计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
46(Suppl.): 198-202, 2009

基于对象存储系统体系结构的研究

单 颖[1] 姚念民[1] 赵建明[2]
[1](哈尔滨工程大学计算机科学与技术系 哈尔滨 150001)
[2](福建师范大学福清分校数学与计算机科学系 福建福清 350300)
(shanyingsc@sina.com)

Research on Architecture of Object-Based Storage System

Shan Ying[1], Yao Nianmin[1], and Zhao Jianming[2]
[1](*Department of Computer Science and Technology, Harbin Engineering University, Harbin* 150001)
[2](*Department of Mathematics and Computer Science, Fuqing Branch Campus of Fujian Normal University, Fuqing, Fujian* 350300)

Abstract With the fast development of object-based storage systems, the requirement for quality of service becomes higher and higher. Architecture SOBSS, which extends the function of metadata server in the architecture and simplifies internal data interactive model of the architecture, is proposed after analyzing the traditional architecture of object-based storage systems. The experimental results demonstrate that adopting the architecture of SOBSS has obvious advantages in promoting the performance of systems through comparing with the traditional architecture of object-based storage systems.

Key words object-based storage system; architecture; interactive model; metadata management

摘 要 随着基于对象存储系统的快速发展,存储系统服务性能要求越来越高.在分析传统基于对象存储系统体系结构的基础上,提出一种体系结构SOBSS,通过扩展元数据服务器功能,简化体系结构内部数据交互模式.通过与传统基于对象存储系统体系结构的性能比较,实验结果表明,采用SOBSS体系结构在提高系统性能方面有明显的优势.

关键词 基于对象存储系统;体系结构;交互模式;元数据管理

中图法分类号 TP393.02

随着当前工业的发展与学术研究的不断深入,存储技术得到快速发展,存储设备从非智能化和外部管理走向智能化、自我管理和关注其提供的存储应用的道路上来[1].然而,随着数据量的持续增长和存储设备成本的不断增加,存储系统影响着计算机存储系统的发展并成为计算机存储系统的瓶颈,成为提高计算机存储系统性能新的挑战.

基于对象存储系统(object-based storage system, OBSS)概念的提出克服了传统存储体系结构,如DAS(direct attached storage),NAS(network attached storage)与SAN(storage area storage)的种种缺陷,提供直接的、存储设备的文件存取[2],使可升级、高性能、跨平台与数据安全共享的体系结构成为可能[3].

基于对象存储的概念来源于CMU PDL的NASD项目,从而成为新的研究热点并不断引领这

收稿日期:2008-09-16
基金项目:国家自然科学基金项目(60503055)

一领域的发展方向. UCB、MIT、中国科学院计算技术研究所工程中心、清华大学高性能研究所等国内外研究机构在这一领域进行深入研究. IBM Haifa[3]实验室开发的基于对象的控制器原型、Cluster File System 公司的 Lustre[4] 文件系统、HP,Cisco 以及 Panasas[5-8] 公司的研究项目推动着存储技术产品化.

随着网络和计算机技术的不断发展,新的存储技术和存储体系结构不断出现,采用何种基于对象存储体系结构成为解决现有计算机体系结构中存在的存储器屏障和提高存储系统性能亟待解决的问题. 本文在研究当前基于对象存储系统体系结构现状和存在问题的基础上,探讨一种基于对象存储体系结构 SOBSS,该体系结构通过扩展元数据服务器功能,简化体系机构内部交互模式,提供一种高效的基于对象存储体系结构.

1 基于对象存储系统 OBSS

在基于对象存储系统中,对象是数据存取过程中基本的逻辑单元,是拥有类似文件接口的存储容器. 对象将文件(高度提取实现跨平台数据共享和基于策略的安全性)与块(直接存取和设备交换结构的可扩展性)相结合[3]. 对象由数据、用户可存取属性和设备管理元数据组成. 对象的长度可变并被用来存储诸如文件、数据库记录、医学图像或者多媒体等任何种类的数据[1]. 存储在对象中的数据对于基于对象的存储设备即 OSD 来说是不透明的,用户可存取属性用来描述对象的特点. 设备管理元数据是存储设备为管理对象物理存储而记录的信息. 基于对象的存储设备即 OSD 作为存储对象的设备提供存储服务,可以通过对设备上数据和需求信息的理解和优化实现存储设备智能化.

1.1 OBSS 体系结构

图 1 所示为 OBSS 的体系结构. 图中虚线箭头表示数据在网络上的传输流向. 典型的 OBSS 由 3 部分组成:客户端、元数据服务器 MDS(metadata server)和对象存储设备 OSD(object storage device). 3 个部分通过存储网络(如 TCP/IP 网络)连接进行协同工作. 其中,客户端作为应用服务器运行应用程序,提供与 MDS 交互的元数据信息或者与 OSD 交互数据的接口程序. MDS 连接客户端与 OSD,提供对象元数据映射信息及相关的授权访问信息. OSD 是对象存储的物理设备,提供客户端访问对象的接口与存储对象的信息.

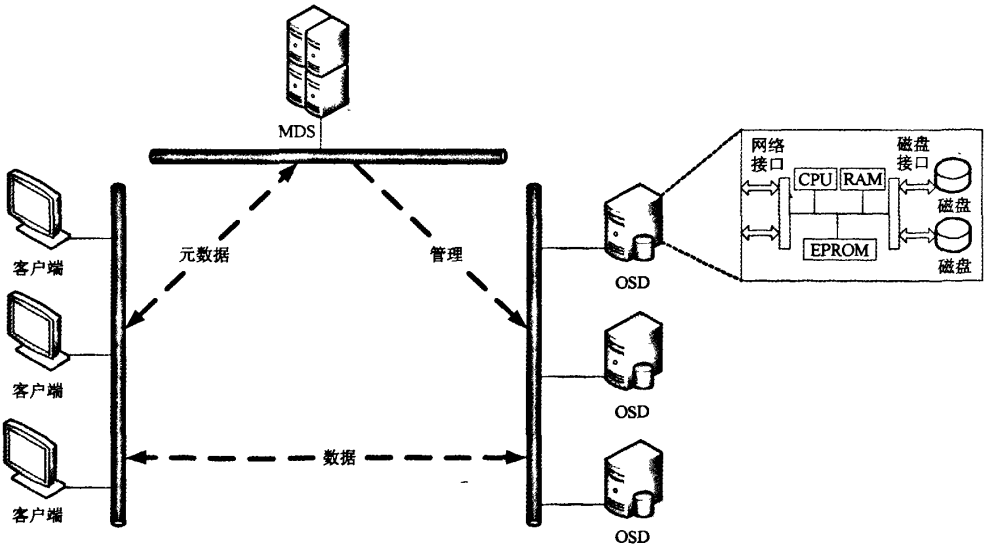

图 1 OBSS 的体系结构

1.2 OBSS 数据交互模式

图 2 所示为 OBSS 的数据交互模式. 该交互过程在 OBSS 体系结构的基础上描述客户端、MDS 与 OSD 之间实现对象存储的数据交互过程,以下是其交互过程的简要描述.

1) 客户端根据需求信息向 MDS 发送请求,请

求的内容为对象的唯一标识 OID.

2) MDS 中存放对象的元数据信息,元数据信息提供对象与文件之间的一种映射,MDS 通过 OID 找到对应信息后,将该信息与提供给用户的授权访问信息一起返回给客户端.

3) 获得元数据和授权访问信息后,客户端向 OSD 发送请求.

4) OSD 通过请求数据确定客户端是否拥有访问权限,验证通过则将数据返回客户端.

5) 整个过程中,MDS 对 OSD 的配置和操作进行管理.

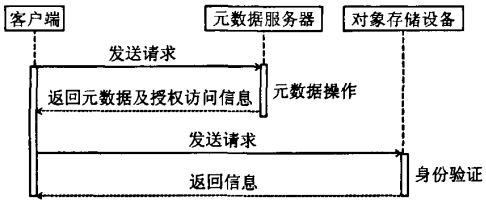

图 2 OBSS 的数据交互模式

1.3 OBSS 体系结构中存在的问题

在 OBSS 体系结构中,MDS 通过管理元数据的操作和文件到对象的映射提升整个系统的性能,实现客户端与 OSD 之间进行有效的并发传输. 然而,这种体系结构在某些方面增加系统的复杂程度.

1) 在时间方面

客户端向 MDS 发送请求后,需要等待 MDS 返回的授权访问信息及元数据信息,其中包括 3 个时间:客户端向 MDS 发送请求的时间、MDS 响应请求返回给客户端的时间及 MDS 查找元数据的时间;客户端向 OSD 再次发送请求数据,需要等待 OSD 将请求数据返回给客户端. 其中包括 3 个时间:客户端向 OSD 发送请求数据的时间、OSD 进行授权认证和查找数据的时间及 OSD 将数据传递给客户端的时间.

2) 在数据传输方面

客户端向 MDS 发送请求后需要等待 MDS 的返回的授权访问信息及元数据信息,其中同样包括两次数据传输;客户端向 OSD 再次发送请求数据,需要等待 OSD 将请求数据返回给客户端;MDS 与 OSD 管理中包括的数据传输.

3) 在存取安全方面

目前的绝大部分存储系统要求客户首先从元数据服务器取得数据的描述,然后客户再向具体的存储设备请求数据. 实际上,客户不需要知道元数据的具体细节,它只要求所需要的数据. 在这种设计中,客户对于元数据只是一个中转站,不但增加了延迟,而且由于暴露了元数据具体细节,增加了安全隐患.

如上所述,由于带宽与延迟这两方面的问题,使该体系结构出现存储器屏障,这种情况的发生将严重影响和限制 OBSS 系统性能,成为性能提升的主要障碍. 因此,当前的 OBSS 数据交互模式无论从时间、从数据传输还是从安全方面,都迫切需要进行改进,从而提高 OBSS 的性能与存储器工作效率.

2 SOBSS 的基本思想

在传统的 OBSS 体系机构中,MDS 的主要任务是对元数据的组织和管理,客户端通过 MDS 实现与 OSD 之间直接的存取操作. 基于效率的考虑,元数据与其描述的数据在其存放的物理位置上距离较近,将元数据信息与实际的数据分离的方式在一定程度上能够减轻系统工作负载,使 OBSS 的整体性能和效率得到提升. MDS 在基于对象的存储系统中的地位非常重要,往往成为 OBSS 的瓶颈. 如本文第 1.3 节所述,由于当前 OBSS 数据交互模式存在很高的复杂程度,因此我们在研究 OBSS 体系结构的基础上,从提升服务性能的角度提出采用 SOBSS 体系结构的基本思想. 该体系结构改进原有 OBSS 体系结构的工作模式,进一步将客户端向 OSD 请求数据的任务由 MDS 完成,客户端只需要一次请求即可等待 OSD 将其所需数据返回,其他工作均由 MDS 与 OSD 之间的操作来完成,这点对于客户端是透明的,从而简化了客户端的操作,发挥 MDS 在 OBSS 中的作用. 对于客户端于 OSD 来说,不需要增加任何复杂程度,在系统元数据的管理中,一方面增加了 MDS 与 OSD 之间传递信息的功能,另一方面对 MDS 结构和功能进行了扩展,从而提升 MDS 服务性能,有效解决 OBSS 中体系结构的整体性能问题.

2.1 SOBSS 体系结构

图 3 所示为 SOBSS 的体系结构. 图中虚线箭头表示数据在网络上的传输流向. 描述如下:当一个客户要求访问数据时,它首先向一个元数据服务器发出请求,元数据服务器根据该请求检索出所需数据的元数据,此时它已经知道哪些存储设备上存有该数据,则它直接将这些元数据发给这些存储设备,再由这些存储设备直接将数据发给客户.

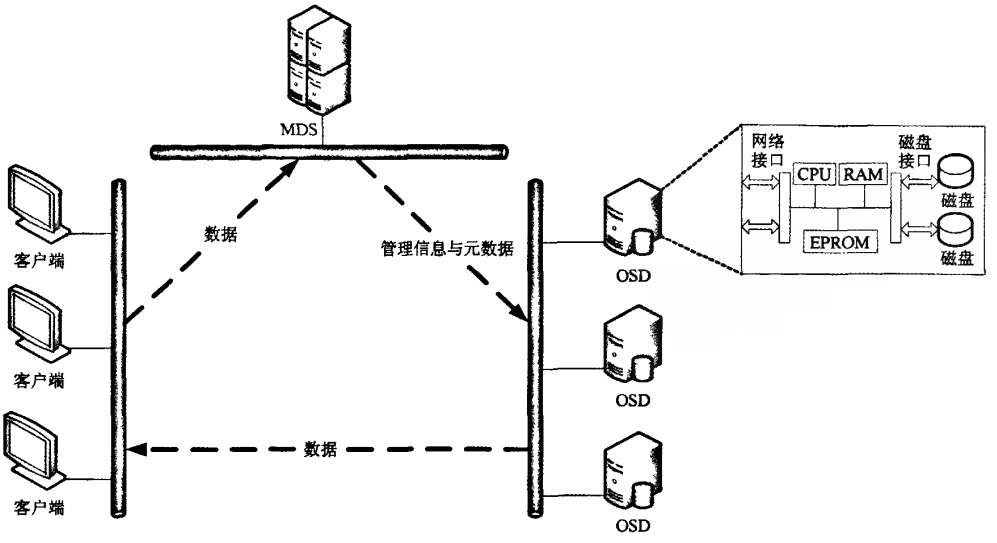

图 3　SOBSS 的体系结构

2.2　SOBSS 数据交互模式

图 4 所示为 SOBSS 的数据交互模式. 可以看出,SOBSS 对传统的数据交互模式进行如下优化:

1) 客户端只需要向 MDS 发送一次请求便可以等待接收数据;

2) MDS 不再是元数据的中转站,而且元数据不再流出存储系统的网域,增加了数据的安全性;

3) 由于客户端进行安全认证由 MDS 完成,OSD 信赖 MDS 发出的请求即不再需要对客户端进行认证;

4) 该体系结构在数据访问过程中至少减少两次网络传输,减少数据传输延迟与系统复杂程度.

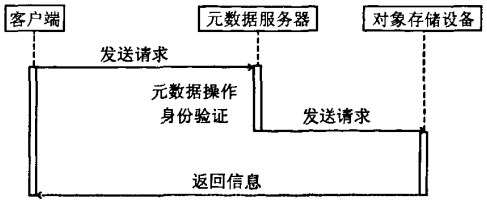

图 4　SOBSS 数据交互模式

2.3　SOBSS 元数据管理

与传统 OBSS 体系结构相比,SOBSS 对应的元数据服务器结构有较大变化. 图 5 所示为元数据服务器的结构图.

元数据服务器由 3 个组成部分:数据信息管理模块、存储设备管理模块及数据库模块. 文件管理模块与元数据管理的关系较大,主要实现接收客户端请求信息、认证、分配对象、记录客户端信息和元数据的操映射等功能. 设备管理模块与设备管理的关系较大,主要实现向存储设备请求数据信息、设备操作调度和设备信息存储等功能. 由于 LDAP 协议基于 X.500 标准,其数据库采用树型目录结构,具有读取的速度优于普通的关系数据库且存取的安全性高等特点,因此采用 LDAP 数据库存放存储设备信息、元数据信息以及客户端信息.

1) 数据信息管理模块:向客户端提供接口、管理与原数据及客户端有关的信息. 其中,客户请求模块负责接收客户端发送的对象请求. 文件管理中心模块对客户端请求进行分析后决定调用相应的功能模块,调用认证模块会对客户端进行安全认证,同时客户端信息管理模块记录客户端信息到数据库中作

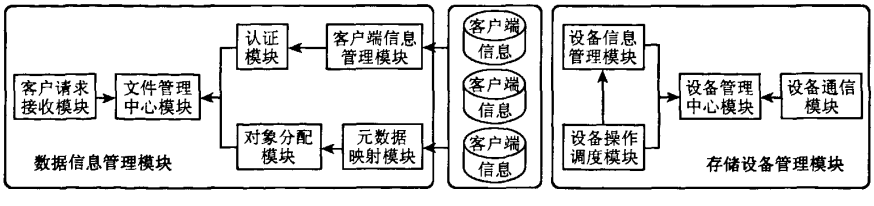

图 5　元数据服务器的结构

为传递给 OSD 的数据. 调用对象分配模块实现对元数据操作的管理. 元数据映射模块实现对象与文件之间的映射,可以直接操作数据库中信息.

2) 数据库:LDAP 数据库负责存放原数据信息与设备信息与客户端信息,这些信息均采用树型目录结构存放. 其中,存放客户端信息是 SOBSS 体系结构功能的关键,MDS 将客户端信息传送给 OSD, OSD 根据客户端的信息(如 IP 地址等)将数据传回客户端.

3) 存储设备管理模块:负责与 OSD 进行通信、管理 OSD 和相应的操作. 其中,设备通信模块负责将管理与元数据信息传递给 OSD. 设备管理中心模块负责对各功能模块进行调度,调用设备操作调度模块对 OSD 各操作进行调度. 设备信息管理模块负责管理与 OSD 操作相关的信息,可直接与数据库进行交互.

3 实验与性能分析

为测试 SOBSS 体系结构的性能,参照 Intel 开放存储工具实现 OSD[9],客户端访问速率为 200Kbps 与访问量相同,由百兆以太网互联的实验条件下,对采用典型 MDS 结构的 OBSS 与采用本文提出的 MDS 结构的 SOBSS 的实验结果进行比较. 图 6 显示了 SOBSS 与 OBSS 在不同的 MDS 结构下吞吐量的比较. 从图中可以看出,采用 SOBSS 进行实验的 MDS 吞吐量(曲线 MDS1)较大,而采用 OBSS 体系结构进行实验的 MDS 吞吐量(曲线 MDS2)与之相比则较小. 一方面,采用 SOBSS 中的 MDS 从与客户端交互数据到接收客户端请求减小负载,并将请求信息的工作承担下来,减小交互的复杂程度;另一方面,SOBSS 对 MDS 的结构和服务方式作相应的改进,能够提高 MDS 的服务效率. 从图中可以看出,SOBSS 体系结构在提高 MDS 吞吐量及提升整个体系结构的工作性能方面具有明显的优势.

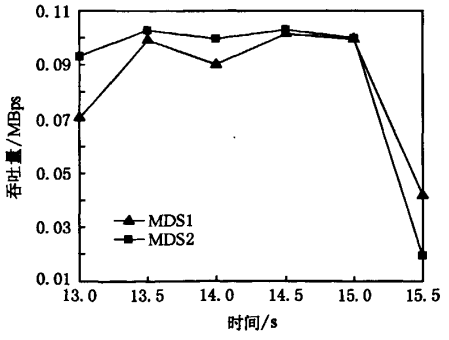

图 6 两种体系结构中 MDS 吞吐量比较

4 结论与未来工作

随着基于对象存储技术的快速发展,如何解决存储器屏障从而提高 OBSS 的服务性能成为基于对象存储技术的核心. 本文提出基于对象存储系统体系结构 SOBSS,通过扩展当前体系结构中元数据服务器功能,简化体系机构内部交互模式,提供一种高效的基于对象存储体系结构. 该体系结构的提出是利用体系结构提高 OBSS 性能的一次实践. 下一步的研究工作主要针对该体系结构的安全性、MDS 负载平衡策略、原型系统及文件访问方式等方面展开.

参 考 文 献

[1] Mesnier M, Ganger G R, Riedel E. Object-based storage. IEEE Communications Magazine, 2003, 41(8): 84-90

[2] Dan Feng, Hanbo Liu. I/O scheduling in huge object-based storage system //Proc of Japan-China Joint Workshop on Frontier of Computer Science and Technology (FCST'06). 2006: 34-45

[3] Mike Mesnier, Gregory R Ganger, Erik Riedel. Object-based storage pushing more functionality into storage. IEEE Communications Magazine, 2005, 24(2): 31-34

[4] Braarn P J. The Lustre storage architecture. [2004-08-04]. http://www.lustre.org/

[5] Gibson G, Nagle D, Amiri K, et al. File server scaling with network-attached secure disks //Proc of the ACM Int Conf on Measurement and Modelling of Computer System. 1996: 272-284

[6] Gibson G, Nagle D, Amiri K, et al. Filesystems for network-attached secure disks, CMU-CS-97-112. Carnegie Mellon University, 1997

[7] Factor M M, Meth K, Naor D, et al. Object storage: The future building block for storage systems //Proc of the 2nd Int IEEE Symp on Mass Storage Systems and Technologies. 2005: 119-123

[8] Panasas Inc. Object storage architecture. [2008-05-06]. http://www.panasas.com/

[9] Intel OSD/iscsireference implementation. [2008-05-06]. http://sourceforge.net/projects/intel-iscsi/

单 颖 女,1981 年生,博士研究生,主要研究方向为网络存储技术、高性能体系结构.

姚念民 男,1974 年生,教授,博士生导师,主持了多个国家科研项目,中国计算机学会信息存储技术专业委员会委员,主要研究方向为网络存储技术、高性能体系结构、服务器、网络存储,性能分析等(yaonianmin@hrbeu.edu.cn).

赵建明 男,1976 年生,博士,主要研究方向为网络存储技术.

基于对象存储系统体系结构的研究

作者： 单颖， 姚念民， 赵建明， Shan Ying， Yao Nianmin， Zhao Jianming
作者单位： 单颖,姚念民,Shan Ying,Yao Nianmin(哈尔滨工程大学计算机科学与技术系,哈尔滨,150001)， 赵建明,Zhao Jianming(福建师范大学福清分校数学与计算机科学系,福建福清,350300)
刊名： 计算机研究与发展 ISTIC EI PKU
英文刊名： JOURNAL OF COMPUTER RESEARCH AND DEVELOPMENT
年，卷(期)： 2009，46(z2)
被引用次数： 0次

参考文献(9条)

1. Mesnier M. Ganger G R. Riedel E Object-based storage 2003(8)

2. Dan Feng. Hanbo Liu I/O scheduling in huge object-based storage system 2006

3. Mike Mesnier. Gregory R Ganger. Erik Riedel Object-based storage pushing more functionality into storage 2005(2)

4. Braarn P J The Lustre storage architecture 2004

5. Gibson G. Nagle D. Amiri K File server scaling with network-attached secure disks 1996

6. Gibson G. Nagle D. Amiri K Filesystems for network-attached secure disks[CMU-CS-97-112] 1997

7. Factor M M. Meth K. Naor D Object storage, The future building block for storage systems 2005

8. Panasas Inc Object storage architecture 2008

9. Intel OSD/iscaireference implementation 2008

相似文献(4条)

1. 学位论文 陕振 OSD文件系统研究与设计 2005

基于对象存储体系结构是一种新的网络存储体系结构，具有高性能、高可扩展性、良好的安全性。基于对象存储设备(OSD)是基于对象存储系统中的基本存储设备，主要承担对象属性和数据的管理、设备安全以及同外界的网络通信功能，其中对象的属性、数据管理功能要通过一个设备内置的文件系统——OSD文件系统来实现。本文研究的目的是要结合OSD本身的特征设计一个适用于管理小型对象的OSD文件系统。

本文深入地分析了OSD本身的特征，并在此基础上研究了面向小型对象的OSD文件系统。在OSD文件系统的设计中，一方面借鉴了现有通用文件系统中的优秀设计思想，另一方面考虑使文件系统充分发挥OSD本身的优势。本文重点研究了对象ID到对象数据的映射机制、存储空间的分配与回收策略以及存储空间的动态优化策略，提出了一种基于“盘区记录表”的对象空间映射方式、一种基于大块分配和基于盘区分配相结合的空间分配策略和一种主动的对象空间动态调整策略。

最后依据本文的设计方案在Linus系统下实现了一个OSD文件系统的仿真系统，再将仿真的OSD文件系统置入一个基于对象存储系统中，替换了原系统中以Ext3文件系统实现的OSD文件系统，并使用标准的测试工具postmark对采用不同OSD文件系统的新、老存储系统的整体性能进行了对比测试。测试结果表明，本文设计的 OSD 文件系统在以小型对象为主的负载下能提高系统15%到60%的整体性能，并且文件系统性能可以随系统使用时间的延长保持稳定。

2. 学位论文 孙丽丽 共享对象存储并行文件系统的元数据管理研究 2005

当前的高性能计算已经由传统的主机方式逐渐向机群方式演变。机群体系结构的采用一方面使得系统的计算能力大大加强，另一方面也对当前的存储系统提出了更高的要求：在保证数据共享和易管理性的前提下，要求存储系统在存储容量和I／O性能方面具有很好的可扩展性。传统的基于主机的存储架构已经远远不能满足这些要求，研究新的存储体系结构和相应的文件系统具有十分积极的意义。本文基于对象存储系统，提出一种新的共享对象存储设备的并行文件系统(命名为SOPFS)设计，其目标是为高性能计算机群提供高性能、可扩展、高可用的机群存储系统。在文中给出了SOPFS的总体描述，内容包括分布元数据管理、并行数据访问等关键技术；结合SOPFS中动态散列分区的元数据组织方法，设计实现了元数据的访问管理；针对高性能并行计算中经常出现的对同一文件／目录的高并发访问情形，提出了一种动态的元数据复制策略，通过多个存放元数据副本的MDS同时响应对同一文件元数据的并发访问请求，提高了高并发访问情形下的元数据访问效率；利用SOPFS的结构优势，即文件系统的数据通路与控制通路分离，提出了懒惰的元数据更新策略，使得文件元数据的更新独立于文件数据读写过程，进一步保证了高的I／O吞吐率；在SOPFS中设计实现了元数据的事务日志机制，以提高系统的可用性和提供系统的快速失败恢复能力。

3. 学位论文 史伟 对象存储原型系统设计及相关实现 2006

数字以强大的信息表达能力以及单一的处理、传输和存储方式，融合了整个信息技术。半个世纪以来，作为数字信息载体的存储技术得到飞速发展。不断增长的存储需求和管理成本催生了基于对象的存储技术，而“对象”也有望成为下一代存储技术的标准接口。

在分析当前流行的网络存储体系结构及存储协议的基础上，对基于对象存储技术以及T10的SCSI OSD（Object-Based Storage Device Commands）标准作了深入研究。“对象”是传统块接口和文件接口的折中，基于对象存储系统在I/O性能、跨平台、可扩展性以及安全性等方面都表现不错。

实现了一个符合T10 SCSI OSD标准的对象存储原型系统，包括对象存储设备和客户端。客户端的SCSI对象设备驱动是一个Linux SCSI上层驱动，基于Linux块设备子系统实现，用来管理所有检测到的OSD设备。iSCSI启动设备是Linux SCSI协议栈的底层驱动，为客户端提供通过IP网络访问iSCSI目标设备的iSCSI通路。iSCSI目标设备实现iSCSI传输协议的Target部分。对象存储服务模块负责管理物理存储介质和处理OSD命令，以上模块均在Linux内核空间实现。

测试并分析了基于对象存储原型系统的性能，得出的结论是：通过在对象存储原型系统引入聚合读/写机制可以大大提升系统的I/O带宽。

4. 学位论文 娄成龙 对象文件系统中元数据管理技术研究 2005

随着信息技术的进一步发展，以及网络的大规模应用，带来了数据的爆炸性增长，也给网络存储带来了巨大的发展机会。如何构建一个扩展性强、

可靠性高、易管理的高性能存储系统成为目前研究的一个重要课题。

基于对象的存储技术是存储领域的新兴技术，它提出了一种新型的存储结构，数据对象是这种存储结构的核心，数据对象封装了用户数据(文件数据)和这些数据的属性(元数据)，他们分别由不同的系统管理。以对象存储结构为基础构建的大型分布式文件系统，可扩展性强、可靠性高，能提供较强的并发数据处理能力。元数据服务管理在对象存储文件系统中尤为重要，采用集群管理元数据是大型对象存储系统中的一种趋势，本文致力于研究对象存储结构中的元数据集群管理技术，所做的主要工作如下：1.分析研究基于对象存储系统的体系结构，设计并实现了一个小型的对象存储文件系统原型OCFS。

2.研究对象存储文件系统中的元数据管理，设计原型改进的文件系统OCFSⅡ，对元数据管理集群实行层次化管理，分别以目录路径索引服务器DPIS集群和元数据服务器MDS集群管理目录元数据和文件元数据。

3.在研究集群负载均衡的基础上，设计和实现OCFSⅡ元数据管理集群静态负载分配与动态反馈重分配相结合的负载均衡方案。通过静态元数据分割算法和元数据分布存储，实现元数据服务负载分流；采用动态反馈服务器负载信息，实现不均衡负载重新分配。保证元数据管理集群的负载均衡，并解决了“热点”数据访问问题。

4.设计实现了OCFSⅡ元数据管理集群可用性保障方案。目录路径索引服务器DPIS集群中采用共享容错磁盘阵列和节点容错机制解决共享存储数据和节点故障问题；元数据服务器MDS集群采用备份服务器保证服务器节点出现故障时元数据服务工作的接替和数据备份的重建。实现了元数据管理集群在单点失效和特定的多点失效情况下的容错和恢复，保证了系统的可靠性和可用性。