CN43-1258/TP
ISSN 1007-130X

计算机工程与科学
COMPUTER ENGINEERING & SCIENCE

2004年第26卷 第3期
Vol. 26, No. 3, 2004

文章编号: 1007-130X(2004)03-0025-04

网络存储环境下的I/O负载*

The I/O Workload in Network-Attached Storage Environments

王 芳, 冯 丹, 张江陵

WANG Fang, FENG Dan, ZHANG Jiang-ling

(华中科技大学计算机学院外存储重点实验室, 湖北 武汉 430074)

(National Laboratory of Storage Systems, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China)

摘 要: 网络存储改变主机系统与存储设备间的连接关系, 利用存储设备的计算能力直接向网络用户提供存储服务。这种结构的改变使系统存储处理模式相应发生变化, 导致存储设备I/O负载特征发生变化。本文提出存储系统负载不仅与应用环境相关, 还与系统存储处理模式相关的观点, 并据此分析了传统存储系统和网络存储系统的I/O负载特征。

Abstract: Network-attached storage devices change the connection relationship with the host system, and provide direct storage services to network users using its own computing capability. This structural change makes the storage disposal mode change, thus makes the workload characteristics of storage device change. This paper considers that storage workload relates not only to application environments, but also to the storage disposal mode, and analyzes the characteristics of the I/O workloads of both traditional and network-attached storage systems.

关键词: I/O负载; 网络存储; 存储处理模式

Key words: I/O workload; network-attached storage; storage disposal mode

中图分类号: TP333 文献标识码: A

1 引言

存储系统特征与其I/O负载特征相匹配时, 系统性能才能实现最优[1,2]。网络存储系统改变了与主机的连接方式, 在网络存储系统与用户负载间出现多重接口, 这使得加载其上的I/O负载特征发生变化, 因而必须调整存储系统的调度策略以适应新的负载特征。本文以大数据流量为特征的文件共享服务为例, 分析了存储系统连接结构和处理模式的改变对I/O负载特征的影响。

2 系统工作模式对I/O负载的影响

网络存储系统的体系结构示意见图1。存储系统负载特征不仅与应用环境相关, 还与系统的存储处理模式紧密相关。若用 $A(x)$ 表示由应用本身决定的负载特征函数, 用 $P(x)$ 表示存储处

* 收稿日期: 2003-01-03; 修订日期: 2003-06-27
基金项目: 国家自然科学基金资助项目(60273074)
作者简介: 王芳(1972-), 女, 河北深县人, 博士, 研究方向为计算机系统结构、计算机网络、网络磁盘阵列、计算机存储系统等; 冯丹, 博士, 教授, 研究方向为计算机系统结构、并行存储系统、容错等; 张江陵, 教授, 博士生导师, 研究方向为磁记录理论及外存储系统。
通讯地址: 430074 湖北省武汉市华中科技大学计算机学院外存储重点实验室; Tel: (027)87542463; E-mail: wangf72@hotmail.com
Address: National Laboratory of Storage Systems, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, P.R.China

理方式对负载产生的影响, 则最终施加在存储系统上的负载特征函数为 $Z(x) = P(A(x))$。过去的研究中通常将 $A(x)$ 直接作为存储系统最终负载特征, 而忽略系统处理行为 $P(x)$ 对它的变异作用。这种忽略与以往存储设备功能简单、系统存储处理方式单一、应用环境单纯有关, 且系统对用户 I/O 请求处理更多的是一种复述传递功能。但是, 随着新的计算机处理技术的引入、应用环境的复杂化, 以及存储设备功能的复杂化, 系统 I/O 处理模式对负载特征产生的影响日渐明显[3]。

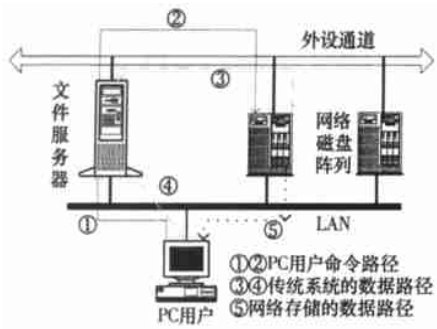

图 1　网络存储系统的体系结构与数据路径

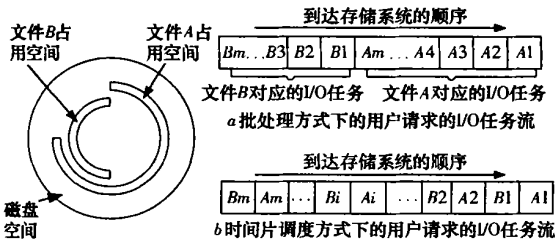

图 2　不同系统处理模式对存储系统实际负载特征的影响

图 2 示出了两个用户分别请求读取长度均为 L_F 的大文件 A 和 B 的例子。假设两文件各占有一连续磁盘空间, 且相隔较远。A 的起始磁道号为 S_{a1}, 跨 S_A 个磁道; B 的起始磁道号为 S_{b1}, 跨 S_B 个磁道。设 A_i、B_j 分别为经过文件系统和 I/O 驱动程序处理后产生的 I/O 任务量, 则文件 A 派生任务为 $\{A_1, A_2, \cdots, A_m\}$, 文件 B 派生任务为 $\{B_1, B_2, \cdots, B_m\}$, $m > 2$, A_i、B_j 所在磁道号分别为 S_{ai} 和 S_{bj}。文件 A 或 B 的第 i 次 I/O 任务的执行时间均可表示为 $t_i = S_i + R_i + B_{io}/r_d + B_{io}/r_s$, 其中, B_{io} 为一次 I/O 任务的数据量, 即 A_i 或 B_i; S_i 和 R_i 分别为寻道时间和旋转延迟; r_d 和 r_s 为磁盘的数据传输速率和通道数据传输速率。

当采用 FCFS(First Come First Service, 简称 FCFS)的作业级单道批处理调度算法时, 系统按请求到达顺序调度。从存储系统接收端看, 其 I/O 负载为两组连续长数据请求 $\{A_1, A_2, \cdots, A_m\}$ 和 $\{B_1, B_2, \cdots, B_m\}$, 如图 1a 所示。两组请求间有一次寻道延时, 组内各 I/O 任务因数据的空间连续性, 寻道时间和旋转延时减到最小, I/O 执行效率最高。可求得服务两组请求的总 I/O 服务时间 T_1 为:

$$T_1 = S_{seek_A} + S_A \cdot S_{\min} + (S_{a_1} - S_{b_1}) \cdot S_{\min} + S_B \cdot S_{\min} + \sum_{i=1}^{n} t_{r_i} + \sum_{j=1}^{m} t_{r_j} + 2\left(\frac{L_F}{r_d} + \frac{L_F}{r_s}\right) \quad (1)$$

其中, 第一项为执行读 A_1 的寻道时间, 第二项为连续执行 $\{A_2, \cdots, A_m\}$ 的总寻道时间, 第三项为读 B_1 的寻道时间, 第四项为连续执行 $\{B_2, \cdots, B_m\}$ 的总寻道时间, $S_{\min}$ 为相邻柱面寻道时间, 其余项为旋转时延和文件传输时间。

当系统在进程级对两组请求按时间片轮转调度时, 从存储系统接收端看, 其实际负载为一组连续请求流 $\{A_1, B_1, A_2, B_2, \cdots, A_m, B_m\}$, 如图 1b 所示。相邻请求 A_i、B_j 间存在较大的寻道和旋转延时, 由于磁头不断往复运动, I/O 执行效率不高。其总 I/O 服务时间 T_2 为:

$$T_2 = S_{seek_A} + \sum_{j=1}^{2n-1}(S_{a_j} - S_{b_j}) \cdot S_{\min} + \sum_{j=1}^{2m} R_j + 2\left(\frac{L_F}{r_d} + \frac{L_F}{r_s}\right) \quad (2)$$

其中, 第一项为执行读 A_1 的寻道时间, 第二项为连续执行 $\{B_1, A_2, B_2, \cdots, A_m, B_m\}$ 的寻道时间, 其余为旋转时延和文件传输时间。

统计平均状态下, $T_2 - T_1 = \sum_{j=2}^{2n-1}(|S_{a_j} - S_{b_j}|) \cdot S_{\min} - (S_A + S_B) \cdot S_{\min}$, 文件 A、B 相隔较远时, 有 $|S_{a_i} - S_{b_i}| > S_A$, $|S_{a_i} - S_{b_i}| > S_B$, 则 T_2 必大于 T_1。在一般情况下, 该不等式也会成立, 即按前一种方式执行的 I/O 效率高于后者。这表明, 相同用户请求经不同的系统处理后, 存储系统实际 I/O 负载特性会不同, 从而使获得的服务性能不同。因此, 在存储系统设计中, 应同时考虑用户行为与系统行为, 以决定存储系统内的各级调度策略, 使系统整体性能达到最优。

3　传统的 I/O 负载特征

传统存储系统 I/O 处理模型如图 3 所示。设

LAN 中同时有 N_c 个用户对不同文件的读写请求, 这些请求都要经过文件服务器预处理, 文件系统建立在由阵列形成的逻辑磁盘上。理想状态下, 同一文件占据逻辑磁盘的一片连续空间。对于单一用户文件读写请求, 经服务器应用软件和文件系统处理后, 形成一系列面向逻辑磁盘的 I/O 任务, 其操作数据大小, 即“访问块” L_{io} 的大小由处理软件、文件系统及 I/O 驱动程序决定。I/O 任务经外设通道到达阵列后再由阵列按分块大小 B 分派到阵列各磁盘上。各磁盘 I/O 负载以顺序请求为特征。

当多用户同时发出文件请求时, 操作系统通常按分时轮转策略调度。用户进程发出一次 I/O 请求后, 就进入 I/O 等待队列, 转为睡眠状态, 系统选择就绪态进程队列中头一个进程运行。因此, 当系统中 N_c 个进程分别处理不同用户文件请求时, 实时加载在存储系统的 I/O 任务队列为 $\{F_{1,1}, F_{2,1}, F_{3,1}, \cdots, F_{1,2}, F_{2,2}, \cdots, F_{i,j}, F_{(i+1),j}, \cdots\}$, $i = 1, 2, \cdots, N_c$, $F_{i,j}$ 表示第 i 个用户请求文件的第 j 个访问块。根据前面同一文件连续放置的假设, $F_{i,j}$ 与 $F_{(i+1),j}$ 间存在明显跨区域访问, 即随机访问。由于服务器预处理所导致的这种“混洗”作用, 使各用户请求本身所包含的顺序性, 在到达存储系统负载接收界面时已成为一组随机 I/O 任务, 性能因磁盘驱动器的定位和旋转延迟而很差。为提高存储系统效率, 可采用增加 I/O 任务数据量, 即加大访问块 L_{io} 的方式, 使磁盘在一次定位、旋转延迟后, 执行更多的 I/O 数据访问, 同时也减少总的 I/O 任务次数。

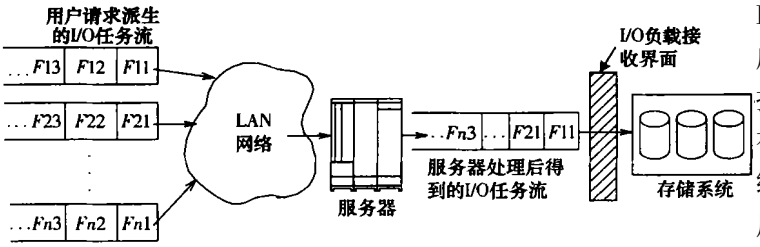

图 3 传统系统中的负载处理模型

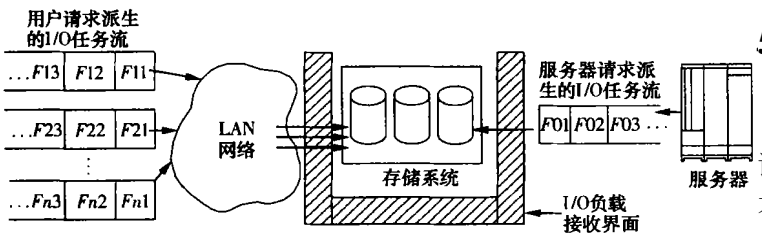

图 4 网络存储系统中负载的处理模型

4 网络存储的 I/O 负载特征

网络存储系统负载处理模型如图 4 所示。网络存储设备的网络接口和外设接口共同构成了它的 I/O 负载接收界面。在处理网络用户文件请求时, 先由服务器获取文件分布的地址表信息, 再在一定的管理权限下, 由用户按地址表信息直接在存储设备上存取文件数据。这时, 存储设备通过网络接口直接“看到”多个完整的文件 I/O 任务队列, 每一任务队列内各 I/O 任务是全顺序的, 或近似全顺序的, 请求并发度由并发的用户数 N_c 决定。磁盘阵列网络接口的存在, 使用户请求数据的 I/O 空间顺序性被完整地保留下来, 并直接呈现在存储设备的负载接收界面上。

当存储系统依次服务不同用户请求时, 可以完全利用用户请求数据的顺序性, 使总寻道延迟和旋转延迟最小, 存储设备执行效率最高, 但用户的实时响应性能则较差。当存储系统采用轮流服务策略时, 每次只对用户请求的一部分服务, 若每次服务的 I/O 任务量等于访问块长度 L_{io}, 则存储系统的性能与传统方式的性能相似, 即在存储系统内部进行了类似于传统服务器的“混洗”功能。如何利用用户请求数据顺序, 提高存储服务效率, 并保证一定的用户响应性能, 是网络存储设备中用户请求调度算法应权衡的问题。从图 4 可见, 网络存储系统除接收用户 I/O 任务流外, 还接收服务器 I/O 任务流。其来源有两类: 一是在处理用户请求过程中, 为获取文件空间分布而派生的 I/O 任务; 二是服务器自身访问文件形成的 I/O 任务。前者请求次数少, 且数据量较少, 加之服务器端运行的高效缓存机制, 在内存中保存了重要的文件系统结构数据, 实际派生的 I/O 任务更少。后者则可将它视为普通用户文件 I/O 任务来处理。

5 结束语

存储系统的 I/O 负载特征对其请求调度策略起着决定作用。I/O 负载特征不仅受到应用环境的影响, 还受到系统处理模式的影响。网络存储系统因其特有的存储体系结构, 使系统的 I/O 处

理模式发生新变化，在设计网络存储系统的调度策略时，应充分考虑各种影响因素，才能在获得较好的整体性能。

参考文献：

[1] A L Narasimha Reddy, Prithviraj Bantrjee. An Evaluation of Multiple-Disk I/O Systems[J]. IEEE Trans on Computers, 1989, 38(12): 1680-1690.

[2] P Chen, D Patterson. Storage Performance-Metrics and Benchmarks[J]. Proc of the IEEE, 1993, 81(8): 1151-1161.

[3] 王芳. 网络磁盘阵列系统的研究:[博士学位论文][D]. 华中科技大学, 2001.

(上接第 3 页)

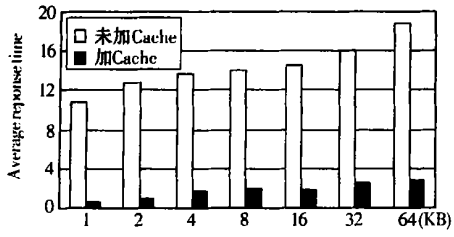

图 4 响应时间对比测试

3 可以看出性能有明显的提升，其提高幅度最高可达一倍。这不仅是由于 Cache 对数据的缓存作用，而且是由于 Cache 的存在减少了许多需在网络传输的 SCSI 命令和状态。图 4 为平均响应时间的对比结果。可以看到，平均响应时间大幅下降，都在 3ms 以内。

5 结束语

目前，基于 IP 技术的存储网将广泛应用到存储环境中。为改善由于TCP/IP 协议栈的执行开销带来的性能问题，本文提出并实现了一种新的缓存结构，它能很好地提高 iSCSI 协议的性能。最后的测试结果证明，对数传率、I/O 请求的平均响应时间等指标有大幅提高。

参考文献：

[1] B Phillips. Have Storage Area Networks Come of Age? [J]. IEEE Computer, 1998, 31(7): 10-12.

[2] Xubin He, Qing Yang, Ming Zhang. A Caching Strategy to Improve iSCSI Performance[A]. IEEE Annual Conf on Local Computer Networks[C]. 2002. 6-8.

[3] Peter M Chen, Wee Teck Ng. The Rio File Cache: Surviving Operating System Crashes[A]. Proc of the 7th Int'l Conf on ASPLOS[C]. 1996.

[4] C Gray, D Cheriton. Leases: An Efficient Fault-Tolerant Mechanism for Distributed File Cache Consistency[C]. Proc of the 20th ACM Symp on Operating System Principles[C]. 1989. 202-210.

[5] Randal C Burns, Robert M Rees. Safe Caching in a Distributed File System for Network Attached Storage[A]. Proc Int'l Parallel and Distributed Processing Symp(IPDPS)[C]. 2000. 155-162.

(上接第 10 页)

得其功能。如果系统要适合多种网络和系统环境，只需要修改相应的网络通信接口。因此，用 COM 技术来实现 H.323 系统中的网守对于软件实现多媒体会议系统具有很重要的意义。

参考文献：

[1] ITU-T Recommendation H.323v1[EB/OL]. http://www.packetizer.com/iptel/h323/drafts/H323v1.zip, 1996-11.

[2] ITU-T Recommendation H.323v2[EB/OL]. http://standard.pictel.com/ftp/avc-site/till_0012/9801_Gen/H323v2_D.zip, 1998-01.

[3] ITU-T Recommendation H.323v3[EB/OL]. http://standard.pictel.com/ftp/avc-site/till_0012/9909_Gen/h323v3_decided_991022.zip, 1999-09.

[4] ITU-T Recommendation H.323v4[EB/OL]. http://standard.pictel.com/ftp/avc-site/till_0012/0011_Gen/H323v4-final_010206.zip, 2000-11.

[5] ITU-T Recommendation H.225.0v4[EB/OL]. http://standard.pictel.com/ftp/avc-site/till_0012/0011_Gen/H2250v4-final_010317.zip, 2000-12.

[6] 潘爱民. COM 原理与应用[M]. 北京: 清华大学出版社, 1999.

[7] Martin Gudgin. 宋亚男译. IDL 精髓 Essential IDL[M]. 北京: 中国电力出版社, 2002.

(上接第 17 页)

参考文献：

[1] 卢昱, 林琪. 网络安全技术[M]. 北京: 中国物质出版社, 2001.

[2] 涂序彦. 大系统控制论[M]. 北京: 国防工业出版社, 2000.

[3] 郑大钟, 赵千川. 离散事件动态系统[M]. 北京: 清华大学出版社, 2001.

[4] 王万良. 自动控制原理[M]. 北京: 科学出版社, 2001.

[5] 卢昱. 网络控制论浅叙[J]. 装备指挥技术学院学报, 2002, 13(6): 60-64.

[6] Richard C Dorf, Robert H Bishop. 谢红卫, 等译. 现代控制系统[M]. 北京: 高等教育出版社, 2001.