第32卷 第15期 计 算 机 工 程 2006年8月
Vol.32 № 15 Computer Engineering August 2006
·软件技术与数据库· 文章编号：1000—3428(2006)15—0078—03 文献标识码：A 中图分类号：TP338.6

并行文件系统自适应的文件条带化技术

魏文国[1,2]，陈潮填[2]，谢赞福[2]，陈国华[2]
(1. 华南理工大学计算机科学与工程学院，广州 510641；2. 广东技术师范学院计算机科学系，广州 510665)

摘 要：研究并行文件系统自适应的文件条带(Striping)策略对改进文件访问性能的影响，并开发动态的文件条带分析模型，利用自动访问模式分类和实时文件系统性能数据为文件条带策略选择模糊逻辑规则库，来优化文件访问性能。研究结果表明：当文件系统负载低时，可以尽量将文件分布到所有磁盘上来最小化I/O的反馈时间；反之，在系统负载高时，使文件分布的范围小一些以便最大化文件系统整体的吞吐量。并通过实验给出了请求大小、请求宽度、请求到达率与系统性能的相互关系，实证了自适应规则库的正确性。
关键词：并行文件系统；文件条带化；自适应；模糊控制

Automatic-adaptive File Striping of Parallel File System

WEI Wenguo[1,2], CHEN Chaotian[2], XIE Zanfu[2], CHEN Guohua[2]
(1. School of Computer Science & Engineering, South China University of Technology, Guangzhou 510641;
2. Department of Computer Science, Guangdong Polytechnic Normal University, Guangzhou 510665)

【Abstract】 This paper studies a fuzzy logic rule base for adaptive striping of files across multiple disks, and the rule base is based on an analytical model of disk contention that includes disk physical parameters and file request sizes; based on auto classifying of access patterns and real-time monitored data of file system. As the file system load is low, the rule base stripes files aggressively to minimize response time. At high loads, it stripes less aggressively to maximize aggregate throughput. Experiments results show how do request size, request width, request arrival rate affect I/O performance, and verify the correctness of auto-adaptive rule bases.
【Key words】 Parallel file system; File striping; Automatic adaptive; Fuzzy control

为了支持大规模计算的程序处理海量数据集，下一代文件系统必须将数据条带化分布到成千上百的磁盘上，既需要使用条带技术来减少传输时间，又要进行多个独立的传输。这些复杂的任务需要文件系统智能地作出文件的自适应分布决定。

将底层的文件系统的策略和应用程序的访问模式进行匹配能显著地改进输入/输出的性能。其中关键研究问题包括：

(1)通过磁盘条带分析模型研究超大规模计算系统的I/O性能。

(2)基于请求模式和系统负载来自适应选择数据条带策略。

(3)通过存储多份冗余、条带化的文件数据在磁盘容量和带宽之间取得平衡的技术。

磁盘条带优化研究现状：磁盘条带优化的效果依赖存储系统的配置和工作负载特征。文献[1]指出可扩展的磁盘I/O算法需要弹性的条带参数。文献[2]注意到使用最适合的条带单元的重要性，它们将并行性(parallelism)定义为服务一个请求的磁盘数量；将并发性(concurrency)定义为平均用户请求数。它们证明了高层的并发性需要底层的并行性来减少资源的争用。

磁盘块级别并行化的研究已经较普遍和深入，我们的研究兴趣不在此处，而是研究文件系统逻辑块的并行化如何根据系统的负载动态调整，以便达到最小的访问延迟，或者是最大化文件系统整体的吞吐量。

1 文件条带的分析模型

一个访问请求被一个磁盘服务的模型是 $M/G/1$，一个请求被多个服务器响应不存在一般的分析模型。我们开发了一个基于自适应条带的分析模型。

1.1 服务时间分布

假设磁盘服务时间是寻道时间、旋转延迟、数据传输时间、控制器接口处理时间和一系列的软件处理时间的总和；同时当磁盘分布在网络上时还要将网络处理时间考虑进来。所有这些时间都不是重叠的。为了简化处理，假设磁盘旋转一周需要cs，旋转延迟均匀分布在区间$[0,c]$上，旋转延迟的期望值为$c/2$s、方差为$c^2/12$。假设磁盘内最大距离的寻道时间为f秒，最大网络连接时间假设为ks，同样服从上面的分布。为了简单起见，假设所有磁道上的数据密度都是t块（事实上外磁道包含的数据块多于内磁道）。对每一数据块花费在磁盘控制接口和网络传输的时间是is，最后，因为客户串行地分发子请求，所以假设每个子请求需要hs的服务时间。

1.2 分布条带模型

在该模型中，假设通过网络有 m 个磁盘相连，平均请求大小为 l 块(假设是文件系统的块)，系统的请求到达率为 λ_m，假设是 Poisson 到达，因此到达的时间间隔服从指数分布。假设每个请求划分成 D 个子请求，分布在 D 个磁盘上，每个磁盘的条带大小为 l/D 块，D 称为请求宽度。对给定的请求大小，被访问的磁盘数量依赖于条带单元。条带宽度指特定的文件分布的磁盘数量。

基金项目：广东省高校自然科学研究基金资助项目(Z03060)
作者简介：魏文国(1968—)，男，副教授、博士生，主研方向：集群，计算机网络和高性能计算；陈潮填，教授；谢赞福、陈国华，副教授
收稿日期：2005-08-22 **E-mail：**wgwei@21cn.com

在该模型中，多个无关的子请求可能在磁盘等待队列中。假设请求均匀分布在磁盘中，并且每个磁盘的请求到达率 λ_D 与汇总的到达率的关系如式(1)所示。

$$\lambda_D = \frac{\lambda_m D}{m} \tag{1}$$

因为均匀分布就意味着没有访问热点，所以实际的响应时间要大于理论值。每个磁盘模型遵守 $M/G/1$ 队列模型，则可能的队列网络模型如图 1 所示。

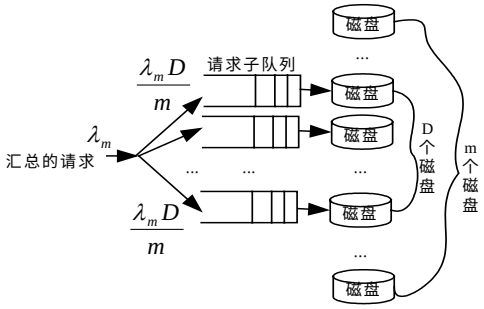

图 1 分布条带模型

在该模型中，子请求被异步响应，首先找出每个子请求的服务响应时间，然后将最大的子请求响应时间作为主请求的响应时间。子请求的服务时间的均值分为 6 个部分：

$$S_D = \frac{f}{2} + \frac{c}{2} + \frac{k}{2} + \frac{c}{t}\frac{l}{D} + i\frac{l}{D} + hD \tag{2}$$

子请求的服务时间的方差为

$$\sigma^2_{S_D} = \frac{(f^2 + c^2 + k^2)}{12} \tag{3}$$

那么，对模型为 $M/G/1$ 的队列延迟时间 W_D 和响应时间 R_D 为

$$W_D = \frac{\lambda_D(\sigma^2_{S_D} + S_D^2)}{2(1-\lambda_D S_D)} \tag{4}$$

$$R_D = W_D + S_D = \frac{\lambda_D(\sigma^2_{S_D} + S_D^2)}{2(1-\lambda_D S_D)} + S_D \tag{5}$$

主请求的响应时间的期望值是D个子请求的响应时间的最大值。式(5)就是所求，即用户的I/O请求响应时间R_D由哪些因素决定，下面给出一个理论模型。

从理论上分析，若客户读/写文件的大小不变，随着请求宽度 D 的增加，即文件的并行性增加，服务时间降低；但是当请求宽度 D 增加到某个阈值之后，因为进程之间的通信和同步开销也随之增加，服务时间反而增加。进一步，大的请求利用磁盘的效率更高，即大请求中单位数据块的响应时间更低。

对固定大小的请求、固定的总请求到达率和固定的磁盘数，增加请求宽度 D，也会增加子请求的到达率。因为服务时间的减少不能补偿请求到达率的增加，在经过某一临界点后，反馈时间快速增加。值得注意的是，小的请求到达率能容忍大的请求宽度，因为小的请求到达率意味着网络比较空闲，可以将文件足够分散存储到多个网络磁盘上来改进 I/O 性能；相反，若请求到达率很大，网络可能已经是系统的性能瓶颈，因此文件的条带分布范围要相对小一些，避免产生更大的网络拥塞。

2 模糊逻辑控制及文件条带化的规则库

经典的控制技术和决策表/树需要深入了解控制领域的相关知识，也依赖于一致的参数空间的划分。而模糊逻辑控制刚好相反，它能处理不准确的系统定义和彼此冲突的目标，利用该领域的常识来控制系统，使得它成为 I/O 系统的合适的解决方法。

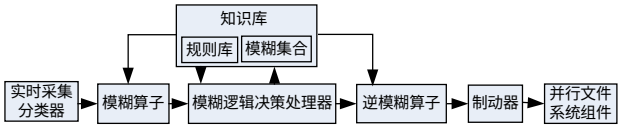

图 2 模糊逻辑控制体系结构

图 2 是我们提出的自适应文件条带的模糊逻辑控制体系结构。各部分的功能如下：

(1)实时采集分类器：搜集系统的重要状态信息（例如CPU、内存的利用率，网络速率等），进行预处理（包括对 I/O 行为进行大致分类）并转发给模糊算子。

(2)模糊算子：采集分类器的输入值是模糊的，该算子结合知识库的信息将输入值转换成模糊代表值。即将模糊的输入值标准化、合并和归类表示成知识库中的一个元素。

(3)知识库：包含模糊规则集和模糊集合，属于静态信息。一个复杂的、自适应的系统能根据系统的输出信息来自动调整模糊集和修改规则，使系统具有自学习的功能。

(4)模糊逻辑决策处理器：该处理器对模糊输入执行知识库的所有规则，结果产生一个新的模糊集元素。

(5)逆模糊算子：产生一个标量值输出，并对标量值进行后期处理，将它们转换成有用的控制信息并分发给制动器。

(6)制动器：根据逆模糊算子产生的标量值来动态地控制并行文件系统组件的行为。

根据有关条带分析模型的参数研究的结论，我们的模糊控制系统按照下面描述的方法，能自适应地给出最佳的条带单元。

假设有模糊变量请求到达率（RequestRate）和请求宽度(RequestWidth)的模糊值如图 3 所示。

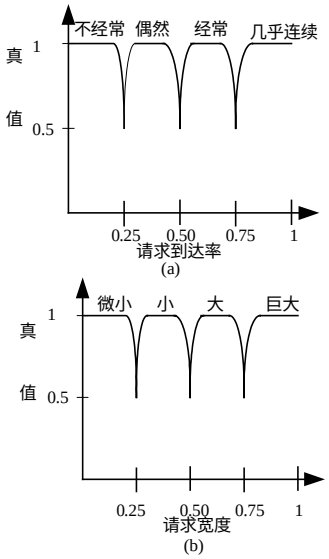

图 3 请求到达率和请求宽度的模糊值

请求到达率有“不经常”、“偶然”、“经常”和“几乎连续”4 个级别，为了使得该规则库能移植到各种系统，将请求到达率和请求宽度都标准化到区间[0，1]。请求宽度为 1 表示所有的设备都被使用。因为三角函数和阶梯函数可以快速地模糊推导，所以只使用它们（本图例使用近似阶梯函数

曲线）。正如前面所说，当请求到达率增加时，独立请求的并行级别降低，在请求宽度和请求到达率之间建立了一个类似“反比例”的关系，这里的“反比例”是近似的、而不是数学意义上严格的反比例关系。为了利用多个磁盘，请求大小必须足够大，这种关系可以用如下的模糊规则表示：

IF(网络性能＝“低”or 磁盘性能＝“高”or 文件并行性＝“高”) THEN 请求宽度＝“小”

IF(网络性能＝“高”or 磁盘性能＝“低”or 文件并行性＝“低”) THEN 请求宽度＝“巨大”

IF(请求大小＝“微小”) THEN

IF(请求到达率＝“不经常”)THEN 请求宽度＝“小”

IF(请求到达率＝“偶然”) THEN 请求宽度＝“微小”

IF(请求到达率＝“经常”) THEN 请求宽度＝“微小”

IF(请求到达率＝“几乎连续”)THEN 请求宽度＝“微小”

ELSE IF(请求大小＝“小”) THEN

……

ELSE IF(请求大小＝“巨大”) THEN

IF(请求到达率＝“不经常”)THEN 请求宽度＝“巨大”

IF(请求到达率＝“偶然”) THEN 请求宽度＝“大”

IF(请求到达率＝“经常”) THEN 请求宽度＝“小”

IF(请求到达率＝“几乎连续”)THEN 请求宽度＝“微小”

END IF

IF(请求到达率＝“不经常”)THEN 文件复制时间＝“在线”

IF (请求到达率＝“经常”)THEN 文件复制时间＝“离线”

该规则库定义的关系可被用于自适应地决定请求宽度。并最终决定条带单元的大小：

条带单元＝平均请求大小/请求宽度 (6)

根据上面的模糊集和模糊规则可以得到：增加请求到达率意味着小的条带数和大的条带单元；当请求到达率固定时，请求大小越大，需要的请求宽度越大，意味着并行性增加。

实时采集和分类器收集系统的性能数据并分类和预测I/O访问模式。以开放源代码的Ganglia[3]为基础设计和实现了一个具有更高可靠性的集群簇/网格实时数据监控与采集系统，该系统能抵御汇集节点和与之相连的线路的失效[4]。

3 实验

为了揭示各种参数对I/O系统性能的影响和检验我们提出的模型的效果，对磁盘仿真工具diskSim[5]进行了局部修改，作为并行文件系统的测试平台。在仿真实验中使用与Western Digital公司型号为WDE4360 的SCSI磁盘近似的参数做了仿真实验。仿真的磁盘参数如下：磁盘的旋转延迟是8.4ms，最大寻道时间是 18ms，每磁道有 22 个数据块，每块为 4kB；并进一步假设仿真的并行文件系统有 256 个磁盘，每个读请求的平均大小为 1 024 个数据块，网络连接建立时间为 30ms，根据其他研究者的经验，假设客户端的每个子请求需要花费 1ms[6]。

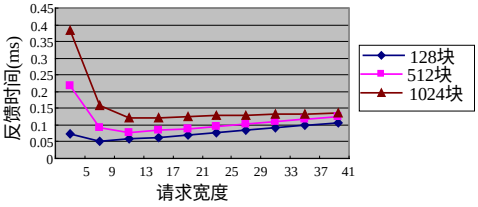

图 4 请求宽度对性能的影响

为了测试请求宽度对并行文件系统性能的影响，假设请求到达率为 30/s 个请求。在 diskSim 上仿真基于网络的 RAID 磁盘，测试结果如图 4 所示，图 4 表示以不同的请求宽度分别读大小为 128 块、512 块和 1 024 块的文件的反馈时间。从图中可以看出：若客户读文件的大小不变，随着请求宽度的增加，服务时间降低；但是当请求宽度增加到某个值之后，服务时间增加。进一步，大的请求利用磁盘的效率更高，即大请求中单位数据块的反馈时间更低。

为了测试请求到达率与最佳条带宽度的关系，我们对固定的请求到达率，以不同的条带宽度写文件，测试给出相应的反馈时间，找出其中最小的反馈时间所对应的条带宽度，就是该请求到达率的最佳条带宽度。测试结果如图 5 所示，从图 5 可以看出：请求到达率越大，最佳条带宽度越小；对相同的请求到达率，请求的文件大小越大，最佳条带宽度越大；并且小的请求到达率能容忍大的请求宽度，文件系统这种自适应的改变能取得最大的性能或者吞吐量。

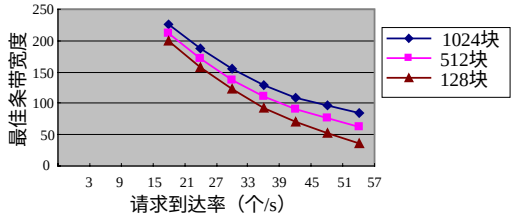

图 5 请求到达率与最佳条带宽度的关系

4 小结

我们研究了面向性能的文件条带分布策略减少 I/O 访问时间的方法，包括实时性能数据采集、自动访问模式分类和模糊逻辑控制来选择和设置弹性的并行文件条带化策略；如何根据请求的大小、一个程序内和多个程序之间的请求的并发性和文件访问模式来动态地改变数据分布在多个磁盘的方式，为了探究磁盘访问的长期趋势，我们构造了分析模型和模糊逻辑库来捕捉访问模式和成千上百个存储设备的关系。通过实验给出了请求大小、请求宽度、请求到达率与系统性能的相互关系，实证了自适应规则库的正确性。该研究结果对真实并行文件系统的开发(如性能的预测，不同的算法取舍、参数调整)都有较大的指导意义。

参考文献

1 Gradecki J D, Ra I. An Adaptive-learning Distributed File System[C]. Proceedings of the 8th International Conference of Knowledge-based Intelligent Information and Engineering Systems, 2004: 637-646.

2 Carballeira G F, Calderon A, Carretero J, et al. The Design of the Expand Parallel File System[J]. International Journal of High Performance Computing Applicatins, 2003, 17(1): 21-37.

3 Massie M L, et al. The Ganglia Distributed Monitoring System: Design, Implementation and Experience [EB/OL]. http://ganglia.sourceforge.net/talks/parallel_computing/ganglia-twocol.pdf, 2005-04.

4 Wei Wenguo, Dong Shoubin, Zhang Ling, et al. An Improved Ganglia-like Clusters Monitoring System[C]. Proceedings of the 2nd International Workshop on Grid and Cooperative Computing, Shanghai, 2003.

5 The DiskSim Simulation Environment[EB/OL]. www.pdl.cmu.edu/DiskSim/, 2005-09.

6 Simitci H, Reed D A. Adaptive Disk Striping for Parallel Input/Output[C]. Proceedings of the 7th Goddard Conference on Mass Storage Systems and Technologies, San Diego, CA, 1999-03-15: 88-102.