DCFS机群文件系统服务器组关键技术研究

Research for the Key Technolog of Server Group in DCFS Cluster Filesystem

中国科学院计算技术研究所智能中心 吴思宁 贺劲 熊劲 孟丹 (北京 100080)

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。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。

关键词：机群文件系统，元数据服务器，存储服务器

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对其分类[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]、国家智能计算机研究开发中心的 COSMOS[7]及 PVFS[12]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：基于物理磁盘与逻辑虚拟磁盘，如 GPFS 等都是基于逻辑磁盘设备；而 NFS、COSMOS 都是基于物理磁盘。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 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。

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。用户可通过两种方式使用 DCFS 机群文件系统：

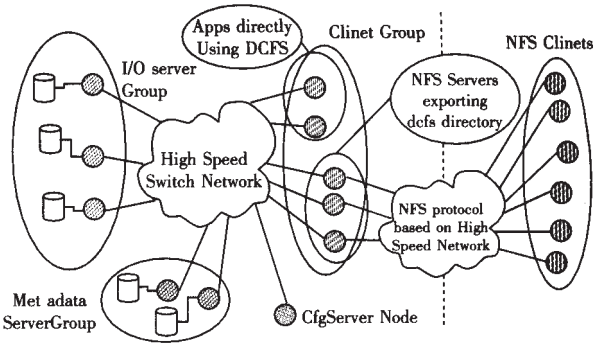

图 1 DCFS 结构示意图

(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。

DCFS 文件系统是一个单一映象的文件系统，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，系统中对应的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减少单个超级管理器节点崩溃时不会影响其它 DCFS 文件系统。

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 用户看来即多个不同

收稿日期：2002-11-08

的 DCFS 文件系统卷。在单个分组内,各存储服务器节点以类似 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

图 2 给出了研究人员在为网络文件系统 NFS 构造一个新的 SPEC 基准(SFS 2.0)前,对 750 台运行 NFS v2(版本 2)的 Auspex 服务器进行专门的负载研究的结果[3]。从图中的数据可以看到,在某些应用模式中,与元数据 6 相关操作所占比重要大于文件的读写操作,所以元数据服务器设计的好坏对于 DCFS 性能有重要的影响。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。

Operation	Cluster 1	Cluster 2
null	1%	0%
getattr	25	15
setattr	1	1
lookup	40	23
readlink	7	2
read	11	32
write	5	17
create	2	1
remove	1	0
readdir	7	5
fsstat	1	2

图 2　数据存取操作比例统计结果

3.1 DCFS 名字空间的组织

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所有低级子目录的元数据都存放在相同该元数据服务器。如图 3 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) 对于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,如果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。

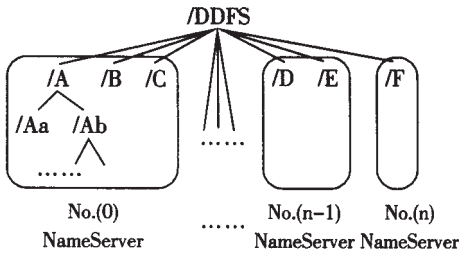

图 3　DCFS 名字空间组织

3.2 目录缓存管理

DCFS 目录缓存有两类作用: 加速名字查找(Lookup 操作)过程;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可能对同一个 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,此时 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。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(dentry)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。每个元数据服务器有一个按名字 HASH 函数组织的 HASH 表,使得 Lookup 操作能根据文件名快速查找到相应的文件的元数据信息。

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 文件系统卷的管理

为了有效地发挥机群中每个节点的存储性能,同时为了便于管理, 在 DCFS 中支持多个文件系统卷, 客户端可以看到多个 DCFS 文件系统卷,并

mount到本地目录下。每个 DCFS 文件系统卷采用了网络存储分组的技术。在 DCFS 中,一个存储分组是存储文件的基本单位,每个 DCFS 文件以普通 Unix 子文件的形式分散存放在某一个存储分组内的所有磁盘上。在网络带宽高于磁盘读写带宽的前提下,存储分布将能提高读写并行度,从而提高读写性能;系统中可以配置多个存储分组,不同的文件会落在不同的存储分组内,多个客户对这些文件的 I/O 请求可由不同的存储分组来响应,从而有效地解决了单一服务器的性能瓶颈问题。

图 4 描述了 DCFS 中的一种典型文件系统卷的构成,从图 4 中可以看出,DCFS 的卷的配置非常灵活,构成文件系统卷的单位是磁盘,而不是存储服务器(Storager),每个 Storager 都可以带多个磁盘,而每个卷可以根据需要跨几个 Storager。

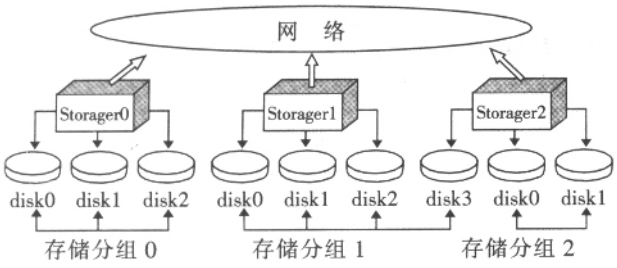

图 4 DCFS 的一种卷的配置

DCFS 提供了简单易用的界面给用户用于创建、删除一个文件系统卷。

4.2 DCFS 存储服务器线程关系

DCFS 存储服务器由多个线程构成(见图 5):

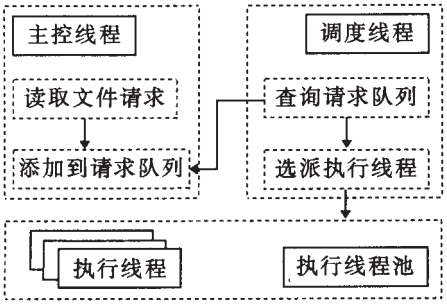

图 5 DCFS 存储服务器各线程关系

(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取一个待处理的请求,选择空闲的执行线程并唤醒它处理相应的请求;

(3)执行线程负责处理调度线程分派的请求。

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(ExecutorThreads)执行这些 I/O 请求。对于数据长度较大的文件读写类型请求,需要指定两个线程来合作完成这个请求:一个线程负责磁盘 I/O 而另一个负责网络 I/O,使得磁盘 I/O 与网络 I/O 能够进行流水操作,以优化读写性能;对于其它类型请求,由一个执行线程完成所有工作;

(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。对于小数据量请求或者文件删除与截断请求,一般由③类型线程完成;大数据量请求则由①与②类型线程协作完成。

4.3 DCFS 存储服务器缓存设计

DCFS 文件系统提供了对多各卷的支持,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,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,用来将存储服务器缓存中被执行线程改写、还未同步到磁盘上的数据块组织在一起,以便定期同步线程来同步这些数据块。

(3) CLEAN 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,存放未被分配的缓存块。

5 性能测试

测试平台是一个 64 节点构成的曙光 Linux 机群系统,每个节点的配置为两个 Intel P4 Xeon 2.2GHz CPU,内存 1GB,硬盘 Seagate 80GB,节点操作系统是 Linux 2.4.18-SMP,节点间通过 Myrinet 网络高速互联。

在图 6、7、8 的测试中,DCFS 的文件服务器配置在 8 个节点上,每个服务器节点既是存储服务器又是元数据服务器。PVFS 的文件服务器配置在 9 个节点,其中一个节点为元数据服务器,其它 8 个节点是存储服务器。测试程序是我们针对分布式/机群文件系统开发的测试程序 FSbench。在图 6、7、8 的性能曲线图中,横坐标单位是 MB,表示每个客户节点进程读写文件的长度;纵坐标的单位是 KB/s,

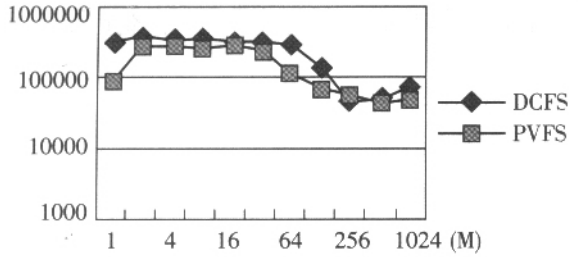

图 6　DCFS 和 PVFS 写聚合带宽比较

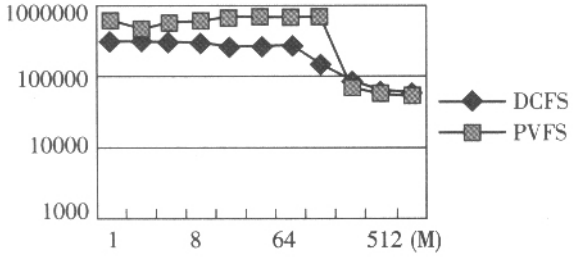

图 7　DCFS 和 PVFS 读聚合带宽比较

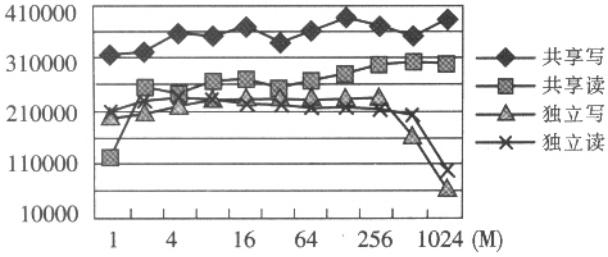

图 8　DCFS 共享读写性能与独立读写性能比较

表示聚合读写带宽,使用对数坐标。

在图 6、7 中给出了 DCFS 和 PVFS 带宽测试的结果,测试程序运行在 16 个客户节点上,在每个客户节点上运行两个测试进程。从图中可以看出,DCFS 的写性能要优于 PVFS,这是因为 DCFS 在存储服务器上实现了缓存管理。DCFS 的读性能在读写文件长度小于 256MB 时差于 PVFS,而对于大文件,DCFS 与 PVFS 性能相当。通过对 PVFS 读操作的分析,我们知道 PVFS 在存储服务器使用了 Sendfile 系统调用直接发送文件数据,而 DCFS 在存储服务器上有缓存,在读性能测试中顺序读文件的条件下,DCFS 多了一次内存拷贝的操作,因此读性能要稍差。对于文件访问局部性较好的情况,DCFS 能获得较好读写性能。图 8 给出了 DCFS 共享读写性能和非共享读写性能的比较,由于存储服务器缓存的作用,在共享的情况下,DCFS 能获得较高的性能。

下面是我们用 FSbench 分别在 NFS 和 DCFS 上进行吞吐率测试得到的结果,测试平台和读写性能测试时相同。配置是 DCFS 的元数据服务器节点为 8 个,NFS 仍是一个服务器节点,客户端为 16 个节点。每个客户节点有一个测试进程,每个吞吐率的测试时间为 10 分钟。

类似于 SFS Benchmark 中采用的表示方法[11],图 9 的横坐标是聚合吞吐率,单位是 Ops/s,它表示的是文件系统在一定的负载水平下得到的实际的吞吐率,纵坐标是平均响应时间 Resp,单位是毫秒,图中在横坐标方向往回拐的点即为饱和吞吐率。在图 9 中,由于 DCFS 的文件服务器可以扩展成多个节点,而 NFS 的服务器无法扩展,所以 DCFS 的负载能力要优于 NFS。

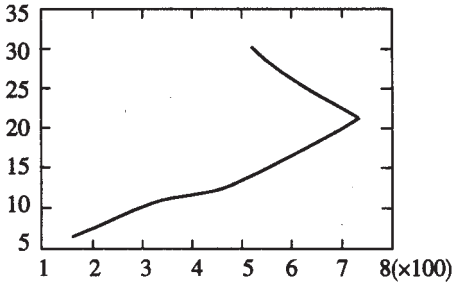

图 9(a)　NFS 吞吐率测试

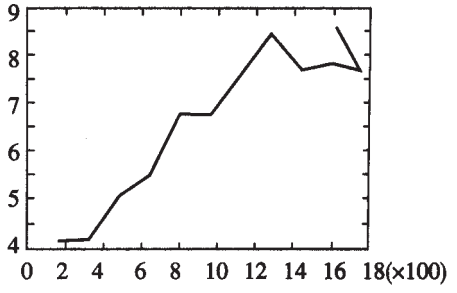

图 9(b)　DCFS 吞吐率测试

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。性能测试表明,我们在服务器组设计中采用的技术是可行的。

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]. 中国科学院计算技术研究所, 2002.6.

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

[3] D Robinson. The advancement of NFS benchmarking: SFS

(下转第 37 页)

tems J, Jan. 1997,5(1):23~38.

[6] H P Dommet, J.J. Garcia-Luna-Aceves. A novel group coordination protocol for collaborative multimedia systems. In Proc. IEEE Int. Conf. On Systems, Man, and Cybernetics, San Diego, CA, Oct. 1998, 2:1225~1230.

[7] H P Dommet, J.J. Garcia-Luna-Aceves. Efficacy of floor control procotols in distributed multimedia collaboretion. Cluster Computing J., 1999, 2(1):17~33.

[8] Timothy K Shih, Lawrence Y Deng, I-Chun Liao, Chun-Hung Huang and Rong-Chi Chang. Using the floor control mechanism in distributed multimedia presentation system, Distributed Computing Systems Workshop, 2001 International Conference on, 2001: 337~342.

[9] Nadia Kausar,Jon Crowcroft.General conference control protocol, Telecommunications, Mar.29-Apr.1, 1998:143~150.

[10] 高旭, 沈苏彬, 顾冠群. 多媒体会议系统的发言权控制协议研究. 计算机学报, 2001.8.

[11] Qian Yi, Hou Yibin. A fuzzy floor control policy for interactive and cooperative system. The 2nd International Workshop on Autonomous Decentralized System (IWADS2002), Beijing, China, Nov. 2002: 147~150.

[12] 黄樟钦, 侯义斌, 田友胜, 黄进锋. 基于 OP 通道技术的远程交互系统设计与实现. 计算机工程与应用, 2000, 2(36):1~3.

[13] 田友胜, 侯义斌. R-J Benun, 黄樟钦, 黄进峰. 基于 Internet/Intranet 的命令通道机制人人交互与协作系统模型. 小型微型计算机系统, 2001, 2(36):430~432.

[14] Tian Yousheng, Hou Yibin, Qian Yi, Chen Yabin. A fuzzy multicast routing algorithm for interaction and co-operative system over the internet, 2001 international conference on Info-tech and Info-net, Beijing, China, Oct.29-Nov.1, 2001: 393~400.

QIAN Yi, HOU Yi-bin (School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049)

Abstract: Floor control is the technology of coordinating and controlling the shared resources concurrently used in the interactive and co-operative environments. It is an important coordinative controlling mechanism in interactive and cooperative systems. Based on the discussions of floor control and its control policies, a fuzzy floor control policy is presented. It integrates many policies basis weights, and makes up their drawbacks, and can be adjusted according to the instances of the system and the requirements of users, through modifying the weights of policies. The testing result indicates that the fuzzy policy is more fairly, interactive and flexible, and it makes the system more general.

Key words: CSCW, Interactive and co-operative, Floor control, Fuzzy control policy

钱 屹 博士研究生。主要研究方向为广义人机交互和多播技术。

侯义斌 教授,博士导师。主要从事 Internet 理论与技术、广义人机交互、多媒体数据库等方面的研究工作。

(上接第 33 页)

2.0. In Proceedings of the 13th USENIX Systems Administration Conference, 1999:175~185.

[4] Thomas E Anderson, Michael D.Dahlin, Jeanna M Neefe, et al. Serverless Network File Systems. ACM Transactions on Computer Systems, 1996,14(1): 41~79.

[5] Kenneth W Preslan, Andrew P Barry, Jonathan E Brassow, et al. 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, et al. GPFS: A Parallel File System. SG24-5165-00. April 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]. 中国科学院计算技术研究所, 2001.6.

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

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

[10] SPEC SFS97 R1 V3.0 Documentation.

[11] Philip H Carns, Walter B Ligon, et al. PVFS: A Parallel File System for Linux Cluster.

WU Si-ning, HE Jin, XIONG Jin, MENG Dan
(National research center for intelligent computing systems, Institute of Computing Technology, Chinese Academy of Sciences, 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

吴思宁 男,(1975-),博士研究生。主要研究方向为机群文件系统与操作系统。