能存储系统

李洁琼 冯 丹
(华中科技大学计算机学院 武汉 430074)

摘 要 广域网智能存储系统针对下一代互联网数据急剧增长、网络资源难以管理和使用的问题,采用多层次、可扩展的分布式存储模式,从改进体系结构着手来提高网络存储系统的性能。其存储管理遵循存储管理计划规范(SMI-S),并针对复杂网络环境下的元数据管理和数据传输问题,提出了有效的负载均衡策略和高速安全的存储中间件解决方案,不仅降低了存储管理开销,加快了数据传输速度,同时也实现了命令与数据分流、扩容与增速同步的目标,从而大大提高了整个存储系统的性能。

关键词 广域网智能存储系统,存储管理规范,元数据管理,存储中间件

WAN Intelligent Storage System for Next Generation Internet

LI Jie-qiong FENG Dan
(Computer Department, Huazhong University of Scienceand Technology, Wuhan 430074, China)

Abstract Aimed at the rapid data growth and the difficulties of network resource management and use in next generation Internet, the WAN intelligent storage system(WISS) adopts multi-level and scalable distributed storage mode to improve the performance of the network storage system based on system architecture optimization. Its storage management follows SMI-S specification. And it proposed an effective load balancing strategy and high-speed secure storage middleware solution for metadata management and data transmission in complex network environment. WISS both improves data transfer rate, and reduces management cost. It achieves the goal of separating control and data flow and accelerating data transfer while expanding storage capacity at the same time. As a result, WISS enhances the performance of the whole storage system enormously.

Keywords WISS, SMI-S, Metadata management, Storage middleware

1 引言

网络数据存储发展到今天,国际学术界和工业界提出了许多新思路来迎接下一代互联网对存储提出的挑战,其中不乏很有价值的思路,例如:加州大学伯克利分校正在研究在临近磁盘位置进行计算的智能存储[1];卡内基梅隆大学提出使磁盘驱动器能下载执行代码到设备端完成的主动磁盘;威斯康星大学麦迪生分校利用磁盘驱动器能力进行文件系统分析并对数据进行优化存放的研究项目智能磁盘等。这些存储系统的出现缓解了数据急剧增长的需求,但也产生了网络资源难以管理、调度和难于使用的问题。

另外值得一提的是基于对象的存储[2]这一思路。卡内基梅隆等多所大学和IBM等公司正在进行该项目的研究,它的思想是将文件系统中操作磁盘的部分程序裁剪并移入到磁盘存储器中,使之成为一个包含数据和操作的对象,而驻留服务器中的文件系统上层只作文件的属性管理。这种思想虽有其优越处,但仍旧局限于设备层次,未能在体系结构上做出重大改变。

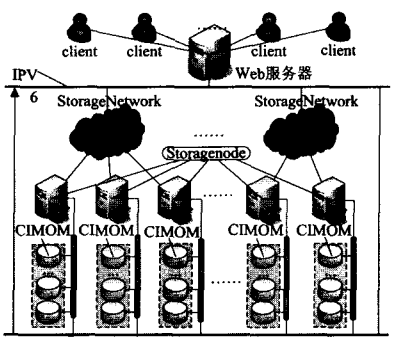

图1 系统硬件结构图

本文在国际存储领域前沿研究的基础上,认为应当从方便网络资源的管理、有利于存取调度和易于使用的角度出发,建立一种新的存储模式,即"多层次、可扩展的分布式存储"模式,并在此基础之上构建面向下一代互联网的广域网智能存储系统。

2 存储系统总体设计

2.1 体系结构

到稿日期:2009-11-06 返修日期:2010-02-20 本文受国家九七三重点基础研究发展计划项目(2004CB318300),国家八六三高技术研究发展计划项目(2009AA01A401/2009AA01A402)资助。

李洁琼(1978—),女,博士生,讲师,主要研究方向为网络存储技术、系统结构,E-mail:jieqiong_li@yahoo.com.cn;冯 丹 女,教授,博士生导师,主要研究方向为网络信息存储和系统结构。

面向下一代互联网的广域网智能存储系统支持 PB 级的总存储容量，支持至少 256 个分布式节点。如图 1 所示，整个系统由 3 个层次构成，分别为：存储子网(StorageNetwork)、存储节点(StorageNode)和存储设备(CIMOM)。系统由若干个存储子网(StorageNetwork)构成，每个存储子网又由若干存储节点(StorageNode)组成，存储节点一般搭建在局域网上，包含若干个存储设备(CIMOM)，这些存储设备在物理上的差异性一般很大。系统通过 Web 方式管理和展示分布在广域网中的存储节点，通过提供一个 PB 级虚拟存储服务器为用户提供存储服务。

系统软件结构如图 2 所示，存储节点通过代理由存储管理模块统一管理，存储管理模块建立在 SMI-S 协议之上，负责对各种异构的存储设备进行统一管理，并收集各存储节点的相关信息提交给元数据管理部分；相关信息被元数据模块处理成为全局视图的变量，交给上层的动态负载平衡模块，该模块协调各个存储节点之间的负载情况，并向存储中间件提供用户存储空间的相关元数据信息；存储中间件为用户提供存储服务，主要是数据管理和数据快速、安全的传输。整体系统采用分布式结构。

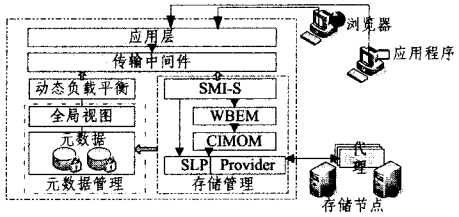

图 2　系统软件结构图

3　存储系统的详细设计与工作流程

3.1　存储管理

实验平台存储管理采用 JAVA 语言实现，在分布式的存储结构中，实现存储资源和拓扑的发现、网络存储资源和拓扑展示、网络系统信息的获取和展示。该管理软件遵循 SMI-S 的管理规范[3]，SMI-S 的目标是在存储网络的存储设备和管理软件之间提供标准化的通信方式，从而使存储管理实现厂商无关性，提高管理效率，降低管理成本，促进存储网络的发展。

3.2　元数据管理

元数据管理是广域网智能存储系统中的重要组成部分，它负责用户存储空间的分配，并向存储中间件提交用户存储空间的相关元数据信息。每个存储节点设有元数据查询表，它记录该存储节点中的各个存储设备中所存储数据的各种属性，是用户与其数据的交互接口。存储节点是由各种不同的存储设备构成，统一由 SMI-S 协议管理，因此这些设备的异构性对上层透明。

原型系统元数据管理流程如图 3 所示。首先根据用户登录的 IP 地址判断其所在地域，然后从所有存储子网中选择距离用户最近的一个作为服务区域。

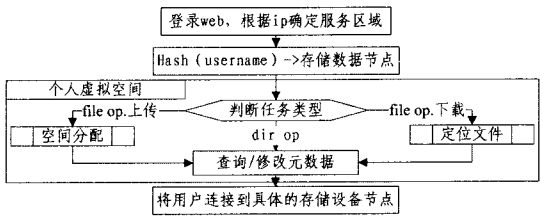

图 3　元数据管理流程图

再用哈希法确定分配给用户的存储节点，并在该存储节点建立一张元数据表，该表记录分配到该存储节点下所有用户的信息以及用户所管理文件的相关信息。

接着为用户在该存储节点上创建一个虚拟空间，该虚拟空间为每个用户在存储节点映射一个逻辑视图，显示用户在本系统中所上传的所有文件的属性，方便用户管理。

另一方面，用户在虚拟空间的操作分为两类，目录操作(dir op)和文件操作(file op)，目录操作是指文件移动、文件重命名、改变文件属性、删除文件、新建文件夹和打开文件夹之类的操作，这些操作不涉及到具体的数据传输，将不会和存储设备节点发生交互；文件操作是指上传、下载文件，这些操作将和存储设备节点发生交互。将用户操作分为两类的原因是，根据实际操作调查显示，目录操作的操作数量占用户操作总量的 80%左右，对于目录操作，只需修改存储节点的元数据表中的相应数据即可，从而实现命令与数据的分流，减少广域网带宽的占用。

存储空间的分配必须考虑系统动态负载平衡问题，设计有效的负载平衡策略[4,5]。

第一步计算负载指数，综合考虑各存储设备的剩余容量、带宽和上传文件的大小等影响存储设备负载的主要因素来计算负载指数。

第二步确定启动策略，采用任务触发启动策略。每一次数据上传任务触发负载平衡策略的启动，并由存储节点根据各个在线的存储设备的负载指数，为数据指定合适的存储设备。

第三步确定信息策略。需要解决信息收集间隔时间、信息收集方式和信息保存方式这 3 个问题。原型系统在系统网络拓扑发生变化时收集各存储设备的负载信息，采用数据库收集的方式，每个存储设备的负载信息由其上层的存储节点管理。

第四步确定数据迁移策略。采用阈值上限和下限。当某存储设备节点的负载指数高于阈值上限，则通知管理该设备的存储节点，要求迁移该设备最近接受的任务；如果某存储设备的负载指数低于阈值下限，则通知管理该设备的存储节点，要求分配新任务。阈值可根据系统运行情况动态调整。

3.3　存储中间件

数据快速安全传输的存储中间件，通过分析存储管理模块提供的设备信息和元数据管理模块提供的元数据信息，在广域网环境下实现远程数据的快速存取和安全可靠传输，以保证数据的及时性、可靠性、安全性和高服务质量。

目前存储中间件主要包括 3 个部分：数据传输模块、移动缓存模块和数据安全模块。数据传输模块在自定义的数据传输协议基础之上实现了文件管理和数据传输等基本的存储服务功能；在数据传输模块之上加载移动缓存模块，可以优化数据传输的性能；数据安全模块也可以加载在数据传输模块之上，从而对用户身份进行检验并防止重要数据在传输过程中被窃取和篡改。

(1)传输协议

用户和服务器之间的存储服务建立在文件系统之上，传输协议建立在 TCP/IP 协议之上，服务器与用户之间共有两种连接：控制连接和数据连接。控制连接是建立在客户端和服务器之间用于应答的通信链路。数据连接是传输数据的全

双工连接，完成传输后无需保持。数据连接只传输数据，控制连接传送命令和响应，从而实现了命令与数据的分流。数据传输协议示意图如图 4 所示。

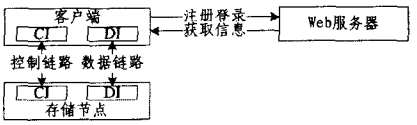

图 4　数据传输协议示意图

(2) 移动缓存

数据传输模块实现了基本的存储服务，而为了获得最佳的数据传输性能，需要加载移动缓存模块。

移动缓存设计的基本思想是：保证数据总是离使用者最近。虽然系统在用户注册时已经按照用户登录的 IP 地址为其指定了最近的存储子网作为服务区域，但当用户访问存储空间的网络位置发生变化时，用户在存取数据时便不能得到最佳的传输性能。移动缓存模块可以为用户动态地选择一个最优的存储节点作为代理缓存节点，用户写数据时，将数据写入代理缓存节点后，由代理缓存节点在后台将数据写入用户存储空间；用户读数据时，存储节点先将数据发送至代理缓存节点，再由代理缓存节点转发至用户，从而确保数据离使用者最近的原则。图 5 为加载了移动缓存模块的数据传输流程图。

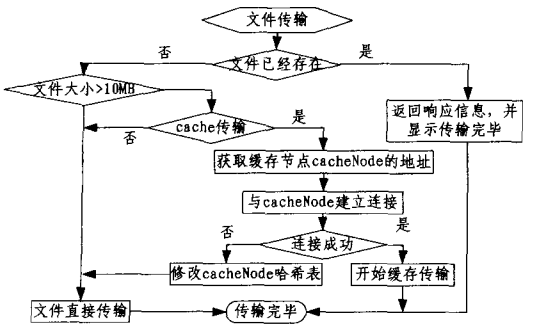

图 5　加载了移动缓存模块的数据传输流程

(3) 安全机制[6,7]

①身份认证

用户的管理控制和授权由三方身份认证机制来保证用户合法、合理地使用该系统，如图 6 所示。

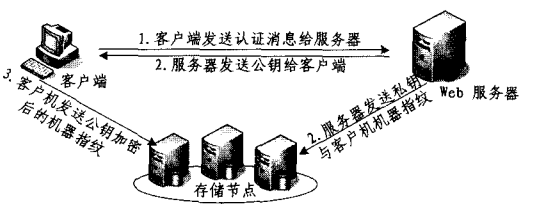

图 6　身份认证示意图

第一步是用户通过浏览器登录 Web 服务器，将用户名、密码和机器指纹等登录信息提交给 Web 服务器，其中机器指纹为用户注册本系统的详细时间。

第二步是 Web 服务器在数据库中提取用户的用户名和密码，检测是否为合法用户。如果为合法用户，则利用 RSA 算法产生一个公钥 G 和一个私钥 S(或者从事先已经生成的密钥文件里面取出来一对密钥)，然后将公钥 G 发送给用户作为登录成功确认信息；与此同时，Web 服务器将私钥 S 和该用户的机器指纹发给即将为该用户服务的存储节点。

第三步是用户收到登录成功确认信息(即公钥 G)后，利用该公钥 G 将自己的机器指纹进行加密(采用 RSA 算法)，并将密文传输给存储节点。存储节点收到该用户发送来的密文，用 Web 服务器第二步发来的私钥 S 解密该密文，然后与明文进行对比匹配，如果二者相同，则是匹配成功，那么就可以认为该用户为合法的，否则为非法的用户，那么立即中止与该用户之间的会话。

②安全传输

身份认证只能保证用户合法的进入本系统，用户的数据在传输过程中的安全则通过安全传输机制来保证。安全传输机制保证合法用户所上传的数据在传输时候不被他人恶意截获、随意篡改，保证数据安全保密地存储到本系统的某个存储节点上(该存储节点对用户透明)。

安全机制采用 SSL 协议，认证方式为证书认证，证书格式为 pkcs12，证书的内容包括主题名、颁发者、有效期、公钥、根证书(受信任证书)的签名等。通信双方持有自己的证书和根证书(受信任证书)，在 SSL 协议握手过程中互相验证证书，从而保证通信的安全。

4　性能测试与评价

为了获得广域网智能存储系统的性能参数，我们构造了一个系统原型，其中 Web Server 采用 HP Proliant DL320 G2 服务器，配置为：CPU：Intel(R) Pentium(R) 4 CPU 3.06 GHz，MEM：512MB。网络客户机采用多台 PC 机，配置为：CPU：Intel® Xeon™ CPU 3.00GHz，MEM：512MB。网络环境为 1000Mb 交换以太网。服务器安装 Linux AS 3.0，Web Server 使用 lampp(Linux + Apache + MySQL + PHP)套件搭建，客户机也在 Linux 下。对照系统为传统的 FTP 文件传输系统。MEM：512MB。网络环境为 1000Mb 交换以太网。服务器安装 Linux AS 3.0，Web Server 使用 lampp (Linux + Apache + MySQL + PHP)套件搭建，客户机也在 Linux 下。对照系统为传统的 FTP 文件传输系统。

从表 1 数据可以看出，广域网智能存储系统的数据传输速度明显高于传统的存储系统。这是由于移动缓存机制的加入，使系统传输性能得到较大的提高，虽然加入安全机制，增加了一些系统开销，但传输性能的优势仍非常显著，这一点在数据上载的测试实验中体现得尤为突出。最好的情况下，系统的上传下载传输速度分别提高 177%和 23%，值得一提的是，由于移动缓存机制是为了加快广域网环境下远程数据的存取速度而设计的，其整体设计思路及实现方式主要是针对广域网下客户机与服务器之间巨大的地理位置和网络环境的差异，但由于实验室测试环境下客户机与服务器之间的地理位置和网段的差异甚小，使得移动缓存机制在下载实验中未能充分发挥作用，而在真实环境中，客户机与服务器之间的地理位置和网络差异变大，移动缓存机制的效果会更加明显。

表 1　加载移动缓存和安全传输模块的改进系统与传统系统的性能比较

文件大小	操作类型	改进系统		传统系统	
		传输速度(KB/s)	传输时间(s)	传输速度(KB/s)	传输时间(s)
500MB	上载	10557	48	5300	96
1GB		13551	77	4900	210
2GB		10777	194	6100	340
500MB	下载	13580	37	11000	46
1GB		12145	86	12000	89
2GB		13430	156	15000	140

表 2 是广域网智能存储系统集合带宽随系统规模扩大而增长的情况。测试过程中将存储节点数由 1 个增加到 6 个，并模拟 100 个并发用户的访问，测试不同系统规模情况下集合带宽的变化情况。结果表明当存储节点的数目不断增加，系统的集合带宽拟线性增长，这一特点在图 7 中表现得尤为明显，这充分体现了广域网智能存储系统扩容与增速同步的思想。另一方面，广域网智能存储系统中实现了命令与数据的分流，避免了大量的数据在存储节点和元数据服务器之间往返拷贝，将高性能的存储设备和网络带宽充分提供给网络用户，而不被传统文件服务器的瓶颈所限制。

表 2　改进系统集合带宽随存储节点数增加而增长情况

并发用户数	存储节点数目	集合带宽(MB/s)	
		上载	下载
100	1	14.41	109.98
	2	33.76	220.61
	3	73.52	331.24
	4	86.74	420.39
	5	135.15	530.36
	6	153.06	624.67

从表 2 的数据中我们还可以看到，当多用户并发上载数据时，网络已经不是影响传输性能的主要因素，集合带宽受制于存储节点的 I/O 能力；当多用户并发下载数据时，系统的集合带宽比较稳定，基本能达到网络传输的峰值。因而，如果能进一步提高存储节点的写性能，将会使整个存储系统数据上载的能力得到很大提高。

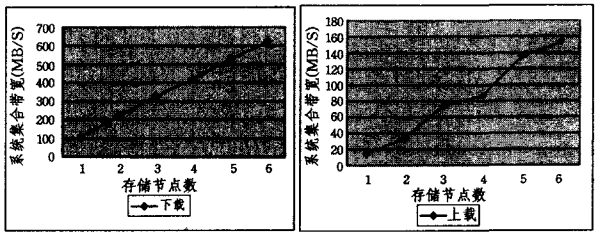

图 7　集合带宽随系统规模扩大拟线性增长示意图

结束语　广域网智能存储系统采用了多层次、可扩展的分布式存储模式，从改进系统结构着手整体地提高网络存储系统性能，提出基于异构多通道存储设备的、命令与数据分流的网络存储系统，该系统将存储节点通过网络接口直接接入下一代互联网，建立存储节点与用户间的直接数据传输，实现扩容与增速同步，保证服务器与存储节点、存储节点与存储节点间高效的控制和数据管理。其存储管理遵循 SMI-S 的管理规范，可以在广域网存储系统中提供标准化的通信方式。创新性地设计了广域网存储中间件，以处理复杂网络环境下的数据传输问题，兼顾了广域网数据传输的高速和安全。

参 考 文 献

[1] Seung Woo Son, Chen Guangyu, Kandemir M, et al. Energy Savings through Embedded Processing on Disk System[C]// Proceedings of the 2006 Asia and South Pacific Design Automation Conference. Yokohama, Japan, 2006: 128-133

[2] 刘群. 基于可扩展对象的海量存储系统研究[D]. 武汉：华中科技大学图书馆，2006

[3] Deng Ze, Feng Dan, Shi Zhan, et al. Scalability Support for SMI-S with Chord[C]// Proceeding of the 10th IEEE International Conference on High Performance Computing and Communications. Dalian, China, 2008: 219-225

[4] 王娟，冯丹，王芳，等. 一种元数据服务器集群的负载均衡算法[J]. 小型微型计算机系统，2009，30(4)：757-760

[5] Brandt S A, Miller E L, Long D D E, et al. Efficient Metadata Management in Large Distributed Storage Systems[C]// Proceedings of the 20th IEEE/11th NASA Goddard Conference on Mass Storage Systems and Technologies. San Diego, California, USA, 2003: 290-298

[6] Kher V, Seppanen E, Leach C, et al. SGFS: Secure, Efficient and Policy-based Global File Sharing[C]// Proceedings of the 23rd IEEE/14th NASA Goddard Conference on Mass Storage Systems and Technologies(MSST 2006). College Park, MD, 2006

[7] Kaminsky M, Savvides G, Mazieres D, et al. Decentralized user authentication in a global file system[C]// Proceedings of the 19th ACM Symposium on Operating Systems Principles. Bolton Landing, New York, 2003: 60-73

(上接第 227 页)

[3] 徐久成，安秋生，王国胤，等. 边界不确定信息的处理—Fuzzy 集和 Vague 集[J]. 计算机工程与应用，2002，16：24-26，82

[4] 闫德勤，迟忠先. 粗糙集与 Vague 集[J]. 计算机科学，2004，31(8)：133-135

[5] 刘金良，闫瑞霞，姚炳学. 粗糙 Vague 集的不确定性度量[J]. 系统工程与电子技术，2008，30(1)：104-107

[6] 朱六兵. 粗糙集与 Vague 集的理论及应用研究[D]. 成都：西南交通大学，2006

[7] Zhu W, Wang Feiyue. Reduction and axiomization of covering generalized rough sets[J]. Information Sciences, 2003, 152: 217-230

[8] 张文修，吴伟志，梁吉业，等. 粗糙集理论与方法[M]. 北京：科学出版社，2001

[9] 徐菲菲，苗夺谦，李道国，等. 基于覆盖的粗糙模糊集的粗糙熵[J]. 计算机科学，2006，33(10)：179-181

[10] 张倩倩，徐久成，胡玉文. 基于覆盖的粗糙 Vague 集模型研究[J]. 广西大学学报，2009，35(5)：653-657，662

[11] 李凡，徐章艳，饶勇. Vague 集[J]. 计算机科学，2000，27(9)：12-14，28

[12] 胡军，王国胤. 覆盖粒度空间的层次模型[J]. 南京大学学报，2008，44(5)：551-558

[13] Liang Jiye, Li Deyu. Information measure of roughness of knowledge and significance of attribute in rough set theory[J]. Journal of Engineering Mathematics, 2000, 17(supp): 106-108

[14] 刘纪芹，史开泉. 基于知识含量的粗糙集不确定性度量[J]. 计算机科学，2007，34(7)：171-173

[15] Zhang Q H, Wang G Y, Hu J, et al. Approximation partition spaces of covering space[C]// IEEE International Conference on Granular Computing. Silicon Valley, 2007: 199-204