

一种高性价比的 PVFS 并行文件系统

A parallel file system based on PVFS with high performance and low cost

(深圳大学)罗秋明 雷海军

Luo, Qiuming Lei, Haijun

摘要:分析 PVFS 并行文件系统的构成,得出客户机软件、元数据服务器软件和数据服务器软件之间的接口关系,然后研究一种由 PC 客户机、PC 元数据服务器和低价数据服务器共同构成的 PVFS 系统,其中客户及与元数据服务器不做重要改变,数据服务器软件需要开发修改以适应新的硬件平台,使得以更低的成本实现相同的系统或者以相同的硬件成本实现更高的性能。

关键词:PVFS;数据服务器;性价比

中图分类号:A

文献标识码:TP391.41

Abstract: The interface between client, meta-data server and file data server software is drawn by analyzing the PVFS system. A new PVFS system is given with high performance on low cost, which is composed of PC client, PC meta data server and a low cost file data server. The client and meta server are remain running on PC hardware, but the file data server is porting form PC hardware to other low cost platform.

Keywords: PVFS, IOD, price performance ratio

1 引言

PVFS(Parallel Virtual File System)并行文件系统有三种灵活的用户访问模式、简单方便的系统配置过程,与事实上的并行编程标准 MPI(Message Passing Interface)的主流实现 MPICH2 的 MPI-IO 有直接的接口——ROMIO,是迄今为止 Linux PC 集群系统中最成功的并行文件系统之一,被广泛的应用和研究。

PVFS 往往是在 Linux PC 集群上,与计算用的节点共享 PC 硬件,这样部署时,元数据服务器(Meta Server)MGR 软件和数据服务器(Data Server)IOD 软件是直接安装在计算节点上的,也就是说集群上的每个节点既是计算用的节点也适用于并行 IO 的节点,因此这种方式会因为计算节点进行 IO 操作而影响正常的计算处理,降低系统的效率。另外一种方式是形成单独的数据服务系统利用多台 PC 安装 MGR 和 IOD 来专门提供并行文件服务,这样可以实现计算任务与并行文件 IO 任务的分离,类似于通道处理器减轻主处理负担的作用一样,可提升系统的计算性能和文件 IO 性能。

第一种方式影响性能而第二种方式的 PC 硬件代价太高,因此需要开发一个更高性价比的 PVFS 系统以实现不影响计算节点的性能并降低系统整体成本。首先分析系统中三个部件之间的互动与接口关系,然

后设法通过软件开发与移植来替换昂贵的 PC 硬件基础上的数据服务器。

2 PVFS 系统分析

并行文件系统 PVFS 的系统构成形式如所图 1。系统中存在三类节点:运行 libpfvs 并行文件库的客户机、运行 MGR 的元数据服务器和运行 IOD 的数据服务器。在客户端的文件请求处理过程中,需要先元数据服务器获得文件数据在数据服务器上分布情况的有关信息,然后客户端才通过网络并行地向一个或多个数据服务器请求所需的数据块。

图 1 PVFS 系统构成

图 1 中 CN0-CNn 是客户节点,它们通过网络与运行 MGR 程序的元数据服务器 Meta Server 和运行 IOD 的 ION0-IONn 数据服务器 IO Server 节点相连。节点之间通信的数据结构是多种 req 请求和 ack 响应

罗秋明:助研 博士

基金项目:广东省自然科学基金项目(No.5301033);

深圳大学科研启动基金 200501

结构,因此在理解 PVFS 原理时关键是要抓住三种类型的节点以及它们之间的请求和响应的数据结构和通信流程。客户端的接口有三种形式:使用 libpvfs 库的专用函数、通过 VFS 接口、在 MPI 环境中的 ROMIO (MPIO 的一种实现)接口。

客户机进行一次文件操作的过程如下:向运行 MGR 程序的 Management 节点发出元数据请求,获得文件在数据服务器上的分布情况;然后根据分布情况和数据偏移及数据量,可能对一个或多个数据服务器进行一次或多次读写,由于是多个数据服务器同时向所有客户机服务,因此系统的总吞吐量是各个数据服务器的总和。

由于 IOD 与客户机和元数据服务器之间的联系完全通过网络进行,因此从客户机和元数据服务器的角度上看,IOD 就是一个能够处理特定 req 请求并返回合适的 ack 响应数据的网络服务器,任何能够实现相同服务的实体替换这些 IOD 时,对于客户机和元数据服务器都是不可见的。同理,用相同功能的其它实体替换客户机或元数据服务器也对另外的两种部件是不可见的。

3 系统整体性能价格分析

如果利用一台完整的 PC 专门来做数据服务器,若只带几块硬盘则显得比较浪费,即使配置 RAID 卡来控制更多硬盘也使得该节点的数据服务受限于网络接口。因此,当存储容量或者说硬盘数目一定时,采用前一种方式系统总带宽较大,但是硬件成本很高,采用后一种方式则使系统的总带宽有所下降。

假设由 N 个容量为 S_d 磁盘构成,则系统总容量为: $S_t = N * S_d$,如果每台 PC 带 M 个硬盘,系统地最大数据吞吐量为: $B_M = (N/M) * B_{net}$,其中 B_{net} 为网络接口带宽,系统总造价为: $C_M = (N/M) * (C_{pc} + C_{raid})$,其中 C_{pc} 是 PC 价格 C_{raid} 是 RAID 卡价格。由此可以看出当 M 值上升时,成本造价下降了但是系统的总吞吐量也下降了,反之,则带宽是高了但是造价也高了,因此在这种系统架构下存在这一对矛盾。

因为数据服务器不需要 PC 的强大计算能力,因此 PC 的高性能在此是一种浪费,也是造成价格高昂的原因之一。利用计算能力比 PC 弱的其他低价硬件系统来专门提供数据服务是另一种选择,例如可用 ARM+ASIC RAID 方式形成嵌入式盘阵列,此时系统的带宽仍为: $B_M = (N/M) * B_{net}$,但是系统造价变为: $C_M = (N/M) * (C_{arm} + C_{asic-raid})$,由于新系统的价格 C_{arm} 远远小于 PC 造价 C_{pc} ,因此在相同系统带宽的条件下造价将下降很多,或者反过来说相同造价下,利用新系统的 M 值可以比 PC 平台取得更小从而获得更大的系统带宽。

图 2 两种不同构架的性能价格比较图,其中假设 PC 服务器价格 16K RMB,RAID 卡 2K RMB,ARM 平台 1K RMB,ASCI RAID 价格 2K RMB,每个磁盘按

1K RMB 计算,纵坐标是系统总造价(含磁盘)

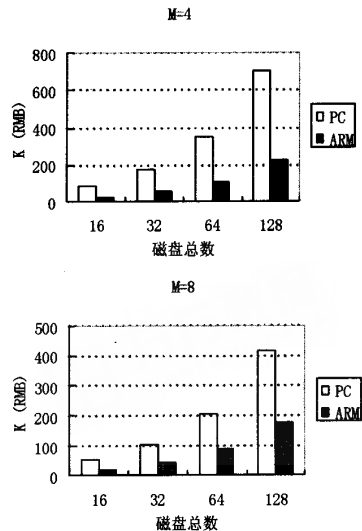


图 2 两种架构下的系统硬件总价

因此将 PVFS 数目众多的数据服务器从 PC 平台转换为其他低价硬件平台对降低系统总造价非常有用。

4 数据服务器软件

需要开发修改数据服务器软件以便能在其他平台上运行,并保证新代码的功能与 PC 客户机和运行 MGR 的元数据 PC 服务器之间的网络通信行为完全不变,即移植后的 IOD 对客户机传送来的 ireq 请数据包的解释和处理与原来相一致。这时 IOD 是一个网络守护进程 daemon,它接受客户机的 ireq 请求,处理完毕后向客户机返回 iack 响应,客户机所需的数据在 i-ack 的数据段中。

所编写的软件的主要功能是:在 acc 端口上监听新的连接,如果有新连接发起则接受连接请求并准备后续工作,如果在已网络连接上新来的 req 则进行立即处理或者需要建立 jobs 在 do_jobs()中去完成,后者现在只有读写两个请求需要 jobs 支持,其他的 req 都是直接处理并返回 ack 数据包。这些来自客户机或者元数据服务器的不同请求,被接受进来后首先分析其请求类型,然后根据函数跳转表跳转到相应的处理函数中去,跳转函数表如下:

```
static int (*reqfn[])(int, ireq_p, iack_p) = {
    do_open_req,
    do_close_req,
    do_stat_req,
    do_unlink_req,
    do_rw_req,
    do_shutdown_req,
    ...,
    do_ftruncate_req,
    do_truncate_req,
    do_fdatasync_req,
```

```
do_stats_req,
do_list_req );
```

编写好相应的执行函数完成 IOD 应有的功能,这些功能函数依据具体的请求完成相应的文件操作,然后将操作结果通过网络数据包返回,如果是读操作则还要将从此判读来的数据在返回数据包中传回,写操作则需要将要写的数据从网络上接受完全并写入磁盘。对于读写操作,由于比较复杂和耗时,因此需要为每个请求生成一个 Job 列表,因此 IOD 中的一个 socket 可能对应多个 job 任务,这些任务又具体细分成更小的“访问”Access 数据结构,这些关于文件数据请求的 job 和 Access 数据都构成链表,具体构成如下图 3:

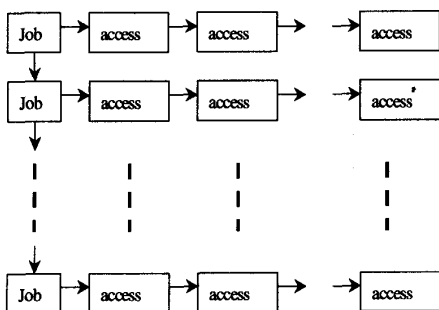


图3 任务 Job 和 Access 关系

如果选择的硬件平台支持内存管理 MMU,那么许多在无 MMU 的硬件平台上的许多移植问题都不存在,比如虚存和线成的问题,因此软件移植并不十分困难。

IOD 软件开发工作是在 ATMEL 公司 ARM920 TARM 处理器的开发平台上进行,开发板上集成了 IDE 控制器,可携带普通 IDE 硬盘。该开发环境需要宿主 PC,在宿主机上编写软件以及进行编译,由于目标板为非 PC Intel 架构,因此开发环境中选用的编译器不是默认编译器而是交叉编译器,在 PC 上编译出适用于开发平台的目标代码,然后通过串口或者网络下载到目标板的 Flash 盘上。

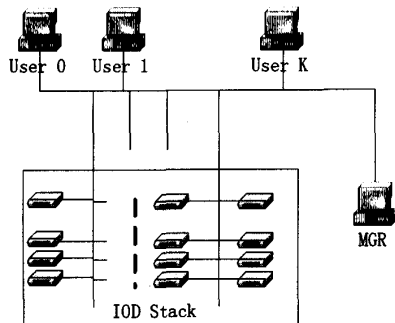


图4 嵌入式 IOD 数据服务器堆

4 结语

本文的创新点:通过开发修改 PVFS 的 IOD 软件,将 PVFS 系统的数据服务器从全部基于 PC 硬件平台

替换成客户机与元数据服务器仍在 PC 平台上而数据服务器运行于其他低价的平台上,关键是开发新的数据服务器软件并保证它能对客户机传来的 ireq 网络数据请求包的正确解释和处理,由于数据服务器的造价所占比例很大,利用低价硬件平台可以明显降低系统总造价,而文件系统功能是不涉及科学计算(浮点性能)的磁盘访问和网络访问,整体性能不会比 PC 平台的有所降低,是一种高性价比的 PVFS 并行文件系统。

参考文献:

- [1]Robert B R, Philip H C, Walter B L III, et, al. Using the Parallel Virtual File System [EB/OL].<http://www.parl.clemson.edu/pvfs/>, 2005-05-01
- [2]PVFS2 Development Team. Parallel Virtual File System, Version 2[EB/OL]. <http://www.pvfs.org/pvfs2/pvfs2-guide.html>, 2005-05-01.
- [3]罗秋明,雷海军,结合隐式元数据和 List IO 的并行文件系统[J]微计算机信息,2006,6-3:246-248

作者简介:罗秋明(1974-),男(汉族),广东人,深圳大学超级计算中心助理研究员,博士,主要研究方向:计算机体系结构,集群计算等;雷海军(1968-),深圳大学超级计算中心副教授,主要研究方向:计算机应用、并行计算。

Biography:Luo, Qiuming, born in 1974, Ph.D., research field: computer system architecture, cluster computing; Lei Haijun, born in 1968, Ph.D. vice-professor of Shenzhen University.

(518060 深圳大学超级计算中心)罗秋明 雷海军
(SuperComputing Center, Shenzhen University, 518060)Luo, Qiuming Lei, Haijun

(投稿日期:2005.12.5) (修稿日期:2006.1.6)

(接 282 页)参考文献:

- [1]侯振挺,郭青峰.齐次可列马尔可夫过程[M].北京:科学出版社,1978
- [2]Craig Goren. Visual Basic 6.0 开发指南[M].北京:清华大学出版社,1997
- [3]马靖,陈江林,张兆林.利用 API 增强 VB 的图像处理功能[J].微计算机信息,2003,(9):102-104
- [4]韩於冀.应用数理统计[M].北京:航空航天大学出版社,2002

作者简介:秦翔宇(1981-),男,河北邢台人,石家庄军械工程学院管理科学与工程学科硕士研究生,研究方向为系统决策与仿真、决策优化等;宋一中(1963-),男,江西万载人,副教授,石家庄军械工程学院装备指挥与管理教研室主任,主要研究领域为管理信息系统开发、军事运筹学、系统决策与建模等。


(050003 河北石家庄军械工程学院装备指挥与管理系)秦翔宇 宋一中

(Ordnance Engineering College, Department of Equipment Control and Management, Shijiazhuang 050003, China) Qin, Xiangyu Song, Yizhong

通讯地址:(050003 石家庄军械工程学院装备指挥与管理系装备管理教研室) 秦翔宇

(投稿日期:2005.12.7) (修稿日期:2006.1.11)

一种高性价比的PVFS并行文件系统

作者: [罗秋明](#), [雷海军](#), [Luo, Qiuming](#), [Lei](#), [Haijun](#)
作者单位: [518060, 深圳大学超级计算中心](#)
刊名: [微计算机信息](#) 
英文刊名: [CONTROL & AUTOMATION](#)
年, 卷(期): 2006, 22 (22)
被引用次数: 1次

参考文献(3条)

1. Robert B R, Philip H C, Walter B L III. et, al [Using the Parallel Virtual File System](#) 2005
2. PVFS2 Development Team [Parallel Virtual File System](#) 2005
3. [罗秋明, 雷海军](#) [结合隐式元数据和List IO的并行文件系统](#)[期刊论文]-[微计算机信息](#) 2006

相似文献(9条)

1. 学位论文 [吴一波](#) [并行文件系统负载均衡技术的研究与实现](#) 2009

随着集群技术的不断发展,并行文件系统作为集群的I/O系统也越来越得到重视。然而,在很多应用中并行文件系统的性能受到削弱,其重要原因就是负载不均衡,使得某些部件成为瓶颈,制约了整个系统的吞吐量。因此,负载均衡对提高系统的性能具有重要作用。本文以并行虚拟文件系统PVFS (Parallel Virtual File System)为基础,深入研究了并行文件系统中数据访问的负载均衡问题,并且在研究分析的基础上提出了平衡负载的方法。

PVFS是一个开放源码的并行文件系统,是迄今为止Linux集群系统中最成功的并行文件系统之一。PVFS具有良好的并行性,较高的可用性、可扩展性和性能,但是其数据服务器缺少负载均衡能力,降低了PVFS的性能,严重制约了系统的吞吐量。本文在PVFS文件系统的基础上,深入研究了分别基于副本和数据迁移的负载均衡技术,分析两者的优缺点,并最终提出和实现了一种数据迁移与副本相结合的BRM (Based on Replication and Migration)负载均衡技术,可以有效的解决PVFS文件系统由于负载不均衡导致的性能瓶颈问题。主要研究工作包括以下几个方面:

- (1)提出了BRM负载均衡技术的体系结构。在分析了负载均衡技术的关注点及其要求的基础上,结合PVFS文件系统本身的特点提出了BRM负载均衡技术的体系结构,采用模块化的思想,将系统分成了热点监测、数据迁移与复制、负载均衡调度三个模块。
- (2)根据BRM负载均衡技术总体框架,首先深入分析了热点监测、数据迁移与复制、负载均衡策略三个关键技术,然后详细描述了BRM负载均衡技术的核心策略,通过选择源文件时权衡文件的热度与大小以降低数据迁移与复制的开销,并根据数据访问方式选择进行迁移或副本,将热点数据以较小代价和适当的方法转移到较空闲的服务器上,有效地提高了整个系统的数据吞吐量。
- (3)实现了BRM负载均衡技术。依据BRM负载均衡技术的总体框架和模块化设计,并结合相关研究,详细描述了实现热点监测、数据迁移与复制、负载均衡策略的关键技术,给出了相应的代码。最后借助系统的应用平台对BRM负载均衡技术进行了相应的试验和测试,通过对试验结果的分析讨论验证了本文工作的可行性和有效性。

关键词: 负载均衡, 热点监测, 迁移, 副本, 策略

2. 期刊论文 [罗秋明, Luo Qiuming](#) [一种算法无关的PVFS负载均衡机制](#) -[计算机工程与应用](#)2006, 42 (19)

并行文件系统PVFS的数据服务器负载分配缺少均衡机制,因此无法减轻热点数据服务器负荷。通过文件数据备份的方式将任务迁移,同时解决在迁移过程中的数据服务器与元数据服务器上的数据一致性,将热点数据服务任务转移到较空闲的服务节点上,可以提高整个系统的数据吞吐量。主要涉及了热点监测与报告、迁移源-目的选取以及数据备份与任务迁移过程三个部分的工作。由于采用了机制与策略分离的设计思想,在该框架下可以采用不同的均衡策略(算法)以适应不同的应用。

3. 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 key 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,给出了可调度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受的。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写带宽性能上比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%。

4. 期刊论文 [吴一波, 赵英杰, 肖依, 刘波, Wu Yibo, Zhao Yingjie, Xiao Nong, Liu Bo](#) 一种基于副本的PVFS负载均衡

机制 - 计算机研究与发展2009, 46(z2)

并行虚拟文件系统PVFS的数据服务器缺少负载均衡机制, 因此存在热点服务器, 降低了系统整体性能. 提出了一种基于副本的负载均衡机制, 通过文件数据备份的方式进行负载迁移, 以解决这一瓶颈问题. 其通过选择备份文件时权衡文件的热度与大小以降低数据备份的开销, 将热点数据以较小代价转移到较空闲的服务器上, 有效地提高了整个系统的数据吞吐量. 其主要涉及了热点监测、数据备份源-目的节点选择以及备份文件策略3个部分的工作. 实验结果表明: 提出的负载均衡机制有效地提高了系统的整体性能, 最高达到了24%.

5. 学位论文 [付印金](#) PB级文件系统元数据管理关键技术与实现 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.

6. 期刊论文 [罗秋明, 欧阳凯, LUO Qiuming, OUYANG Kai](#) PVFS元数据服务器的并行化设计与实现 - 计算机工程

2006, 32(12)

PVFS并行文件系统采用集中式的元数据服务器, 这使得元数据服务器在大量的文件操作情况下成为I/O瓶颈. 该文通过增强PVFS的客户端和元数据服务器的相关功能, 使得客户的文件请求按照文件名的Hash变换转向不同的元数据服务器, 在保留原来的用户访问方式和系统配置不变的情况下实现元数据服务器的并行化, 达到明显提高元数据服务器的总吞吐量和可扩展性.

7. 学位论文 [何飞跃](#) 并行文件系统元数据管理研究 2004

随着高性能微处理器、高速网络的出现和对计算能力需求的增大, 以廉价硬件和软件支撑的集群系统越来越被广泛地使用, 引起集群技术的迅猛发展. 集群文件系统是集群的一个重要组成部分, 作为一种集群体系结构上的并行文件系统, 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结点上, 文件的定位需要借助元数据来完成, 元数据的管理成为管理数据的一个关键. 为了提高元数据管理的可靠性, 需要具有容错能力的元数据管理系统. 为此, 我们针对集群文件系统的元数据管理, 设计了一个双元服务器系统. 该系统内部由两台元数据服务器组成, 通过对元数据的镜像产生副本, 保证元数据的可靠性; 通过主服务器失效后从服务器接管服务来屏蔽故障, 保证元数据服务的连续性. 系统具有集中管理方式控制简单、易于实现和维护等优点, 克服了其单一失效点的缺陷, 同时又避免了分布式管理的一致性维护设计与开销. 在Linux内核空间实现元数据镜像技术、故障检测技术、IP接管技术和恢复技术, 具有对应用程序透明性的特点. 系统的最终目的是将其结构推广到多机情况下, 进一步提高容错能力, 实现高可用性. 为了提高元数据服务器的处理效率, 提出了一种寄生式元数据存储管理方法. 并行文件系统的元数据寄生在本地文件系统内核中, 通过增加系统调用实现对寄生元数据的操作, 保证对现有系统的兼容性. 将该方法应用于PVFS (Parallel Virtual File System) 的元数据管理, 元数据操作性能能提高大约5~8倍.

8. 期刊论文 [王梅, 罗秋明, Wang Mei, Luo Qiuming](#) PVFS代码结构及并行Meta服务研究 - 微计算机信息2006, 22(16)

PVFS (Parallel Virtual File System) 广泛应用于PC集群并行计算环境中, 通过ROMIO形式的MPI-IO接口与MPICH结合, 用于提高数据文件的访问性能. 通过对PVFS的源代码分析, 得出PVFS的系统架构、运行机制与采用的策略, 并在此基础上找出元数据服务器并行化的可行性, 以设计出并行元数据服务方案来提高元数据访问的吞吐量.

9. 学位论文 [岳建辉](#) 基于集群的高可用通信中冗余技术的研究 2003

集群系统是一个并行处理系统, 具有性价比高和可扩展的特点. 集群已成为高性能计算和超级服务器的基本构建方法, 它被运用在各个领域. 但还有许多具有挑战性问题有待解决. 其中, 针对一些关键性事务处理需要向用户提供不间断服务, 这就要求集群服务器具有容错能力. 冗余副本是容错计算中的一种技术, 其中的主动副本服务器是容错能力最强的一种. 组通信是主动副本服务器的基础, 现有的组通信是建立在用户层. 这种组通信不但对用户不透明而且消息的延迟较大. 讨论使用这些组通信技术构造基于PVFS的主动副本元数据服务器. 另外, 针对副本中的资源利用率较低的问题, 还提出面向高可靠和高性能的多IP链路的技术. 该技术在提高集群内部通信可靠性的同时也提高通信带宽. 由于基于IP, 它独立于底层通信链路从而也突破了现有port trunking技术有端口限制的要求.

引证文献(1条)

1. [姚大勇, 杨广文, 卢琳](#) 基于Impulse C的软硬件协同设计及应用 [期刊论文] - 辽宁工学院学报 (自然科学版)

2007(3)

本文链接: http://d.g.wanfangdata.com.cn/Periodical_wjsjxx200622090.aspx

授权使用: 中科院计算所(zkyjsc), 授权号: 7eb0ac7d-8d8d-4b95-addf-9e400128d814

下载时间: 2010年12月2日