分布式网络存储系统元数据服务器群的设计与实现

林 凌，陈展虹
（福建教育学院信息技术系 福建 福州 350001）

【摘 要】：为了满足分布式网络存储对元数据服务的要求，本文采用多个元数据服务器组成元数据服务器群，并且提出分区散列管理方案对元数据服务器群进行有效的管理。并在不涉及元数据的物理移动的前提下，实现元数据服务器群的负载平衡机制、可扩充性以及高可用性等性能。

【关键词】：网络存储；元数据服务器

1.引言

在分布式网络存储系统中，元数据服务器模块地位非常的重要，因为所有的文件操作都要经过元数据服务器，它负责管理元数据，执行元数据操作。如何实现数据和元数据的分离，以及元数据的分布式管理是一个重要的任务。另一方面，我们发现目前大多数的研究工作是针对数据存储的性能优化的，而对元数据服务器的性能优化方面的工作却相对有限，因此，在存储数据激增的条件下，元数据服务器很可能成为存储系统的瓶颈。我对元数据服务器模块提出了新的机制和实现一分区的散列管理方案，它使用多台元数据服务器组成元数据服务器群，实现元数据服务器之间的负载平衡，并且还在可扩充性、容灾性等方面满足大规模存储的应用要求，提高了存储系统的整体性能。

2.相关研究

分布式文件系统的服务器结构发展到专有服务器系统占据主流的阶段，将元数据管理和文件数据存储分开，其典型代表有Slice文件系统[1]，Storage Tank[2]以及Lustre [3]文件系统等等。

但是，目前在元数据服务器的可扩充性能方面还存在局限。一些分布式文件系统，比如，"燕星"[4]系统，他只有一个元数据服务器，长期使用该系统，在数据量增加的情况下，元数据服务器将成为瓶颈。lustre分布式文件系统，虽然他是有两个元数据服务器，但是其中的一个元数据服务器只是一个备份服务器而已，并没有可扩充性。

3.元数据服务器群的设计与实现

3.1 总体设计和实现

1.总体设计

元数据服务器群模块要实现文件散列，元数据划分和元数据存储的功能，主要如下：

- 接受从应用服务器发来的文件系统请求，并返回响应
- 管理和处理元数据，包括查找、修改、移动等
- 负载均衡，当某个元数据服务器负载过重时，其它的元数据服务器服务器将分担它的部分工作
- 容灾处理，当某个元数据服务器故障，其他的元数据服务器将接替它的工作
- 扩充处理，当增加了新的元数据服务器，在新旧元数据服务器之间合理分配逻辑分区的管理权
- 重构处理，当添加了新的逻辑分区时，能够合理地在元数据服务器群间进行新逻辑分区的分配

为了完成上述功能，由两部分模块组成：基本功能子模块和性能优化子模块。主要功能如下：

① 文件散列管理器把一个文件名杂凑成一个整数，这整数可以被映射成一个存储着这个文件的元数据的那个分区。

② 映射管理器可以算出当前装入这个分区的元数据服务器的ID号。

③ 客户机可以向元数据服务器发出带有路径名散列值的元数据请求。

④ 逻辑分区管理器就存取位于通用存储空间的逻辑分区中的元数据，并向应用服务器返回元数据

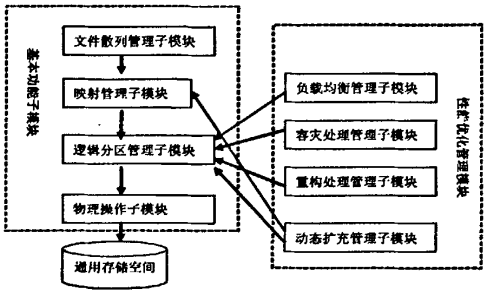

图1 元数据服务器群管理模块

2.软件结构

原型系统中的元数据服务器群的软件结构如下图2所示：

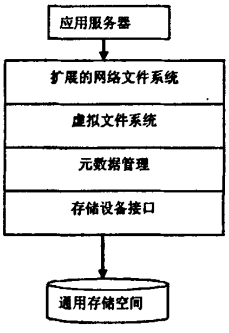

图2 原型系统中元数据服务器模块的软件结构

3.2 主要功能模块的设计和实现

一个元数据服务器可以安装多个分区，但是一个分区只能被安装到一个元数据服务器上。文件的元数据根据文件名的散列结果决定其在逻辑分区上的存储。主要算法如下：

```
Mapping(filename,PWi,MWi)
{
Pi=f(H(filename));        // H为文件名的散列函数；映射函数f将文件名的散列结果映射为分区号(Pi)，完成了第一个映射功能；
MDSi=ML (Pi,PWi,MWi);     //函数ML利用分区号和相关参数(PWi和MWi在后面说明)，计算出分区号Pi所对应的元数据服务器号MDSi，完成了第二个映射功能；
Return MSDi;
}
```

此外，映射函数要保证 Pi∈{0,Pn}，其中 Pn 是分区的总数；H(filename) ∈{0,Hn}，其中 Hn 是最大散列值，MDSi∈{0,Mn}，其中 Mn 是元数据服务器的总数。而且满足以下不等式：Hn≥Pn≥Mn>0。

从映射算法中我们可以看出，当 PW 和 MW 是固定时，映射管理器将映射函数 ML 简化成了一张映射表 MLT(如表 1)。通过 MLT，映射管理器可以了解到哪一个元数据服务器安装了那个分区，它管理着这个文件的元数据。最后，客户机和选中的元数据服务器连接以获得该文件的元数据，文件到目标的映射关系以及安全信息。

逻辑分区号（Pi）	元数据服务器 ID 号（MSDi）	元数据服务器的权重
0－15	0	300
16－31	1	300
32－47	2	300
48－63	3	300

表 1 MLT 示例

3.2.1 性能优化功能子模块的设计和实现

性能优化功能子模块的主要设计目的是提高元数据服务器群的性能，可用性和可扩充性。我们主要克服了传统的元数据服务器群的群内元数据存取和移动问题，降低交叉的元数据服务器元数据移动，以获取在负载平衡和可扩充性方面的高性能。性能优化功能子模块又由以下四个小模块组成：

1、负载平衡管理子模块

我们提出基于权数的动态负载平衡算法，算法如下：

根据每一个元数据服务器的 CPU 的处理能力、存储容量和带宽为 i 元数据服务器分配权数(MWi)。分区权数(PWi)反映 i 分区的存取频度，根据下列等式(1)决定哪一些分区分配给哪一个元数据服务器：

$$\frac{\sum PWi}{MWi} = \frac{\sum_{a=0}^{Pn} PWa}{\sum_{a=0}^{Mn} MWa}$$

此外，每个元数据服务器还需要负责维护自己的负载信息和所有安装在该元数据服务器上的分区的负载信息。

当一个元数据服务器出现了过载，就会向主节点发出警报。主节点会根据等式 1，将分区的控制权从过载的元数据服务器转移到一些低负载的元数据服务器上。

2、容灾处理管理子模块

传统的容灾处理是用备用服务器接管故障服务器。而本系统是依靠自己的元数据服务器"群"。当某个元数据服务器故障时，根据等式 1 来决定由群中的其他元数据服务器去接管故障的元数据服务器的逻辑分区。

3、动态扩充管理子模块

当负载增加时，系统可以动态地添加新的元数据服务器。将一些分区从现存的元数据服务器上卸下，分配给新加入的元数据服务器。不涉及任何元数据的物理移动。

4、重构管理子模块

本系统可以通过添加存储硬件来建立新的逻辑分区，以提高扩充的容量。

4.结论

本系统的元数据服务器群管理模块，可以管理分布式网络存储系统中的元数据服务器群。利用散列方法避免了大量的元数据存储，使用文件名散列机制可以消除过量的多元数据服务器间的通信问题。而且，在通用存储空间的逻辑分区的概念的基础上，元数据服务器群模块方法极大地简化了元数据服务器群的实现，也为负载平衡，容灾性和可扩充性提供了有效的解决方法。

参考文献：

1.D.Anderson,J.Chase,A.Vahdat.Interposed request routing for scalable network storage.Duke University,Tech Rep:CS-2000-05,2000.

2.J.Menon,D.A.Pease,R.Rees,et al.IBM storage tank-A heterogeneous scalable SAN file system.IBM Systems Journal,2003,42(2):250~267

3.P.J.Braam.The luster storage architecture.http://www.lustre.org/docs/luster.pdf,2003.08

4.郭朝阳，代亚非，韩华 燕星系统的设计及其实现中的技术问题.计算机工程与应用.2003.9,147.

(上接第 176 页)

在 Flash 元件制作过程中，如果一些图片不满足我们的要求，我们还可以通过 Flash 的打散命令打散图形，删除掉不满足我们要求的地方，然后再转换成元件的方法实现

3、从网络上收集矢量图

以上两种方法在处理图片上都有一定难度，我们还可以自己收集满足我们要求的图片，由于 Flash 是一款矢量图制作软件，所以，到网上下载现成的矢量图是最快捷的方法，下载完成后只要将其导入到库就可以随意调用了。

除了这些元件，还要将音乐导入到库备用，用文字工具输入文字，分段做成一个个元件备用。还要制作一些控制影片的按钮等等。

(3)课件的整合

课件的整合合可以用两种方法：就是放在一个场景或运用多个场景。如果已经制作了影片剪辑的小片段，就只需要一个场景就能解决问题，我们只需要按照课件的结构和次序将各个影片剪辑分配合理即可；另外一种情况是对大多数比较复杂的课件而言，对于内容比较复杂的课件来说，为了方便我们制作课件，我们可以采用多个场景整合课件的方法来实现。以上两种整合课件的方法，一般通过以下两种方法实现：

1、帧跳转

这是在 Flash 课件中最常用而且也是最简单的一种方式，在场景的时间轴上每隔一段(如 10 帧)插入一个关键帧，在关键帧上放置 Stop 命令，每个关键帧放置不同的交互内容，然后用按钮用 play 进行跳转，或者直接在按钮上加上命令 gotoAndStop (帧数) 来实现，这种方法适合在一个场景或一个影片剪辑里使用。

2、场景跳转

和帧跳转一样，对于内容比较复杂的课件而言，我们最好的方法就是分场景来实现，因为这样不仅可以把课件分成各个部分来分步实现，而且还降低了课件结构的复杂性。常见的场景跳转命令用 gotoAndPlay("场景",帧)来是实现，这种方法适合在多个场景之间进行切换。

(4)课件的发布

发布课件主要有导出.swf 和打包.exe 两种方式。swf 是 Flash 的打包影片格式，没有包含播放器程序，发布为 swf 有以下几种方法：使用"控制->测试影片"命令，使用"文件->导出影片"命令，单击"发布->发布设置"命令。在播放课件的计算机上没有 Flash Player 时，使用这种方式是最好的了。单击 Flash MX 的"文件->发布设置"命令，打开"发布设置"对话框，勾选"Windows 放映文件(.exe)"选项，单击"发布"按钮即可将课件打包成 EXE 文件。

参考文献：

1.黄冈 于林.(Macromedia Flash MX 标准教程)，北京希望电子出版社，2002 年版

分布式网络存储系统元数据服务器群的设计与实现

作者： 林凌， 陈展虹
作者单位： 福建教育学院信息技术系,福建,福州,350001
刊名： 福建电脑
英文刊名： FUJIAN COMPUTER
年，卷(期)： 2008，24(4)
被引用次数： 0次

参考文献(4条)

1. D.Anderson. J.Chase. A.Vahdat Interposed request routing for scalable network storage.[Tech Rep:CS-2000-05] 2000

2. J.Menon. D.A.Pease. R.Rees IBM storage tank-A heterogeneous scalable SAN file system 2003(02)

3. P.J.Braam The luster storage architecture 2003

4. 郭朝阳. 代亚非. 韩华燕 星系统的设计及其实现中的技术问题[期刊论文]-计算机工程与应用 2003(09)

相似文献(10条)

1. 学位论文 宋冬梅 基于网络的新型存储服务器系统中元数据服务器的设计与实现 2002

当前,数据密集型应用和互连网的快速发展对网络和数据存储能力提出了更高的要求.计算机系统结构在经历以CPU为中心和以内存技术为中心的阶段之后,逐渐转变为以I/O特别是存储系统为核心的阶段.而现存服务器系统主要依靠自身所带的存储设备,随着信息量和访问量的增长,造成传输、数据共享和管理共享等各方面的问题,使其不能满足日益增长的网络服务需要,因此研究新型可扩展、高性能、高可用的以存储为核心的服务器系统具有重要的实际应用价值.该文作者曾在国家高性能计算机工程技术中心参与以存储为核心的新型网络服务器系统的研发工作,承担了系统中元数据服务器的模块设计任务以及最小系统中元数据服务器的实现设计、编码、调试工作.该文首先介绍了存储技术发展的几个方面以及几种基于网络的存储系统上的分布式文件系统;其次,描述了以存储为核心的新型网络服务器系统的设计思路,剖析了该系统的软硬件体系结构,着重论述了新型网络存储系统中元数据服务器的设计和实现.

2. 学位论文 张顺达 对象存储系统的元数据管理 2006

随着网络技术和信息数字化的快速发展，面向海量数据的大型应用纷纷涌现，进一步对存储系统性能提出更为苛刻的要求。尽管磁存储技术仍在不断发展中，但受到块级存储访问接口制约，无法改变I/O性能远落后于CPU和内存速度的状况。对象存储系统（Object-Based Storage System）以对象为接口，将有望解决这些问题。容纳海量用户数据的对象存储系统中高效的元数据管理成为了新的挑战和研究课题。

对象存储系统由客户端、元数据服务器和各个对象存储节点三部分组成。用户数据存放在直接联网访问的智能存储节点上。元数据服务器在对象存储系统中的位置非常重要，是整个系统潜在的瓶颈。在这种具有分布式体系结构特征的对象存储系统中，文件被映射到一个或多个对象存储节点上。合理的对象分布策略对系统性能显得尤为重要。针对常用对象分布策略哈希（Hashing）算法和分片（Fragment-Mapping）算法存在的优缺点，提出一种能够结合两者优点、又尽量避免其缺点的柔性对象分布算法，同时分析了影响对象存储系统性能的主要因素。

元数据服务器的设计及元数据的组织和存储是面向对象系统中元数据管理的重要组成部分。元数据服务器使用了轻量级目录访问协议（Lightweight Directory Access Protocol，LDAP）作为存放元数据的平台，针对这个平台设计了相应的数据分配算法和数据转换模块，针对元数据访问特征，构建缓冲机制优化元数据访问性能。通过测试验证了柔性对象分布算法和元数据组织管理模式在对象系统中是行之有效的，并对系统性能的提升起到了重要作用。

3. 学位论文 陈文华 IP SAN非对称存储虚拟化的研究与实现 2004

本文以IP-SAN为基础，研究不同存储设备的虚拟化技术，提出IP-SAN的非对称存储虚拟化的架构和解决方案。

文中分析现在的主流存储技术的特点，对比不同存储技术虚拟化的发展趋势。比较不同的IP-SAN虚拟化方案，总结各种方案的特点，分析非对称虚拟化的优势。分析iSCSI协议，解剖iSCSI协议启动器和目标器结构和部分实现细节。提出IP-SAN非对称虚拟化架构，分析IP-SAN启动器端虚拟化的功能和接口；构建非对称元数据服务器模块，设计该服务器的软件组成和结构。研究IP-SAN非对称虚拟化的实现方案。提出非对称虚拟化管理中启动器翻译引擎的实现方案和元数据服务器的具体实现。

4. 学位论文 李小利 基于InfiniBand的网络存储系统结构与卷分配策略研究 2008

面对网络化环境下数据快速增长与存储容量的急剧膨胀，存储系统的I/O(输入/输出，Input/Output)性能成为衡量外存储系统的主要指标。由于InfiniBand(无限带宽，简称IB)具有高带宽、低延迟的优点，被广泛应用于高性能计算节点的互联网络，如果将该网络连接拓展到存储系统中，一方面可以显著提高存储系统的I/O性能，另一方面便于将计算节点间的网络和存储网络统一起来，给集群系统的管理提供了极大的方便。本文基于此目的，主要研究InfiniBand网络存储系统结构与存储卷的分配策略。研究内容如下：

网络存储系统结构方面，基于SRP(SCSI RDMA Protocol)协议，将InfiniBand体系映射到SCSI体系，概括出InfiniBand存储网络的I/O路径，利用通信队列对、内核旁路技术、远程直接存取技术设计了一个基于InfiniBand的存储系统目标模拟器。实验表明，远程主机通过InfiniBand网络连接访问目标器存储卷的I/O带宽比通过光纤通道连接或千兆以太网连接的I/O带宽高得多，甚至比主机本地磁盘访问的I/O带宽高5倍。研究还得出了目标器I/O带宽利用率受RAID卡带宽约束的结论，进一步提出了在目标器后端进行多个RAID卡组合，以提高后端磁盘访问的聚合带宽。

逻辑卷授权方面，通过主机通道全球唯一标识号(Globally Unique Identifier：GUID)与逻辑卷之间的动态映射关系实现了存储系统逻辑存储资源的授权访问与屏蔽策略。此算法作为一个授权功能模块添加在InfiniBand目标器中。系统测试表明：动态映射授权算法，有效确保了存储系统逻辑卷访问权限设置的安全性和灵活性。

SAN中的数据共享方面，给出了一个IB-SAN集中控制存储共享管理方案。它采用元数据服务器实现SAN存储管理功能，通过InfiniBand扩展网络、第三方数据传送和软件RAID来提高系统性能。提出了一种动态建立磁盘映射的策略以提供物理存储设备虚拟视图给系统中的节点服务器。测试表明，共享存储管理器保证了存储资源的同步性，支持用户对卷的在线增减，跨平台性好、接口简单、扩容简单。

5. 学位论文 陈俭喜 基于虚拟接口的网络存储系统研究 2006

计算机与互联网技术日新月异，人类社会信息化进程逐步深入，数据密集型应用不断涌现，存储需求也因此不断扩大，存储系统面临着严峻的挑战。存储系统不仅要具有较高的访问速度，同时还要能够合理地利用所有异构的存储资源，具有高可靠性和可扩展性。因此，利用高速网络通道组建网络存储系统并且在存储技术上寻求创新与突破是迎接这一挑战的最佳策略。

虚拟接口体系结构（Virtual Interface Architecture，VIA）是轻量级通信协议的工业标准，它具有通信开销小，传输带宽高等很多固有优点。特别是它缩短了传统的I/O 路径，较好地解决了以FC（Fibre Channel）为基础的存储网络互操作性问题，而且提供了远程DMA（Direct Memory Access）传输机制，可以实现直接存储访问。

因此，将VIA 应用于网络存储系统中，构建基于虚拟接口存储网络系统VISA（VirtualInterface Storage Architecture）可以有效地缓解网络存储系统中的传输瓶颈，解决存储网络的兼容性和互操作性，节省存储开支。VISA 系统中存储资源由存储节点提供，被元数据服务器管理调度，位于VISA 内部的用户可以直接使用这些存储资源，远程用户可以通过iSCSI 连接进行远程存储访问。

VISA 系统利用VIA 作为数据通信传输协议，而已有的VIA 实现要么因为使用专用的特殊网络设备，不仅兼容性不好，而且造价昂贵，要么因为抹煞了VIA 的固有优点而造成性能不高。为了克服这些缺陷，进一步提高VISA 系统的性能并保持其良好的兼容性和互操作性，设计了适合于存储的VI 网络适配器VI-NIC（VirtualInterface Network Interface Card）。它采用以太网作为传输媒介，通过SOC（System OnChip）的方式在适配器端实现VIA 协议功能，从而降低了主机的通信开销，提高了数据传输性能，并且保持了良好的兼容性和低价格。

异构存储资源的管理与存储资源利用率不高是存储系统需要解决的另一难题。

为此在VISA 系统中实现了两级存储虚拟化，第一级存储虚拟化是由存储节点将异构的存储设备虚拟为具有统一访问界面的存储实体，第二级存储虚拟化则是通过存储资源的动态分配与块地址的动态映射实现块级存储虚拟化。第一级存储虚拟化实现了对异构存储资源的统一管理与访问，块级存储虚拟化则既提高了存储资源的利用率，节约了存储成本，又可以通过优化的存储资源分配与块地址映射策略提高存储访问性能。

6.学位论文 钱迎进 基于对象存储的高可用技术的研究与实现 2005

在信息技术高速发展的今天，数据存储具有举足轻重的地位，在互联网的推动下，经济的全球化趋势将企业的经营模式从8×5变成了24×7，为了保持一个企业的正常运行，数据——企业最重要的资产——必须在任何时候都可供使用。如何构建一个高性能、高可用的存储系统，这是摆在研究者面前极富挑战的一个重要课题。

本文对基于对象存储文件系统Lustre的高可用性进行了研究，主要研究内容包括：Lustre文件级网络RAID的设计与实现及其不一致性的恢复、存储对象分配算法和负载均衡技术、元数据服务器集群及集群服务恢复技术等。

首先对Lustre的总体结构及其性能进行了分析，针对Lustre目前采用的高可用方案过于昂贵的问题，提出并设计了Lustre文件级网络RAID(LAID)的冗余存储方案。它不仅能提高了系统的聚合I/O带宽，而且可以同时容忍节点和磁盘失效，大大提高了系统的可靠性和容错性。然后针对存在的LAID不一致性的问题，提出了基于LAID日志记录的分布式协同一致性恢复算法，该算法对I/O性能影响很小，而且可以进行在线恢复。最后对LAID的性能进行了测试，测试结果表明它不仅可以得到很好的聚合I/O带宽，而且提高了系统的可靠性和容错性。

在研究负载均衡的基础上，针对各个存储节点的可用磁盘空间、可用网络带宽及负载的无时无刻都在动态变化，提出了存储对象分配和负载均衡算法，来更加均衡的使用各个存储节点存储空间和计算资源。通过存储对象分配策略：轮循法(round-robine)、剩余空间权重算法及写时创建(CROW)策略使所有的OST的存储空间资源都被均一的使用，从而避免发生单个OST存储空间被用尽，潜在IO负载不均的状况发生；同时提出了基于反馈信息的镜像LAID的负载均衡算法以及特殊情况下LAID5负载均衡算法，它以尽量减少服务器节点工作量均衡IO负载分布为目标，通过捎带机制来反馈信息，根据反馈信息在客户端计算出各个存储服务器的综合负载，选择负载最轻的对象存储服务器来进行文件IO，是一个基于反馈的全局动态负载均衡算法。

针对如何构建高可用、可扩展和高性能的元数据服务器集群，首先对Lustre采用的用的元数据分配算法、元数据服务器集群的元数据对象的管理及目录拆分技术进行研究；然后针对元数据服务器的恢复，给出了基于事务处理的服务请求的重放协议；针对分布式元数据操作可能造成整个文件元数据不一致的问题，设计一个分布式算法将各个元数据服务器节点的磁盘的状态保存到一个映像快照，出现故障后，通过undo日志把各个节点回滚到快照状态，从而保证整个文件系统的一致性和完整性；最后对基于共享存储模式下的集群高可用方案给出了failover服务器选择算法。

开源项目Lustre已经在各个存储领域中取得了巨大的成功。本文的部分研究成果已经融入到开源项目Lustre最新开发版中。

7.学位论文 吴思宁 机群文件系统服务器关键技术研究 2004

机群文件系统作为缓解机群系统I/O瓶颈问题的手段,需要为机群系统的各类应用提供高性能、可扩展的文件服务,因此对机群文件系统的研究是高性能计算机体系结构研究的重要内容.该文结合曙光机群文件系统DCFS的设计和实现,对机群文件系统设计的关键问题进行了讨论,并针对机群文件系统服务器设计的几个问题进行了研究.该文的主要工作如下:1.该文对机群文件系统的体系结构进行了总结,提出了多文件系统卷的结构,该结构具有可扩展、易管理、灵活的特点;该文对多文件系统卷中存储服务器的网络存储分组的组织形式进行分析,提出了网络存储分组模型,并讨论了影响存储分组读写性能的因素;对元数据服务器的组织和元数据的分布与映射策略进行了讨论,给出了可调粒度的元数据分布策略,使得用户可以根据应用程序的模式灵活选择文件系统卷的元数据分布粒度.2.作者对目录操作中的两个问题进行了研究:(1)元数据目录缓存管理;(2)大目录优化.独立的元数据服务器使设计者可以根据目录缓存的特点设计合理的管理方法,作者通过研究发现,客户端目录缓存和元数据服务器上的LOOKUP目录缓存和READDIR缓存构成了一个多级的目录缓存结构,元数据服务器上的LOOKUP缓存和READDIR缓存表现出了不同的访问特性,作者根据LOOKUP缓存和READDIR目录缓存的特性提出了目录缓存的管理方法,试验表明该方法较采用LRU、LFU和FBR替换算法的缓存管理方法具有更高的缓存命中率.作者和该研究小组成员合作对大目录优化进行了研究,提出了LMEH动态HASH的目录管理算法,在DCFS上的试验表明,对于大目录下的元数据吞吐率性能,该方法较线性的目录管理算法平均提高了1.97倍.3.作者结合DCFS元数据分布策略和元数据缓存管理设计了元数据一致性协议,该协议保证了元数据一致性,分析表明其开销是可以接收的.4.在曙光4000L上设计并实现曙光机群文件系统DCFS,给出了机群文件系统性能评价的方法,定义了读写带宽性能和元数据吞吐率的可扩展性度量.在曙光4000L上的测试表明,DCFS与类似结构的PVFS文件系统相比,在读写性能上,DCFS除了在小文件最高读带宽性能上比PVFS差19%,在其余情况下DCFS的最高聚合读写性能优于PVFS,平均高44.4%;DCFS元数据吞吐率的性能平均比PVFS高6.391倍;DCFS在综合负载测试中表现出比PVFS更好的性能,全局响应时间为PVFS的18.2%.

8.学位论文 金伟 备份系统在IP存储网络中的实现 2003

IP存储网络技术是目前存储体系的新的发展方向,它基于成熟的TCP/IP网络技术,通过IP网络传送数据.备份技术是维护系统安全性和可用性的最重要的手段之一.借助IP存储网络技术,构建基于它的网络备份系统,可以有效的满足应用和性能的需求.为了实现异构平台的数据共享,提供单一的管理点,获得所有存储资源的统一逻辑视图,及更好的扩展性和可用性,就必须采用SAN的存储虚拟化技术,目前网络级存储虚拟化是其主流的技术.采用非对称方式的元数据服务器来实现虚拟化,可以获得接近该地文件系统的性能.为了将广泛使用的NAS与SAN融合起来,构建一个统一的IP存储网络,可以采用"NAS Head"的基于NAS的统一存储网实现方式.

9.期刊论文 方圆.杜祝平.胡永奎.Fang Yuan.Du Zhuping.Hu Yongkui 基于对象存储文件系统研究综述 -计算机光盘软件与应用2010,""(5)

网络数据信息爆炸性的增长,使网络存储技术变得越来越重要,已成为Internet及其相关行业进一步发展的关键.近年来,存储技术的发展日新月异,已经由后台走到了前台,从而引发了数字技术的第三次浪潮.本文结合网络存储研究现状,将介绍一些热点网络存储技术及其新进展.

10.学位论文 付印金 PB级文件系统元数据管理关键技术的研究与实现 2008

随着高性能计算技术和因特网技术的不断发展，数据资源迅猛增长，很多应用的存储需求达到PB级。为了消除存储瓶颈，有效地支持高性能计算，继DAS、NAS和SAN三种网络存储技术之后，基于对象的存储技术成为存储领域的新兴技术，并形成了一种新型的存储结构。构建在基于对象存储结构上的PB级文件系统可以有效地管理数据资源，为用户提供一个虚拟化大容量存储器的统一访问接口、高I/O带宽、以及可扩展的存储服务。对文件数据的访问需要借助于元数据，元数据管理对数据管理至关重要。PB级文件系统具有TB级的元数据，为了消除元数据访问瓶颈，必须由元数据服务器集群来管理元数据，使得其元数据管理更具有挑战性。

本课题主要研究PB级文件系统的元数据管理。首先，通过对Lustre和PVFS的I/O性能测试，比较分析了基于对象的文件系统相比于传统并行文件系统的性能优势，并分析PB级文件系统结构和各组成部分的软件模块结构及其功能。其次，提出自适应的动态目录元数据划分方法来有效地平衡元数据服务器集群的负载，同时，最少化平衡负载过程中的元数据迁移量，并通过开发目录级局部性，提高元数据服务器cache的性能。再次，采用基于层次的计数型布隆过滤器数组能够提供快速的元数据查询服务，并能够节省内存开销。最后，根据元数据的特点，在结合文件访问语义和访问历史记录定义文件相关度的基础上，设计新的元数据预取策略来提高缓存命中率，降低平均响应时间。