CN43-1258/TP
ISSN 1007-130X
计算机工程与科学
COMPUTER ENGINEERING & SCIENCE
2005 年第 27 卷第 5 期
Vol. 27, No. 5, 2005

文章编号:1007-130X(2005)05-0054-02

LCFS 中元数据服务器的可靠性分析模型*

A Reliability Analysis Model of the Meta-Data Server in LCFS

王召福,章文嵩,刘 仲
WANG Zhao-fu, ZHANG Wen-song, LIU Zhong
(并行与分布处理国家重点实验室,湖南 长沙 410073)
(National Key Laboratory for Parallel and Distributed Processing, Changsha 410073, China)

摘 要:可靠性问题是研究大规模集群存储系统的一个重要方面,元数据服务器是大规模集群存储的中心。本文针对基于镜像复制和共享存储的服务器实现方案,采用马尔可夫补偿模型研究元数据服务器的状态迁移概率,分析了元数据服务器集群的可靠性以及数据一致性对可靠性的影响,对实现大规模集群文件系统中的元数据服务器有重要的指导意义。

Abstract: The reliability of large-scale clustering storage systems is an important research aspect in the related domain. Moreover, the meta-data server is the heart of a cluster storage system. In this paper, we propose the architecture of the LCFS meta-data server, and then use the Markov reward model to study the reliability of the meta-data server. The consistency can affect the reliability markedly, and the result is useful for the implementation and for improving the LCFS.

关键词:集群文件系统;元数据;可靠性;马尔可夫模型
Key words: clustering file system; meta-data; reliability; Markov model
中图分类号:TP333 **文献标识码**:A

1 引言

传统的文件系统中,元数据与数据本身通常由同一个文件系统管理,保存在同一台存储设备上,并且为提高访问效率,元数据与其描述的数据在物理上尽可能地靠近。而现有的一些大规模分布式存储系统中,数据可以通过高速网络直接访问,为避免元数据访问成为系统访问瓶颈,往往采用多台元数据服务器构成集群来提供系统的元数据服务[1]。在基于对象存储的大规模集群存储系统中,元数据服务器是整个系统正常运转的核心,元数据服务器的可靠性成为系统实现时需要着重考虑的问题之一。尽管元数据的数据量相对于整个存储系统的数据容量而言比较小,但有统计表明,在所有文件系统的访问中,对元数据的访问大约占全部访问次数的 50% ~80%。所以,元数据服务器的可靠性仍然是保证整个存储系统可靠性的重要因素。为了性能和容量的需要,在大规模集群存储中,元数据服务器也需要由多台服务器提供服务。本文将采用马尔可夫模型,对集群结构的元数据服务器进行可靠性分析。

2 LCFS 系统中的元数据管理

2.1 LCFS 的构成

我们设计的 LCFS(Linux Clustering File System,简称 LCFS)系统的构成主要包括三部分,分别为客户端文件系统、元数据控制系统和对象存储系统,如图 1 所示。客户端文件系统是 LCFS 文件系统的使用接口,元数据控制系统管理 LCFS 文件系统的树型目录信息和文件描述信息,而对象存储系统负责数据文件的存储。

2.2 元数据控制系统

元数据服务器 MDS 用于存储文件目录,包含集群文件系统的根目录。将分布在不同对象存储体上的文件对象构造成树型结构的、统一的存储空间。与传统的 Unix 文件系统不同的是,LCFS 的目录不再存储在文件对象中,而是独立地存储在元数据服务器中,以便于目录的集中管理和

* 收稿日期:2003-09-24;修订日期:2004-01-07
基金项目:国家 863 计划资助项目(2001AA111012)
作者简介:王召福(1975-),男,山东泰安人,博士,研究方向为分布式系统、性能评价等;章文嵩,博士,副教授,研究方向为 Linux 操作系统、集群系统等;刘仲,讲师,研究方向为集群存储系统等。
通讯地址:410073 湖南省长沙市砚瓦池正街 47 号并行与分布处理国家重点实验室;Tel:13560497512
Address: National Key Laboratory for Parallel and Distributed Processing, 47 Yanwachi St, Changsha, Hunan 410073, P. R. China

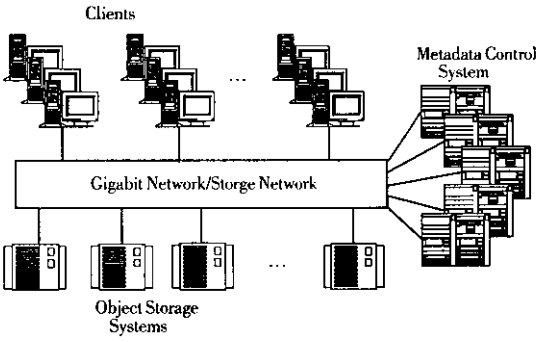

图1　基于对象的集群文件存储系统 LCFS 的结构

目录级的读写访问控制。

随着机群中结点数目的增加，系统失效的概率也将增大。为保证超级服务器提供连续不间断的服务，要求系统具有高可靠性。比较典型的两种元数据服务器集群实现方式包括：

（1）把元数据管理按照信息特征划分到不同的服务器进行管理（包括目录子树法或散列法等）。这时，针对每一个子服务器采用双系统备份，以提高系统的可靠性，双系统的后端存储系统一般采用镜像复制存储。

（2）采用前端动态调度、后端共享存储的模式。因为一般的后端存储设备（尤其 SAN 系统）可以支持很高的并行访问率，因此此方案具有可行性。

3　元数据服务器的可靠性分析模型

3.1　镜像复制技术及冗余存储

在当前的存储领域，镜像复制技术和冗余机制技术被广泛应用，以提高系统的可靠性。

镜像复制快速把数据块备份到与丢失数据的硬盘相关的冗余设备，可以认为是一种最快速的容错方法。它一般是在写数据时以并行方式写相同的两个驱动器。而通过编码技术利用冗余数据，获取系统的检错或纠错能力是存储系统中常用的容错技术。例如，RAID5 就是采用块交叉和校验信息旋转分布，实现可容错的磁盘阵列。

在上一节提到的元数据服务器的不同实现方案中，第一种方案可以认为是基于镜像复制，而第二种实现方案是基于冗余校验信息。在这一类系统中，这两种实现方案具有不同的可靠性。

3.2　马尔可夫激励模型（MRM）

马尔可夫模型和排队论被广泛用于系统的性能评价。对于马尔可夫链模型，首先需要正确建立描述系统性能需求的模型，然后可以通过矩阵式求解需要的性能参数，对系统状况进行评价。

对于系统的可靠性分析，一般采用马尔可夫激励（MRM）模型。MRM 模型把目标系统的状态 S 分为两类：

$S = A \cup N$

其中，A 表示可靠状态，N 表示不可靠状态。

若只考虑可靠状态集，可以通过式（1）、式（2）计算出在到达不可靠状态前的平均时间。

$$L_N(\infty)\,Q_N = -\pi_N(0) \tag{1}$$

$$MTTA = \sum_{i \in N} L_i(\infty) \tag{2}$$

3.3　基于数据复制的双服务器可靠性分析

针对系统的工作特点，可以定义系统的四个有意义的状态：

（1）状态 0 双服务器均处于正常工作状态；

（2）状态 1 一台服务器出现故障，双系统状态不一致；

（3）状态 2 一台服务器出现故障，双系统状态一致；

（4）状态 3 两台服务器均出现故障。

状态变迁定义的参数包括 α、γ 和 c，其意义分别在表 1 中定义。

表1　系统参数及意义

参数	意义	值（小时）
$1/\alpha$	平均故障时间	10^5
$1/\gamma$	故障修复时间	0.5
c	双系统一致性概率	

双系统可靠性状态转换图如图 2 所示。系统的可靠状态包括 0、2 二个，而状态 1、3 为不可靠状态。

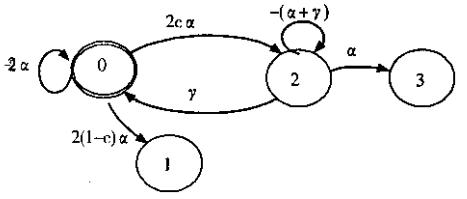

图2　双系统可靠性状态转移图

对有效状态建立状态变迁矩阵，有

$$Q_N = \begin{pmatrix} -2\alpha & -2c\alpha \\ \gamma & -(\alpha+\gamma) \end{pmatrix}$$

$$\pi_N(0) = (1 \quad 0)$$

由式（1）、式（2）求得：

$$MTTA = \frac{\alpha + 2\gamma}{2\alpha^2 + 2\alpha\gamma - 2c\alpha\gamma}$$

当 $c = 1$ 时，认为双系统完全保持一致。这时，系统的可靠性为：

$$MTTA = \frac{\alpha + 2\gamma}{2\alpha^2}$$

当 $c = 0$ 时，认为双系统不能保持一致，任何的系统失败都导致数据丢失。这时，系统的可靠性为：

$$MTTA = \frac{\alpha + 2\gamma}{2\alpha^2 + 2\alpha\gamma}$$

根据典型取值，在 c 分别取 0 和 1 的情况下，系统的可靠性有几个数量级的差别，这是因为 $\alpha\gamma$ 的差别较大。

因此，系统一致性对可靠性的影响比较明显，在依靠冗余复制保证系统可靠性的情况下，保证系统的一致性十分重要。

3.4　基于共享存储的元数据服务器可靠性分析

基于共享存储的元数据服务器的可靠性问题包括两部分，前端提供服务的服务器是 m 台互相备份的元数据处理服务器，而后端是共享的存储系统。元数据可靠性问题可简化地只考虑后端存储设备，此处以基于校验的 RAID5 存储设备为例进行分析。设 RAID5 存储设备中的单元数为 n，则其状态可归并为：

（1）状态 0 存储设备中 n 个单元均可用；

（2）状态 1 存储设备中有一个单元故障，系统可用；

（下转第 58 页）

```
    <author>吴持仁</author>
    <publisher>联合出版社</publisher>
  </book>
  <book year="2003">
    <title>夜玫瑰</title>
    <author>痞子蔡</author>
    <publisher> 媒体联合出版社</publisher>
  </book>
</bib>
```

操作后得到如下结果：

```
<title> XML 数据库的昨天和明天</title>
<title>夜玫瑰</title>
```

我们可以看到，这里仅仅涉及扫描。再来看一个例子：

```
/root/student[/name="Tom"]/age
```

这个例子中多了谓词操作，扫描部分同上例。但是，在 student 元素的路径中，需根据谓词进行过滤，只选择在子女元素 name 的子孙节点里存在的文本节点值为 tom 的 student 元素。最后，在检索到的 student 元素的子女元素里取得 age 元素的路径，该路径就是 NQuery 式的返回结果。

下面叙述使用正则表达式的查询。在现在的网络查询运用中，正则表达式运用越来越普遍，使用正则表达式将提高查询的速度，更可以完成复杂条件的查询。在 NQuery 中，引入了支持正规表达式比较运算符 =~。例如，在使用这个运算符对标题(title)以「夜玫瑰」开始的 book 元素进行查询，其表达式如下：

```
/bib/book[/title =~ "^夜玫瑰"]
```

^ 是表示文章的开头的字母。

查询结果如下：

```
<book year="2003">
  <title>夜玫瑰</title>
  <author>痞子蔡</author>
  <publisher>媒体联合出版社</publisher>
</book>
```

再对标题含有「昨天」字符的 book 元素进行查询，表达式如下：

```
/bib/book[/title =~ "昨天"]
```

查询结果如下所示：

```
<book year="2000">
  <title>XML 数据库的昨天和明天</title>
  <author>吴持仁</author>
  <publisher>联合出版社</publisher>
</book>
```

5 结束语

随着网络技术的迅猛发展和面向海量数字资源管理应用的大量出现，擅长事务处理和并发控制的传统关系数据库模型在处理半结构、非结构的数字资源时遇到了挑战。而 XML 数据库的出现正好可以弥补关系数据库的不足。本文提出了一种 NXD 的逻辑模型——NXD 树，并在此基础上设计了一种查询语言 NQuery，初步构想了 NXD 的实现。

参考文献：

[1] Ronald Bourret . XML and Databases[DB/OL]. http://www.rpbourret.com/xml/XMLAndDatabases.htm,2003-03.

[2] Akmal B Chaudhri, Awais Rashid, Roberto Zicari . XML Data Management: Native XML and XML-Enabled Database Systems [M]. Addison Wesley,2003.

(上接第 55 页)

(3) 状态 2：存储设备中两个单元同时故障，数据丢失。

状态变迁定义的参数包括 μ、λ 和 n，其意义分别在表 2 中定义。

表 2 系统参数及意义

参数	意义	值(小时)
$1/\mu$	平均故障时间	10^5
$1/\lambda$	平均修复时间	1 小时/2GB
n	系统中的设备数	

*RAID*5 可靠性状态转移图如图 3 所示。

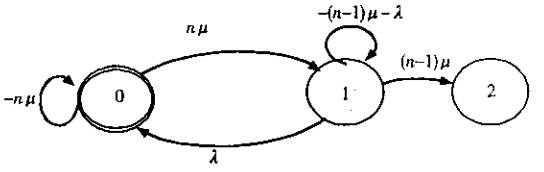

图 3 RAID5 可靠性状态转移图

对可靠状态建立状态变迁矩阵，有

$$Q_N = \begin{pmatrix} -n\mu & n\mu \\ \lambda & -(n-1)\mu-\lambda \end{pmatrix}$$

$$\pi_N = (1 \quad 0)$$

由式(1)、式(2)求得：

$$MTTA = \frac{(2n-1)\mu+\lambda}{n(n-1)\mu^2}$$

在系统参数 μ、λ 取值固定的情况下，系统的可靠性主要受阵列单元数 n 的影响。n 值越大，系统可靠性越低，但此时系统的利用率增高。当然，在存储磁盘阵列中，为保证系统的正确性，n 值不可能取太大。在 n=2 时，系统即为双系统镜像模式。

4 结束语

作为集群存储系统的中心，元数据服务器的可靠性研究是大规模存储系统研究领域的重要课题。对于不同的集群式元数据服务器实现方案，基于数据复制的双服务器模式对双系统的数据一致性程度要求比较高。而在基于共享存储的元数据服务器实现模式中，系统可靠性近似等价于后端存储设备的可靠性。不过，后端共享存储的访问性能必须通过存储局域网才能得到保证。

参考文献：

[1] Gunter Bolch, Stefan Greiner, Hermann de Meer, et al. Queueing Networks and Markov Chains[M]. John Wiley & Sons,1998.

[2] Qin Xin, Ethan L Miller. Reliability Mechanisms for Very Large Storage Systems[A]. 20th IEEE / 11th NASA Goddard Conf on Mass Storage Systems and Technologies[C]. 2003.

[3] Peter J Braam. The Lustre Storage Architecture[M]. Cluster File Systems,Inc,2002.

[4] Scott A Brandt, Ethan L Miller, Darrell D E, et al. Efficient Metadata Management in Large Distributed Storage Systems [A]. The 20th IEEE / 11th NASA Goddard Conf on Mass Storage Systems and Technologies[C]. 2003.

LCFS中元数据服务器的可靠性分析模型

作者： 王召福， 章文嵩， 刘仲， WANG Zhao-fu， ZHANG Wen-song， LIU Zhong
作者单位： 并行与分布处理国家重点实验室,湖南,长沙,410073
刊名： 计算机工程与科学 ISTIC PKU
英文刊名： COMPUTER ENGINEERING & SCIENCE
年，卷(期)： 2005，27(5)
被引用次数： 4次

参考文献(4条)

1. Gunter Bolch. Stefan Greiner. Hermann de Meer Queueing Networks and Markov Chains 1998

2. Qin Xin. Ethan L Miller Reliability Mechanisms for Very Large Storage Systems 2003

3. Peter J Braam The Lustre Storage Architecture 2002

4. Scott A Brandt. Ethan L Miller. Darrell D E Efficient Metadata Management in Large Distributed Storage Systems 2003

相似文献(10条)

1. 学位论文 于洪芬 小规模集群文件系统的元数据管理策略研究 2007

集群文件系统是集群的一个重要组成部分，它为用户提供一个虚拟化大容量存储器的统一访问接口和高I/O带宽。由于集群文件系统的文件数据分散存储在各个节点上，文件的定位需要借助元数据来完成，虽然元数据的数据量相对于整个存储系统的数据容量而言比较小，但有统计表明，在所有文件系统的访问中，对元数据的访问大约占全部访问次数的50%～80%，因而元数据的管理成为了管理数据的一个关键。
本文主要研究小规模集群文件系统的元数据管理策略，给出了一种具体实现元数据控制系统的方案，即两级元数据服务器结构，它由高级元数据服务器（Advanced Metadata Server，AMDS）和双元数据服务器（Double Metadata Server，DMDS）构成，分别管理目录元数据和文件属性元数据，实现数据的层次管理。在访问元数据时，分两级访问：第一级是访问AMDS中目录路径元数据，在文件系统名空间中定位元数据；第二级是访问DMDS，对文件自身元数据进行操作。在元数据控制系统中，AMDS是主体，所有的元数据操作请求都是通过AMDS处理的，只在需要时访问DMDS。这样可以保证元数据服务的高可靠和高扩展，同时能保证元数据访问能够一次定位，提高系统性能。具有容错能力的双元数据服务器采用双系统备份、镜像复制存储技术，从而保证了元数据管理的高可靠。

2. 期刊论文 封仲淹. 万继光. 李锡武. 李旭. FENG Zhongyan. WAN Jiguang. LI Xiwu. LI Xu 多级集群文件系统的全局命名空间的设计与实现 -计算机工程2006,32(21)

随着存储数据的爆炸性增长和集群技术的快速发展,集群文件系统的研究越来越成为一个焦点,该文设计并实现了一个多级集群文件系统(MCFS)的全局命名空间,MCFS系统采用元数据分层管理思想,建立一个统一的元数据树,从而使整个系统运行在一个松耦合、异构的环境下实现全局命名空间,提高了系统的并行性和可扩展性.在介绍了MCFS命名空间管理的同时,进行了相应的试验测试和性能分析.

3. 学位论文 秦航 基于集群文件系统的元数据容错研究 2004

为了解决PVFS中元数据管理的瓶颈，高可用性集群文件系统元数据容错系统MDFTS 以PVFS为基础平台，对系统中元数据的故障进行检测与诊断，并进行检查点恢复。为了达到复杂的元数据管理一致性，采用了一个无集中式服务器的体系结构，保证所有的数据和元数据能够存放到系统的任意地方，并且在操作的过程中可以动态迁移；采用元数据的磁盘日志结构和内存日志结构相结合的方式对元数据进行管理，减少了fsck对庞大的文件系统中元数据的扫描时间；为了实现故障恢复，提出了元数据容错的设置检查点算法和回卷恢复的算法，提高了文件系统元数据服务的可用性；给出了基于元数据故障的随机过程模型，可以通过减少检错时间提高文件系统的可用度。系统在操作系统应用层实现，通过修改元数据结构和相关的系统调用，使得集群文件系统内部各个数据节点和元数据管理节点相互协作，统一调度，支持高可用性。
测试结果表明，元数据容错系统可以针对系统模拟的不同类型的故障进行错误检测，并能够对系统和应用进行切换与恢复。

4. 期刊论文 田俊峰. 于洪芬. 宋玮玮. TIAN Jun-feng. YU Hong-fen. SONG Wei-wei 小规模集群文件系统中两级元数据服务器的设计与实现 -小型微型计算机系统2007,28(6)

在集群文件系统中,元数据服务器是整个系统正常运转的核心,它的可靠性和性能是设计系统时需要着重考虑的问题之一.本文设计了一个具有高可靠性、高性能的两级元数据服务器系统,兼顾了集中式元数据管理和分布式元数据管理的优点.系统中高级元数据服务器负责维护文件系统全局的目录结构和管理整个文件系统的命名空间,双元数据服务器负责维护文件元数据的分布信息,并采用了马尔可夫回报模型对两级元数据服务器系统进行了可靠性分析.实验数据表明,具有两级元数据服务器的集群文件系统能提供高吞吐量.

5. 学位论文 宋玮玮 两级元服务器集群文件系统的负载平衡策略 2007

集群文件系统是利用高速通用网络将一组高性能工作站或高档PC机，按某种结构连接起来，在并行程序设计以及可视化人机交互集成开发环境的支持下，对存储文件进行统一调度，协调处理，实现高效并行处理的系统。随着Internet的迅速发展，客户端对服务器的访问数量急剧增加，这就要求服务器应该具备高速并发处理任务的能力，为了提高资源利用率及缩短任务响应时间，采用好的负载平衡策略成为解决问题的关键所在。
本文给出了一种两级元数据服务器集群文件系统的负载平衡策略，高级元服务器依据二级元服务器的实时负载状态将任务快速分配给二级元服务器，实现了任务的并行处理。同时在存储文件时，给出了一种能正确反映各存储节点I/O流量和存储率的方法：计算文件热量值，并依据文件热量值对待存文件进行了合理的分配存储。文中对该策略进行了详细的说明，最后对其进行了性能分析，实验结果表明，本策略提高了系统的性能，缩短了任务执行时间，取得了较好的效果。

6. 学位论文 黄九鸣 SAN环境下高性能集群文件系统研究与实现 2006

本文在对当今主流的网络存储技术进行研究的基础上，针对SAN环境提出了一种集群文件系统模型SANFS，并对该模型进行了设计实现和测试验证。

SANFS具有高性能、高可扩展能力、低成本、易架设等特点，适用于非线性视频编辑、科学计算、VOD视频服务等以顺序I/O为主且对数据I/O稳定性要求较高的领域。

SANFS模型基于CIFS协议构建，可被实施于当今主流的各种操作系统。模型充分利用SAN网络上各主机共享存储设备的优势，让各客户端在元数据服务器的控制下，直接对存储设备进行I/O操作，使系统I/O吞吐率和I/O速率稳定性达到一个较高水平。模型以元数据“分散缓存集中控制”和文件

数据“分布读写”为主要指导思想，其核心技术包括：块设备对象缓冲、文件映射关系缓冲、元数据预分配及预取、SANFS OpLock机制及SANFS DOOR Lock等。其中，经本文抽象定义的SANFS DOOR Lock模型可被应用于其它场合的缓冲器中，具有较大参考价值。此外，SANFS还在客户端的文件系统一级，实现了对元数据服务器虚拟卷的支持，提高了系统可扩展性。

本文还详细设计了SANFS客户端在WINDOWS上的实现与元数据服务器在Linux上的实现，并研究与分析了其中的关键技术。

最后，通过对SANFS当前实现版本的测试，验证了模型的正确性和各种技术手段的有效性。

7. 期刊论文 李胜利. 陈谦. 程斌. 唐维. LI Sheng-li. CHEN Qian. CHENG Bin. TANG Wei 一种集群文件系统元数据管理技术 -计算机工程与科学2006, 28(11)

本文研究集群文件系统的特征,提出了一种分布式元数据管理技术.该技术通过哈希方式分布元数据对象、自侦测自适应和连续相邻节点备份的方法,实现了元数据的动态扩展和高可用.在我们研制的HANDY文件系统中采用了这项技术.测试结果说明,HANDY的元数据扩展性是令人满意的,实现了动态可扩展和高可用的设计目标.

8. 学位论文 丁亚军 蓝鲸集群文件系统Windows客户端的性能优化研究 2007

随着集群技术和网络技术的飞速发展,网络存储系统成为解决集群I/O性能瓶颈的主要手段之一.集群文件系统作为网络存储系统的核心技术,很好地解决了传统文件系统中存在的性能、容量、共享、可扩展性等问题.I/O性能是衡量集群文件系统的关键指标之一,面对应用的多样性和复杂性,如何提高系统的性能是当前集群文件系统研究的热点之一.

本文结合蓝鲸集群文件系统的特点,对Windows客户端的性能进行了深入研究,重点分析了单点顺序访问、单点并发及多点并发访问时系统中存在的性能瓶颈点,并有针对性地采用了一些优化方法优化系统的性能,使得系统的性能在实际应用环境中提高了30％～50％.取得了如下主要成果:1) 提出了客户端性能优化模型建立正确的性能优化模型是对客户端性能优化的基础.通过对Windows客户端数据流进行深入分析研究,找出了系统中存在的性能瓶颈点,在此基础之上,根据客户端软件各个功能模块的特点进一步进行抽象,建立了客户端性能优化模型,不仅为本文后续的研究打下铺垫,也为今后客户端在不同应用模式下进行性能持续优化打下坚实基础.2) 设计并实现了元数据缓存、ENBD异步模型以及专用的ENBD网络处理接口元数据缓存是集群文件系统客户端提高性能最有效的方法之一.通过客户端元数据缓存技术,一方面降低了元数据服务器的负载,另一方面元数据缓存中的块映射信息缓存有效地减少了数据读写关键路径上与元数据服务器通信的次数,降低了网络延迟,提高了整个系统的性能.ENBD协议处理层原先采用了同步处理模型,对所有的请求串行化处理,在优化过程中,将ENBD协议处理层进一步抽象成一个虚拟设备,采用设备的异步并发机制进行了优化,提高了单个客户端的并发性能.此外,针对ENBD协议的特点,设计了专用的ENBD网络处理接口,消除从ENBD协议到Socket协议间的拷贝开销.3) 提出并实现了基于文件类型动态调整预读粒度的机制预读作为提高文件系统性能的有效手段被广泛使用.通过深入研究蓝鲸集群文件系统在非线性编辑这种应用模型下存在的问题,提出并实现了基于文件类型动态调整预读粒度的机制,通过提供用户可订制的预读策略,很好地解决了系统在并发模式下磁盘抖动的问题,使得蓝鲸集群文件系统在非线性编辑应用模式下性能获得巨大提高.

9. 期刊论文 张敬亮. 张军伟. 张建刚. 许鲁. ZHANG Jing-liang. ZHANG Jun-wei. ZHANG Jian-gang. XU Lu 蓝鲸文件系统中元数据与数据隔离技术 -计算机工程2010, 36(2)

针对文件系统中元数据访问对应用数据访问的干扰问题,基于资源分配分离及访问路由分发技术,在蓝鲸文件系统中实现元数据和数据存储及I/O时的隔离.测试结果表明,将元数锯从数据中隔离并将其独立配置到高速存储后,应用I/O带宽提高了2.6倍～2.8倍,操作吞吐率提高了2倍～5倍.

10. 学位论文 丁定华 蓝鲸集群文件系统日志性能优化研究 2009

随着信息技术的不断发展和广泛应用，数据量呈爆炸式增长，网络存储系统越来越受到人们的关注。作为网络存储系统核心部件的集群文件系统，如何提供高吞吐率、高带宽、低延迟、高可用的文件共享服务成为亟待解决的问题。文件访问包括数据访问和元数据访问。虽然元数据量在整个文件系统中所占比例较小，但其访问频率较高，因此元数据访问的性能对文件系统整体性能有着重要影响，如何提高元数据访问的性能成为当前集群文件系统的研究热点之一。

本文结合蓝鲸集群文件系统的特点，深入研究日志模块对元数据访问性能的影响，重点分析了日志在文件系统同步导出语义下以及现有检查点机制的性能瓶颈，并针对性地采用了一些方法进行优化。测试结果表明，同步导出语义下元数据访问性能最多提高了2倍，此外，还有效地削弱了检查点操作对前端文件系统元数据访问的影响。本文取得如下成果：

1)提出一种日志并行提交机制

在同步导出语义下，文件系统事务提交非常频繁，单一的日志提交线程可能引发事务提交拥塞，使事务提交成为系统性能瓶颈。在分析并行提交相对串行提交具有优势的必要条件之后，本文设计并实现了一种日志并行提交机制，采用多个提交线程避免事务提交拥塞；此外，对并行提交下不同IO调度算法对日志设备IO性能的影响进行了研究，进一步提出并实现了多设备日志结构。

2)设计并实现了日志主动检查点机制和异步恢复机制

现有的日志检查点机制使用Pdflush系统线程作为元数据回刷线程，使用遍历尝试的策略释放内存和日志空间，精确性低、在同步导出时开销大、对于回刷时机的选择不能很好地满足日志的需求。本文一方面通过探测日志空间使用状态主动触发元数据回刷，另一方面通过建立回刷链表实现内存和日志空间的精准释放。大容量日志带来日志恢复耗时较长的问题，在高可用环境中将引起系统中断服务时间过长，本文提出并实现了日志异步恢复机制，Postmark测试表明，主从元数据服务器切换时文件系统上线时间缩短了17.5％～29.6％，模拟测试的缩短幅度达到了51.7％～83.2％。

关键词：集群文件系统，BWFS，日志，性能优化

引证文献(4条)

1. 李学明. 唐相桢 基于3-gram模型和数据挖掘技术的元数据预取[期刊论文]-重庆大学学报（自然科学版） 2008(6)

2. 田俊峰. 于洪芬. 宋玮玮 小规模集群文件系统中两级元数据服务器的设计与实现[期刊论文]-小型微型计算机系统 2007(6)

3. 田俊峰. 于洪芬. 宋玮玮 小规模集群文件系统中两级元数据服务器的设计与实现[期刊论文]-小型微型计算机系统 2007(6)

4. LIU Yuling. YU Hongfen. SONG Weiwei Design and Implementation of Two-Level Metadata Server in Small-Scale Cluster File System[期刊论文]-武汉大学学报（英文版） 2006(6)