DCFS 机群文件系统服务器组的设计与实现*

吴思宁　贺劲　熊劲　孟丹

（中科院计算技术研究所高性能室　北京　100080）

wsn@ict.ac.cn

摘　要　在客户机/元数据服务器/存储服务器结构的机群文件系统中，元数据服务器和存储服务器的设计和实现对整个机群文件系统的性能有重要的影响，本文给出曙光 LINUX 超级服务器机群文件系统 DCFS 元数据服务器和存储服务器的设计和实现，初步的测试结果表明 DCFS 机群文件系统服务器组的设计是可行的.

关键词　机群文件系统，元数据服务器，存储服务器

中图分类号　TP311.13

1　引　言

随着计算机技术的发展，机群系统已经成为高性能计算领域中的主流平台. 但现代大规模计算系统中存在的“I/O 瓶颈”问题极大地影响了系统整体性能的发挥，机群文件系统是目前机群系统解决 I/O 问题的一种重要方式.

许多研究机构对机群/分布式文件系统进行了研究，根据文件服务器结构、数据存储方式以及数据流动方式三个方面可以对其进行分类[1]. 首先，机群/分布式文件系统可以根据服务器构造方式分为专用服务器（Dedicated Server）与无集中式服务器（Serverless）两类. 前者包括 SUN 微系统公司的 NFS[2]、IBM 公司的 GPFS[6]；后者有加州大学伯克利分校的 xFS[4]以及国家智能计算机研究开发中心的 COSMOS[7]等. 其次，根据文件服务器上存储文件的块设备类型还可以分为两类：基于物理磁盘的机群分布式与基于逻辑虚拟磁盘文件系统，如 GPFS[6]等都是基于逻辑磁盘的；而 NFS[2]、COSMOS[7]都是基于物理磁盘的. 最后，数据流动方式可以划分为通过服务器中转与直接从存储设备到使用数据的客户端两种. 在基于 SAN 的 MPFS 系统中，数据直接从链接到光纤交换机的存储设备上传递到客户节点，而一般基于 Ethernet 的网络文件系统，如 NFS[2]都需要从服务器传递给客户节点.

DCFS（Dawning Cluster File Serving/System）是为曙光 3000L 机群及后续曙光超级服务器设计的机群文件系统. DCFS 将为曙光超级服务器上的关键应用，特别是科学计算、Web 应用与信息服务及视频点播服务提供可管理、可扩展及高性能的文件 I/O 服务.

本文将如下组织：第 2 节介绍 DCFS 文件系统的总体结构，第 3、4 节对 DCFS 服务器组(包括元数据服务器和存储服务器)中采用的技术进行了讨论，第 5 节给出了初步的测试结果，我们将在第 6 节中给出结论.

---

* 吴思宁，男，1975 年生，博士生；贺劲，男，1974 年生，博士；　熊劲，女，1969 年生，副研究员，博士生；孟丹，男，1965 年生，研究员，博士；研究方向为超级服务器体系结构、机群系统软件.

2 DCFS 机群文件系统结构

图 1 给出了 DCFS 机群文件系统结构以及使用 DCFS 的方式[1]. 从图中可以看出，DCFS 由客户节点、元数据服务器、存储服务器和配置管理节点构成. DCFS 提供了对多卷的支持，每个卷由多个客户节点、一个元数据服务器组和一个存储服务器组构成. 用户可以通过两种方式使用 DCFS 机群文件系统：（1）通过客户节点提供的文件系统接口直接使用 DCFS 文件系统；（2）客户节点把 DCFS 文件系统作为 NFS 的一个 export 目录提供给 NFS 的客户使用.

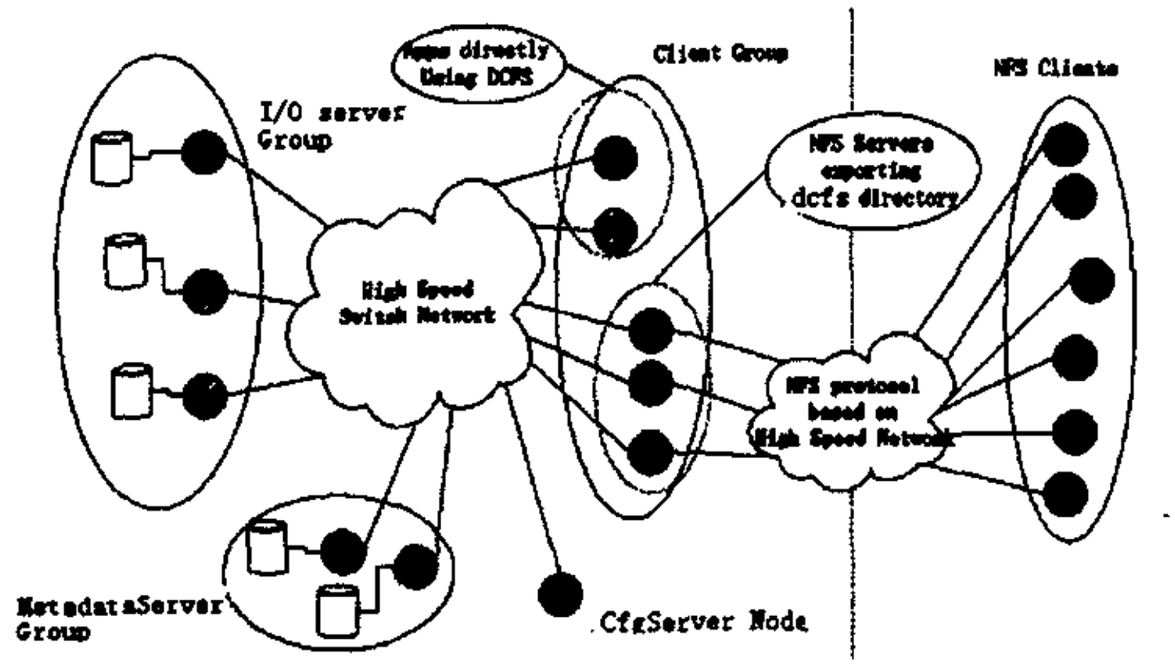

图 1 DCFS 结构示意图

DCFS 文件系统是一个单一映像的文件系统，提供给用户符合 POSIX 标准的文件系统接口，使得用户可以通过系统调用和 shell 命令直接使用 DCFS 文件系统.

元数据服务器负责管理 DCFS 的文件数据分布信息、DCFS 目录文件及普通 DCFS 文件元数据（包括文件长度、权限、日期及其它属性信息）的存储. DCFS 提供了字节粒度的强制文件锁的支持，文件锁的信息也是由 DCFS 元数据服务器负责维护. 每个元数据服务器组中有一个超级管理器的特殊文件服务器进程，它负责管理本卷的超级块以及其它重要信息. 当存在多个卷时，系统对应地会存在多个超级管理器. 在系统配置时，DCFS 的配置协议将协助系统管理员把这些超级管理器进程分散在不同节点上以便单个超级管理器节点崩溃时不会影响其它 DCFS 文件系统.

存储服务器组由连接双端口高性能磁盘 SCSI 阵列的多个节点构成. 为了有效地管理这些存储单元所管理的存储空间，同时也考虑大规模文件的带宽性能，DCFS 系统将它们划分为多个网络磁盘分组（Stripe）来管理，这些网络磁盘分组对于 DCFS 用户说来即为多个不同的 DCFS 文件系统卷. 在单个分组内，各存储服务器节点以类似于 RAID-0 的方式来组织文件数据的存储. 在文件读/写时，通过多存储服务器节点并发读/写而获得比单存储节点结构更好的聚集文件 I/O 性能.

超级服务器的系统管理员通过配置管理器来管理 DCFS. 这些配置与管理活动主要有启动与添加 DCFS 服务器与存储单元组以及扩充服务器磁盘容量等；安装或卸载 DCFS 客户端接口以及监控所有 DCFS 节点状态等.

3 DCFS 元数据服务器

在 DCFS 的元数据服务器组的组织中，我们采用了二级树状名字空间管理. 元数据服务器提供了文件属性缓存，同时为了加速 LOOKUP 操作和目录操作，我们采用了集中式目录缓存管理策略.

3.1 DCFS 名字空间的组织

由 DCFS 元数据服务器组组织成一个二级的树状结构，每个元数据服务器组有一个超级管理器，负责维护 DCFS 超级块、根目录的组织和名字空间的划分. 每个元数据管理器负责管理 DCFS 根目录下的若干子目录，整个 DDFS 根目录的下一级子目录及其所属的所有低级子目录的元数据都存放在相同元数据服务器中. 如图 2 所示，DCFS 根目录下的 A、B、C 子目录以及相应的子目录树由管理器 0 进行管理. 这种名字空间管理具有如下的优点：(1) 当系统发出长路径名字解析请求时，使用查找优化策略的 DDFS 客户进程只需要一次网络通信就可以取回所有元数据来完成整个查找操作[1]；(2) 该名字空间划分方法使得元数据服务器可以同时处理多个客户对不同子名字空间中文件操作请求；(3) 增加了系统的可用性，当某一个元数据管理器不能正常工作时，不会影响别的元数据服务器管理的子目录下的文件操作.

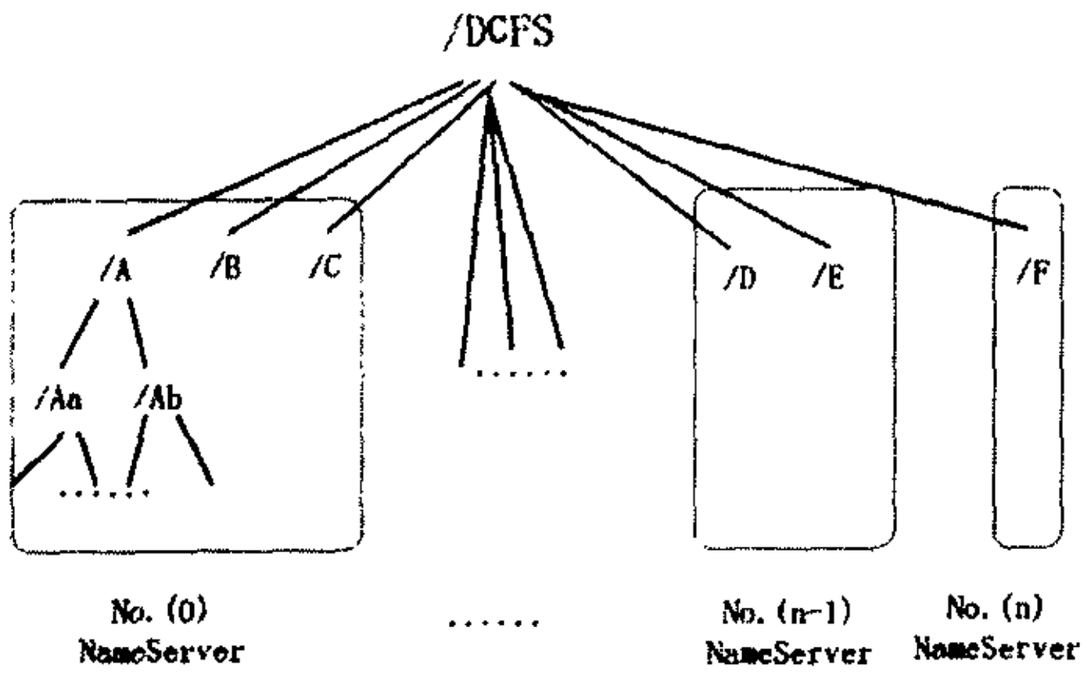

图 2 深度优先名字空间组织

3.2 目录缓存管理

DCFS 目录缓存有两种作用：(1) 加速名字查找（LOOKUP 操作）过程；(2) 加速读目录（READDIR）过程.

在实现中，DCFS 可以选择两种实现方式：(1) LOOKUP 目录缓存与 READDIR 缓存分离，此时 DCFS 客户节点内部目录缓存中对同一个 inode 的名字可能同时存放在两个不同类型的缓存单元中；(2) LOOKUP 目录缓存与 READDIR 缓存集中管理时，DCFS 需要更加复杂的数据结构来管理这些目录缓存.

在 DCFS 中，我们采用了集中目录缓存管理策略. 图 3 给出了集中式目录缓存策略的示

意图. 对于目录中的每一项，在元数据服务器中用目录入口描述结构来表示. 同时，我们用目录数据缓存来存放目录文件的数据块. 每个目录入口描述结构有一指针指向其在目录缓存块中对应目录项的位置. 为了能根据文件名快速查找到相应的文件的元数据信息，每个目录的 Inode 数据结构有一 HASH 表指针，该 HASH 表把该目录下的目录入口描述结构组织起来

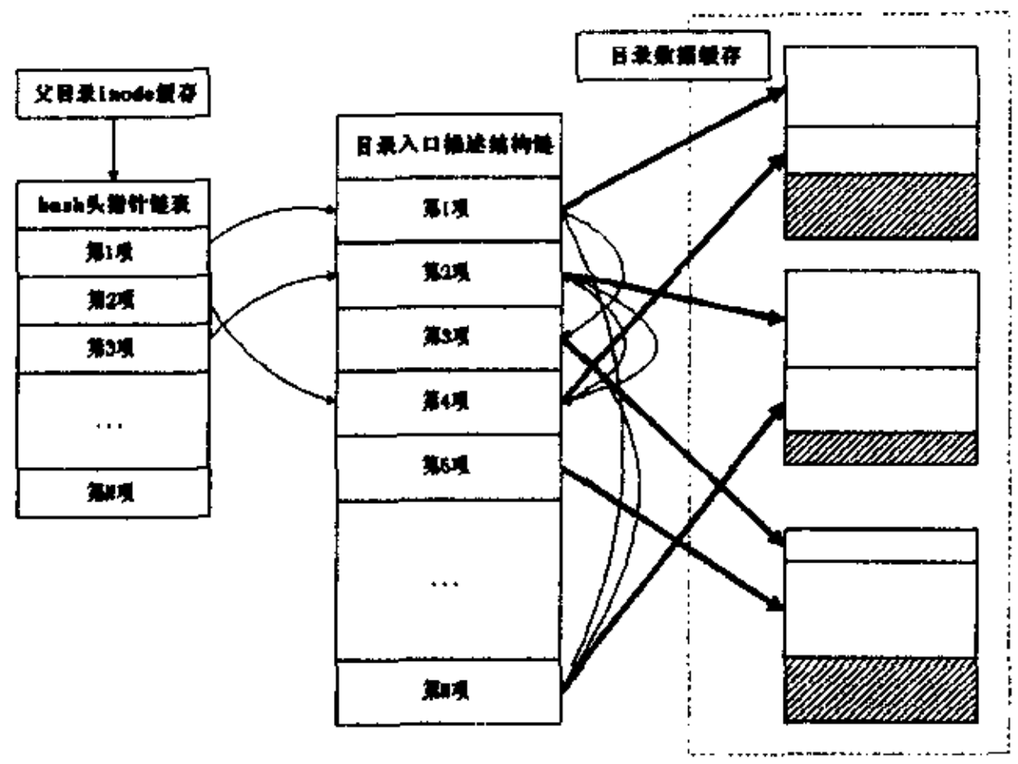

图 3 集中目录缓存组织结构示意图

4 DCFS 存储服务器

DCFS 存储服务器的设计目的是要充分地发挥其所管理的存储设备的性能. 在 DCFS 存储服务器的设计上，我们主要采用了多线程和缓存技术.

4.1 DCFS 存储服务器线程关系

如图 4 所示，DCFS 存储服务器由多个线程构成：（1）主控线程负责接收来自客户节点进程（Clerk）的文件读/写请求、文件同步请求和来自管理器进程的、代替 Clerk 进程转发的文件删除与截断等请求，并把请求放到请求队列上；(2)调度线程从请求队列上取下一个待处理的请求，选择空闲的执行线程并唤醒它去处理相应的请求；（3）执行线程负责处理调度线程分派的请求.

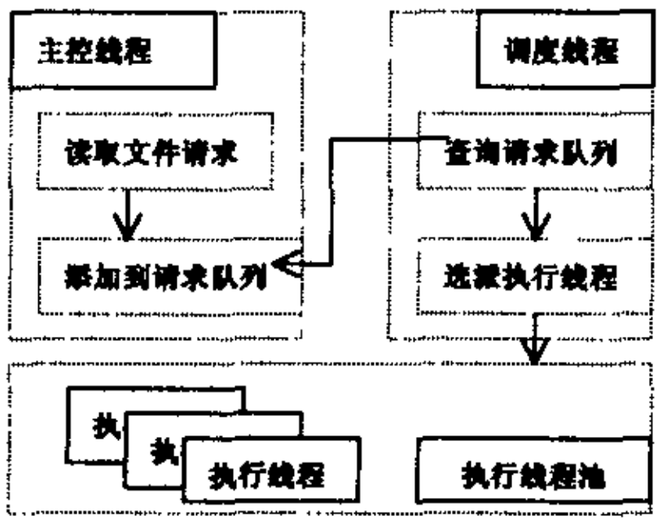

图 4 DCFS 存储服务器各线程关系

对于文件操作请求，存储服务器的基本处理模式如下：

（1）主控线程（masterThread） 负责接收新的 I/O 请求，并建立全局的请求队列.

（2）调度线程（schedulerThread） 读取全局请求队列，选派新执行线程（Executor Threads）执行这些 I/O 请求.

（3）执行线程（ExecutorThreads） 在执行时，按照工作性质执行线程可以分为几类：①通信线程；②磁盘 I/O 读/写线程；③普通文件操纵线程.

4.2 DCFS 存储服务器缓存设计

DCFS 文件系统提供了对多卷的支持，相应于每个文件系统卷，存储服务器提供了一套独立的数据结构进行管理，如图 5 所示，数据结构包括：（1）用于快速定位某个 DCFS 文件块在存储服务器缓存中的位置的 HASH 表. （2）DIRTY 链表. 本链表用来将存储服务器缓存中被执行线程改写，将还未同步到磁盘上的数据块组织在一起，以方便定期同步线程来同步这些数据块. （3）CLEAN 链表. 该链表保存“干净”的数据块，位于该链表中的数据块在没有空闲缓存块可供分配时被替换. （4）空闲链表. 该链表存放未被分配的缓存块.

图 5 DCFS 存储服务器缓存结构示意图

5 性能测试

实验平台是由 6 个节点构成的 Linux 机群，其中 2 个节点同时配置为存储服务器与元数据服务器，4 个客户节点的配置都相同：SMP 节点（双 PⅢ CPU，主频为 1GHz），主存 1GB，硬盘为 IBM 的 DDYS-T18350N SCSI 硬盘. 这 6 个节点通过 100Mbps 的快速以太网连接，节点操作系统为 Turbo Linux 6.5，核心版本为 2.2.18-10 SMP.

我们在上述的 Linux 机群的平台上，对 DCFS 的读/写带宽性能进行了测试，图 6 所示的为测试结果. 所用的带宽测试的测试程序是 HP 公司的 iozone，吞吐率测试的基准测试程序是 Lmbench.

图 7 给出了 DCFS 读/写带宽以及吞吐率性能测试的结果. 每个客户节点读/写的文件大

小为 128MB，读写粒度为 128KB.

图 6 DCFS 带宽性能测试结果

从图 6 可以看出，在一定规模内，客户节点地聚合读/写带宽很好地体现了 DCFS 的可扩展特性. 从测试数据看，DCFS 较好地利用了服务器的磁盘 I/O 与网络资源.

图 7 给出了 DCFS 的吞吐率性能测试结果，左图是创建、打开以及写入少量数据的聚合吞吐率；右图是关闭以及删除操作的聚合吞吐率. 当客户数增加时，它们较好地体现了可扩展特性，但是当文件增大时，在读/写操作时可能涉及到多个服务器，由于通信开销增长使得总聚合性能有所下降.

图 7 吞吐率测试结果

6 结 论

DCFS 是为曙光超级服务器设计的机群文件系统，其目的是为曙光超级服务器上的各种应用提供较好的 I/O 性能. 为了获得较好的 I/O 性能，我们在 DCFS 机群文件系统服务器组的设计中采用了一些技术. 在元数据服务器上，我们采用了二级的树状结构、文件属性缓存、集中式目录缓存管理等技术；在存储服务器上，我们主要采用了多线程和缓存技术. 初步的测试表明，我们在服务器组设计中采用的技术是可行的. 当前，我们在一个 6 个节点的小规模的 Linux 机群系统中实现了 DCFS，对于 DCFS 在由几十、几百个节点构成的大规模机群系统的性能，我们还需要在往后的试验中进一步研究.

参 考 文 献

1 贺劲. 机群文件系统性能与正确性研究: [博士学位论文]. 北京：中国科学院计算技术研究所，2002.

2 Brent Callaghan. NFS Illustrated, Addison-Wesley. April 2000.

3 Robinson D. The advancement of NFS benchmarking: SFS 2.0. In Proceedings of the 13th USENIX Systems Administration Conference 1999, 175~185

4 Thomas E Anderson, Michael D Dahlin, Jeanna M Neefe, David A Patterson, Drew S, Roseli and Randolph Y Wang. Serverless Network File Systems. ACM Transactions on Computer Systems, 1996, 14(1): 41~79

5 Kenneth W Preslan, Andrew P Barry, Jonathan E Brassow,Grant M Erickson, Erling Nygaard, Christopher J Sabol,Steven R Soltis, David C Teigland and Matthew T. O'Keefe. A 64-bit, Shared Disk File System for Linux. The Sixteenth IEEE Mass Storage Systems Symposium.

6 Jason Barkes, Marcelo R Barrios, Francis Cougard, Paul G Crumley, Didac Marin, Hari Reddy. Theeraphong Thitayanun. GPFS: A Parallel File System.SG24-5165-00. 1998.

7 冯军. 机群文件系统性能优化中的关键问题研究: [硕士学位论文], 北京：中国科学院计算技术研究所，2001.

8 Lily B Mummert. Exploiting Weak Connectivity in a Distributed File System，PhD thesis, Carnegie Mellon University.

9 John H Hartman, John K Ousterhout. The Zebra Striped Network File System.

Design and Implementation of Server Group in DCFS Cluster Filesystem

Wu Sining　He Jin　Xiong Jin　Meng Dan

(High Performance Computing Lab,Institute of Computing Technology, Beijing, 100080)

Abstract　In a cluster filesystem, the design and implementation of meta servers and storage servers have great effect on the performance of the cluster filesystem. In this paper, we introduce the design and implementation of DCFS cluster filesystem which is designed for DAWNING super server.

Key words　cluster filesystem, meta server, storage server

DCFS机群文件系统服务器组的设计与实现

作者： 吴思宁， 贺劲， 熊劲， 孟丹
作者单位： 中科院计算技术研究所高性能室(北京)

相似文献(10条)

1. 期刊论文 史小冬.祝明发.叶庆华 单一系统映象机群文件系统可扩展元数据服务器的设计与实现 -微电子学与计算机2002,19(2)

文章介绍了曙光3000超级服务器机群文件系统COSMOS中元数据服务器的设计与实现.针对机群文件系统对单一系统映象和可扩展性的要求,分析了机群文件系统元数据服务器基本操作的特点,找出了多个不同的元数据操作进行并行处理的可行办法,并给出了实现方案及系统性能的测试模型和测试结果.

2. 学位论文 熊劲 大规模机群文件系统的关键技术研究 2006

机群结构已成为高性能计算机的主流结构。随着CPU处理能力和通信速度的迅速提高，I／O成为制约机群应用实际性能的瓶颈。机群文件系统作为解决机群I／O瓶颈的核心技术，其研究具有重要的意义。

机群文件系统的发展趋势为：第一，元数据处理与文件I／O分离；第二，利用大规模网络存储系统来提供多条数据I／O通路；第三，利用一组元数据服务器来提供多条元数据I／O通路。

针对这种结构的机群文件系统，我们研究了其中的几个关键问题，包括元数据的分布问题，元数据的一致性和快速故障恢复问题，以及PB级机群文件系统的相关问题。本文的主要贡献在于：

(1)提出一种高效的大存储空间的管理策略——Bitmap-Extent混合策略。针对PB级机群文件系统，打破了传统文件系统基于一个块设备的限制，提出将机群文件系统与物理存储分离的一种逻辑空间策略，从而解决了文件系统容量受限问题和存储扩展问题等；而且针对PB级存储空间管理，提出一种基于位图与extent链表相结合的大规模存储空间管理机制，以提高存储空间的管理效率。

(2)提出一种基于粒度的动态元数据分布策略。元数据分布问题是决定非集中式元数据处理性能的关键问题。我们提出的基于粒度的动态元数据分布策略以提高元数据处理整体性能为目标，综合考虑元数据分布均衡度和文件系统层次结构关系两个因素对元数据处理整体性能的影响，按照一定粒度来划分名字空间和分布元数据，实验结果表明在模拟真实环境的负载下它的性能高于动态随机分布策略和动态根子树分布策略。

(3)提出一种基于简化的两阶段提交协议的、故障后可快速恢复元数据一致性的分布式元数据处理协议。元数据一致性问题是影响分布式元数据管理的可靠性和高可用性的关键问题。为了解决元数据服务器之间的元数据一致性问题，我们将两阶段提交协议与元数据的处理协议结合起来，提出一种基于简化的两阶段提交协议的分布式元数据处理协议，在元数据服务器失效或客户节点失效时，能够快速恢复文件系统的元数据一致性，保证文件系统的可用性。

(4)设计和实现了面向多用户多任务环境的、支持大规模机群系统的、面向海量数据存储的机群文件系统。DCFS2。在机群文件系统性能评价方面，提出从峰值性能、稳定性能、系统规模扩展性、元数据服务器扩展性、存储设备扩展性和存储I／O带宽利用率六个性能评价指标。并用这六个指标对DCFS2的性能进行全面评价。我们的结果表明，DCFS2能够获得比GFS等文件系统更高的聚合I／O带宽和聚合元数据处理性能。

3. 期刊论文 宋玮玮.马晓雪.Yu Hongfen.Tian Junfeng.Song Weiwei.Yu Hongfen 一种包含两级元数据服务器的机群文件系统 -科技信息（科学·教研）2007,""(12)

本文提出了一种两级元数据服务器机群文件系统,高级元服务器实现任务分配功能,依据与二级元服务器性能最密切相关的负载信息将任务快速分配给二级元服务器,同时在存储文件时,提出了一种文件热量模型,能正确反映各存储节点I/O流量和存储量的方法,并依据文件热量值对待存文件进行了合理的分配存储.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.

4. 期刊论文 吴思宁.贺劲.熊劲.孟丹 DCFS机群文件系统服务器组关键技术研究 -微电子学与计算机2003,20(6)

在客户机/元数据服务器/存储服务器结构的机群文件系统中,元数据服务器和存储服务器的设计和实现对整个机群文件系统的性能有重要的影响.文章给出了曙光LINUX超级服务器机群文件系统DCFS元数据服务器和存储服务器的设计和实现的方法,测试结果表明DCFS机群文件系统服务器组的设计是可行的.

5. 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介组的组织形式进行分析,提出了网络存储介组模型,并讨论了影响存储介组读写性能的因素;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％.

6. 学位论文 李晖 基于日志的机群文件系统高可用关键技术研究 2005

近年来机群系统以其低成本、高性能而逐渐成为高性能计算的主流平台、作为解决机群系统外存储瓶颈上的有效手段的机群文件系统因此得到了很大的发展。一个机群文件系统必须要满足机群计算环境的需要，为应用提供高性能、可扩展、高可用的文件服务。由于机群文件系统本身结构复杂，实现复杂而且整个系统规模很大，这些因素就决定了对高可用技术的依赖。本文将研究基于日志的机群文件系统高可用的关键问题以及解决策略，同时给出了一些评价方法以及具体的评测结果。具体内容以及研究成果如下：

(1)研究了基于日志的机群文件系统高可用技术的关键问题。本文分析了不同类型的机群文件系统的高可用需求以及高可用技术，对机群文件系统高可用相关的概念进行了介绍，描述了机群文件系统高可用领域的研究内容，并在分析的基础上提出了基于日志的机群文件系统高可用技术，分析了其中的关键问题，给出了相应的解决策略，并对其正确性和完备性给予了证明。

(2)实现了DCFS2机群文件系统高可用模块。作为文中策略的一个实际应用，本文给出了DCFS2机群文件系统高可用的设计与实现技术，给出了系统中利用日志来保证机群文件系统一致性的方法。主要内容包括：以DCFS2机群文件系统为原型系统，研究了单一以及多个元数据服务器下如何使用日志来保证文件系统的一致性：研究了机群文件系统日志对元数据操作的性能影响；研究了客户端的高可用问题。

(3)提出了机群文件系统高可用性的分级的定义。机群文件系统的高可用性的高低一直缺乏有效的定性或定量的分析方法，由于软件系统不能象硬件系统那样进行定量分析，我们根据机群文件系统的应用模式，将影响机群文件系统高可用性的因素进行分析，以机群文件系统的故障因素和恢复目标因素为线索，采用分级的方法对机群文件系统高可用性进行了定义，提出了机群文件系统高可用性的分级的定义。

(4)对基于日志的高可用技术进行了评价。目前在高可用技术的评价上尚没有完善的评价体系，本文从功能性，正确性，性能，恢复时间等多个方面对基于日志的高可用技术进行了评价，并给出了各种情况下的具体的测试结果。 文中还讨论了下一步的研究方向，包括多节点故障恢复等方面。

7. 期刊论文 田俊峰.宋玮玮.于洪芬.TIAN Jun-feng.SONG Wei-wei.YU Hong-fen 两级元服务器机群文件系统的负载平衡策略 -计算机工程2007,33(16)

提出了一种两级元数据服务器机群文件系统的负载平衡策略,高级元服务器依据与二级元服务器性能最密切相关的负载信息将任务快速分配,在存储文件时,提出了一种能正确反映各存储节点I/O流量和存储量的方法:计算文件热量值,并据此对待存文件进行了合理的分配存储.实验结果表明,该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.

8. 期刊论文 熊劲.范志华.马捷.唐荣锋.李晖.孟丹.XIONG Jin.Fan Zhihua.Ma Jie.Tang Rongfeng.Li Hui.MENG Dan DCFS2的元数据一致性策略 -计算机研究与发展2005,42(6)

随着集群应用对机群文件系统的性能、容量和规模等需求的日益增长,采用多元数据服务器是机群文件系统发展的必然趋势.基于多元数据服务器的分布式元数据处理是文件系统研究的一个重要问题.机群文件系统DCFS2采用分布式日志技术和改进的两阶段提交协议解决了分布式元数据处理下元数据的一致性问题.性能测试结果表明,DCFS2所采用的基于分布式日志的元数据处理策略能够提供高的I/O性能,并能够保证在元数据服务器失效后文件系统快速恢复.

9. 学位论文 李剑宇 基于对象存储的机群文件系统数据通路关键技术研究 2007

近年来机群系统凭借良好的可扩展性、可用性以及极高的性价比成为高性能计算机和超级服务器的主流结构.然而,磁盘性能的改善远远落后于CPU处理速度、内存性能、互连网络带宽的提高,使得I/O系统成为严重制约机群系统性能提高的瓶颈.机群文件系统作为缓解机群系统I/O瓶颈的重要手段,具有重要的研究意义.

为了提供可扩展的文件服务,大规模机群文件系统的发展趋势是数据通路与元数据通路相分离、基于对象存储的网络存储系统和多元数据服务器的体系结构.

针对这种结构的大规模机群文件系统,本文重点研究了其中的几个关键问题,包括I/O性能的关键优化技术、多负载的高效支持以及数据的高可用.本文的主要贡献在于:1)设计和实现了基于对象存储的大规模机群文件系统LionFS,重点研究了在这种新的存储体系结构下,I/O性能的关键优化技术,包括:异步化、直接递送的数据传输机制以及基于前端负载访问信息的预取技术,测试表明,由于直接递送的方式减少了传输过程中客户端和对象设备的内存拷贝,使得读写性能都有比较大的提高,分别为:读24％,写28％,而且采用预读技术后系统的数据通路达到并发流水,使读带宽增长了70％;2)提出了一种类会话的I/O访问机制以及面向文件访问请求的全局调度机制.本地环境通过连续的文件布局并配合CFQ磁盘调度策略能够很好地改善多个并发顺序流的聚合性能.然而,机群文件系统基于网络传输的数据访问方式使得存储服务器的驱动方式与负载构成不同于本地环境,限制了上述策略的优化效果.为此,我们设计了类会话的I/O访问机制以充分发挥本地系统ext3数据块预留机制的优化效果,并在此基础上提出了面向文件访问请求的全局调度机制以缓解多负载并发访问磁盘时的相互干扰.性能测试表明:综合使用上述策略能够明显改善机群环境下多个并发顺序流的聚合性能.3)提出了一种基于复制技术的数据高可用机制,针对复本一致性、故障记录和数据恢复、在线恢复等问题,该机制扩展了标准的文件锁协议来保持单复本的Unix文件共享语义;利用轻量级的故障记录机制来降低日志操作的性能开销并通过重放更新的修复策略在故障排除后进行高效的数据修复;最后借助文件冻结技术实现在线修复,满足数据高可用的需求,性能数据表明:该机制在故障发生后仍然能够保持系统具有较好的性能,而且在故障排除后能够快速完成数据的恢复.

10. 期刊论文 贺劲.吴思宁.孟丹.徐志伟 网络文件系统中的元数据存取优化研究 -微电子学与计算机2003,20(3)

文章研究了客户机/元数据服务器/存储服务器三层结构网络文件系统中,针对改善元数据服务效率的优化策略,包括客户节点接口中的路径解析加速、元数据服务器上的元数据存取优化及元数据服务器组的结构支持等几部分,作者通过建立分析模型与实际系统模拟的方法,证明这些策略的确可改善系统性能,优化系统效率.