高性能计算机并行文件系统

卢　凯　等编著

国防科技大学出版社

高性能计算机并行文件系统

卢　凯　迟万庆　冯　华
秦　莹　周良源　编著

国防科技大学出版社
·长沙·

图书在版编目(CIP)数据

高性能计算机并行文件系统/卢凯等编著.—长沙:国防科技大学出版社,2005.6
ISBN 7-81099-183-3

Ⅰ.高… Ⅱ.卢… Ⅲ.电子计算机—并行处理—文件系统
Ⅳ.TP311.13

中国版本图书馆 CIP 数据核字(2005)第 045665 号

国防科技大学出版社出版发行
电话:(0731)4572640 邮政编码:410073
E-mail:gfkdcbs@public.cs.hn.cn
责任编辑:耿 筠 责任校对:唐卫葳
新华书店总店北京发行所经销
国防科技大学印刷厂印装
*
开本:850×1168 1/32 印张:6 字数:156 千
2005 年 6 月第 1 版第 1 次印刷 印数:1-1500 册
ISBN 7-81099-183-3/TP·14
定价:12.00 元

前　言

I/O 系统一直是限制计算机系统整体性能提高的瓶颈。目前，并行 I/O 技术是克服 I/O 瓶颈的有效途径之一。但由于要求用户应用直接利用并行 I/O 子系统往往比较困难，所以人们设计了并行文件系统以向用户提供高性能的 I/O 服务。

本书深入研究了当前国际上提出的各种并行文件系统技术，分析归纳了其优缺点。在此基础上，作者从并行科学计算应用的 I/O 访问模式、面向并行科学计算应用的并行 I/O 服务模型及相关界面、并行 I/O 服务算法和 cache 预取算法等方面提出了提高其服务性能的若干技术。

并行文件系统的结构和服务模型同用户应用的访问模式密切相关。目前的 I/O 访问模式分析大多局限于静态的统计分析，因此缺乏对设计并行文件系统的指导意义。本文通过动、静态相结合的分析技术分析了若干典型并行科学计算应用的 I/O 访问特性，提取了并行科学计算应用的常见 I/O 访问模式，提出了可描述并行文件系统性能主要因素的馅饼模型，由此得出了设计和研究高性能并行文件系统的指导原则。

基于对并行科学应用 I/O 访问模式的分析，我们提出了面向磁盘的集成式并行文件系统服务模型(IDDIO)及相关界面。IDDIO 模型克服了现有模型只能对一类请求进行优化的缺陷，它可对同步 I/O 请求和独立 I/O 请求同时进行优化处理，使服务性能获得较大的改善。

在 IDDIO 并行文件系统服务模型框架下，本文进而分别研究

了对同步式 I/O 请求和独立式 I/O 请求的优化算法。针对同步式访问中现有并行 I/O 服务算法在数据细粒度分布时通信消息过多,易造成瓶颈的问题,提出了面向通信的分组并行 I/O 服务算法(CDGIO)。CDGIO 算法充分利用目前磁盘系统和通信系统的特点,采用两级并行,分组打包等策略获得了较好的性能优化。理论分析和模拟结果表明,该算法有效地缓解了现有服务算法中存在的通信瓶颈问题。

并行科学计算应用中独立 I/O 请求对访问服务延迟十分敏感。针对此特点,本文提出了适度贪婪的 cache 预取一体化算法(PGI)。PGI 算法基于用户应用的 I/O 访问模式,在适当时刻发出适度的数据请求,从而回避了现有贪婪或保守预取算法的缺点。PGI 算法采用了全局性的损失估计技术,综合评价淘汰和减少预取数据块的损失,获得了较短的整体服务延迟。同时,针对现有 cache、预取算法不适于并行文件系统环境的缺陷,一体化算法采用了汇集和负载均衡等技术,实现了较好的 I/O 访问性能优化。

最后,本书介绍了作者利用上述研究成果研制的 SkyWalk 并行文件系统的设计和实现。SkyWalk 并行文件系统采用了 IDDIO 并行服务模型、CDGIO 并行 I/O 服务算法和 PGI cache 预取算法。SkyWalk 并行文件系统提供了丰富灵活的用户接口和高性能的I/O服务,并可同时支持远程并行文件传输服务功能。

本书出版,得到了国家 863 重大软件专项服务器操作系统内核项目(2002AA1Z2101)的资助。

作 者

2005 年 5 月

目　录

第七章　结束语

第一章 绪 论

计算机作为当今信息时代典型的信息处理工具，对社会和经济的发展至关重要。高性能计算机系统为解决地质勘探、天气预报、生物医药等与人类生存和发展休戚相关的问题起了决定性的作用。

随着多媒体应用、网络应用、大型数据库系统、科学计算应用等进一步发展，人们对高性能并行计算机系统整体性能的追求是无止境的。从系统工程的角度来看，提高系统整体性能的关键是提高系统中性能受限部分，即瓶颈部分的性能。

从过去十年中计算机系统的发展来看，计算机微处理器的性能以年均35%的速度递增；计算机的访存时间也以30%～80%的速度递减；而以磁盘为代表的I/O子系统性能年增长率仅为7%[115]。表1.1显示了近年来CPU、内存和磁盘的发展对比情况[119]。磁盘、内存和微处理器间性能的不平衡性导致了在大多数计算机系统中I/O子系统成为制约系统性能提高的瓶颈。这种性能差距在提供了高性能处理能力的大规模并行处理系统中更为明显。

表 1.1 不同存储层次间的典型性能差异

	寄存器	cache	主存	磁盘
典型尺寸	< 1KB	< 4MB	< 4GB	> 1GB
访问时间(ns)	2 ~ 5	3 ~ 10	70 ~ 400	5000000
带宽(MB/s)	4000 ~ 32000	800 ~ 5000	400 ~ 2000	4 ~ 32
管理者	编译器	硬件	操作系统	用户/操作系统

单一的磁盘存储系统已无法满足大规模应用对 I/O 存储量、访问带宽的需求。随着并行技术的发展,人们在大规模并行处理系统中采用了并行 I/O 子系统,希望通过并行化的磁盘访问和数据传输向用户提供高性能的 I/O 服务。由于要求用户应用 I/O 直接使用并行 I/O 子系统往往比较困难,所以人们设计了并行文件系统以提供高性能的 I/O 服务。

并行系统的整体性能取决于服务执行的并行度。因此,高性能并行文件系统就成为开发用户 I/O 访问并行性,提供高性能 I/O 服务的主要途径。合适的并行文件系统结构和并行 I/O 服务算法等是将底层 I/O 子系统提供的大物理带宽转化为用户实际有效 I/O 带宽的关键。本文的工作正是在深入分析归纳当前国际上提出的各种并行文件系统技术优缺点的基础上,在 I/O 访问模式、并行 I/O 服务模型、并行 I/O 服务算法和 cache 预取算法等方面,研究了提高服务性能的若干关键技术。

1.1 应用对存储系统的需求

在信息时代中,人们对信息存储容量和访问速度的要求越来越高。随着高性能并行计算机系统处理能力的不断增强,其单位时间内可处理的数据量也不断增大。设每一条指令需要处理 1 个

或2个浮点数,每个浮点数用4字节表示,则2400MIPS的21264处理器需要9.6G的I/O带宽与之匹配[4]。大规模并行处理系统由多个微处理器组成,其峰值处理速度由处理机个数决定,故所需I/O带宽更加巨大。这就对存储系统的性能提出了更高要求。

目前,已有或新出现的要求计算机存储系统提供海量存储空间和大带宽I/O服务的应用主要有以下几类。[17]

1.1.1 多媒体应用[1~3]

多媒体信息的特点决定了多媒体应用除需要存储系统提供较大的存储容量外,还要求提供一定的实时性操作。

多媒体信息所涉及的数据量巨大。一个人体的全身CT切片数据量将达几十G之多。同时,多媒体应用还要求存储系统能提供一定的实时性。例如输出一秒钟的NTSC制式视频信号,画面的分辨率为512×480,每个像素点用24位表示,每秒30帧,则一秒钟画面的信息量为21.1MB。存储系统就必须提供约20MB的I/O带宽才能保证画面的连续播放。即使采用了压缩技术,多媒体应用对存储系统的存储容量和带宽要求也依然很高。

视频点播系统VOD(Video On Demand)是一个典型的多媒体应用,它同时对多个用户提供多条流的实时视频播放服务。其存储系统的I/O带宽需求随服务流数增加而增加。故VOD系统对存储系统的容量、带宽和实时性都有较高的要求。

1.1.2 广域信息服务[16]

短短几年时间里,WWW服务已风靡全球,成为当今最流行的网络信息服务。一个Web站点将服务于来自世界各地的请求,每个请求的数据量从几十字节到几十兆字节不等。有时单位时间内

Web 站点的请求数很少,例如在夜晚和休息日,而有时请求又十分密集。这就要求 Web 的存储系统能对这种变化剧烈的访问模式提供高性能的 I/O 服务。

1.1.3 大规模科学计算应用

除多媒体应用和广域信息服务应用外,科学计算应用对 I/O 系统的容量和带宽有更高的要求。

美国高性能计算和通信(HPCC)[4]计划列出了将利用大规模并行处理系统来解决的重大挑战性课题,如图形图像处理、天气预报、海洋环流、基因工程和流体动力学等。这些应用对存储系统性能都有极高的要求。HPCC 计划明确定义下一代高性能并行机系统应提供 1Tearflops 的处理能力、1TB 的主存储器和 1TB/s 的 I/O 带宽(3T 目标)。实际上,在主存储器和 I/O 系统性能方面,我们距此目标还相距甚远。同时,并行处理系统在运行中还将定时记录系统状态信息(checkpoint)以便故障恢复。checkpoint 时记录的状态信息量很大,并且完成时间越短,系统资源浪费越少,故对 I/O 系统也提出了较高要求。

1.2 并行文件系统研究中的主要技术问题

目前,并行文件系统研究主要存在以下三个方面的问题:

(1) 用户的 I/O 访问模式是什么? 不同的用户应用存在不同的 I/O 访问模式。只有针对用户的访问方式具体确定采用何优化手段,才能提供高性能服务。目前对 I/O 模式的分析工作还难以提供较完整的模式信息和设计指导原则。

(2) 如何开发访问并行性? 开发 I/O 请求的服务并行性主要

有两种方法:由用户明确定义;由并行文件系统自行发掘。上述方法各有优缺点。用户往往不了解底层数据的分布状况,难以提供并行化信息;而在许多访问方式下,如用户访问行为复杂以及请求尺寸小和访问的数据细粒度分布时,并行文件系统又难以自行开发并行性。因此,最佳方案是有机地结合这两种技术。但如何实现是当前的研究难点。

(3) 如何实现优化的并行服务? 优化并行服务的关键是充分利用系统各部分性能,消除瓶颈,同时获得最大带宽和最小延迟。在不同数据分布方式下,现有的服务机制往往因未统筹考虑系统各部分性能,导致新瓶颈部件的出现,如数据细粒度分布时结点间通信很可能成为瓶颈等。因此,在综合考虑全系统各单元特点的基础上,研究合适的并行服务机制是实现优化服务的关键。

在第二章中,我们将具体介绍并分析并行文件系统研究的现状和缺陷。

1.3 主要贡献和创新点

在广泛调研和收集大量资料的基础上,本书详细分析了当前并行文件系统的研究现状。以层次型系统为目标,面向并行科学计算应用,以提高并行 I/O 子系统的实际 I/O 带宽利用率为目的开展了研究工作。本书从应用的 I/O 访问模式分析入手,在透彻分析应用 I/O 需求的基础上,提出了 IDDIO 并行文件系统模型,并进而对其中若干关键技术进行了深入研究。

主要研究工作和创新点包括以下几点:

(1) 深入研究了当前主要的并行文件系统技术,结合当前大规模并行计算机系统的技术水平和发展趋势,分析归纳了他们的优缺点。

(2) 并行文件系统的结构和服务模型同用户应用的访问模式密切相关。但目前的 I/O 访问模式分析大多局限于静态的统计分析,因此缺乏对设计并行文件系统的指导意义。作者通过动、静态相结合的分析技术分析了若干典型并行科学计算应用的 I/O 访问特性,提取了并行科学计算应用的常见 I/O 访问模式,提出了可描述并行文件系统性能主要因素的馅饼模型(PieMod),由此得出了设计和研究高性能并行文件系统的指导原则。

(3) 基于对并行科学应用 I/O 访问模式的分析,作者提出了面向磁盘的集成式并行文件系统服务模型(IDDIO)及相关界面。IDDIO模型克服了现有模型只能对一类请求进行优化的缺陷,它可对同步 I/O 请求和独立 I/O 请求同时进行优化处理,使服务性能获得较大改善。

(4) 在 IDDIO 并行文件系统服务模型框架下,作者进而分别研究了对同步 I/O 请求和独立 I/O 请求的优化算法。针对同步式访问中现有并行 I/O 服务算法在数据细粒度分布时通信消息过多,易造成瓶颈的问题,提出了面向通信的分组并行 I/O 服务算法(CDGIO)。CDGIO 算法充分利用目前磁盘系统和通信系统的特点,采用两级并行,分组打包等策略获得了较好的性能优化。理论分析和模拟结果表明,该算法有效地缓解了现有服务算法中存在的通信瓶颈。

(5) 针对并行科学计算应用中独立 I/O 请求对访问服务延迟十分敏感的特点,我们提出了适度贪婪的 cache 预取一体化算法(PGI)。PGI 算法基于用户应用的 I/O 访问模式,在适当时刻发出适度的数据请求,从而回避了现有贪婪或保守预取算法的缺点。PGI 算法采用了全局性的损失估计技术,综合评价淘汰和减少预取数据块的损失,较当前 cache 预取算法获得了较短的整体服务延迟。同时,针对现有 cache、预取算法不适于并行文件系统环境的缺陷,PGI 算法采用了汇集和负载均衡等技术,实现了较好的I/O

访问性能优化。

(6) 利用上述研究成果,作者设计并实现了 SkyWalk 并行文件系统。SkyWalk 并行文件系统采用了 IDDIO 并行服务模型、CDGIO 并行 I/O 服务算法和 PCI cache 预取算法,向用户提供了丰富灵活的用户接口和高性能的 I/O 服务,并可同时支持远程并行文件传输服务功能。

本书的工作是在并行文件系统研究的前沿领域开展的,其研究成果直接应用于实际系统中,具有重要的理论和现实意义。

1.4 论文结构

本书是对研究工作的总结,共分 7 章。

第一章为绪论。

第二章介绍了当前并行文件系统技术和底层支持技术的发展状况。首先介绍了现代并行计算机系统的主要结构、现代磁盘技术、RAID 系统和并行 I/O 子系统的现状,然后介绍并分析了当前并行文件系统的研究热点和存在的问题。最后简要介绍了目前典型并行文件系统的结构、特点及其优缺点。

分析应用的 I/O 访问模式是设计高性能并行文件系统的前提。在第三章中,我们深入分析了并行科学计算应用的 I/O 访问模式。实现分析了一般情况下并行科学计算应用的主要访问特点,其中包括访问的同步性、独立性和访问方式等。然后通过对若干典型并行科学计算应用 I/O 访问轨迹的分析,提取其运行时的动态 I/O 访问模式,其中包括不同类型的访问时间和空间上的关系等。从而总结出一般的并行科学计算应用常见的 I/O 访问模式,并提出了决定系统 I/O 服务性能主要因素的馅饼模型,并在此基础上提出了设计面向并行科学计算应用的并行文件系统时的原