简单树结构存储系统数据分块标准及其证明

周 可 吉永光 冯 丹 王 芳

(华中科技大学 计算机科学与技术学院, 湖北 武汉 430074)

摘要: 为解决传统磁盘阵列技术受系统总线与设备通道制约的问题,采用树结构和单元控制器构建了新型的存储系统.在树结构存储系统的基础上,通过对简单树结构存储系统数据分块方式的定义、推论,以及冲突系数的证明,给出了简单树结构存储系统的数据分块标准.结果证明:简单树结构存储系统数据分块标准能最大程度降低存储系统有效冲突系数,提升存储系统性能.

关 键 词: 树结构; 数据分块; 有效冲突系数; 分块标准

中图分类号: TP333.3 文献标识码: A 文章编号: 1671-4512(2007)03-0053-03

Data striping criterion for simple tree structure storage system and its verification

Zhou Ke Ji Yongguang Feng Dan Wang Fang

(College of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract: In order to resolve the restricted system bus and the equipment channel based on the traditional redundant array of independent disks (RAID), a new simple tree structure storage system (STSSS) was constructed by RAID with unit controllers. The definitions and deductions are proposed to study STSSS data striping, the efficient conflict coefficient (ECC) is given and the criterion for STSSS data striping is proved. The conclusion shows that the criterion of STSSS data striping can reduce the efficient conflict coefficient and improve the performance of the storage system.

Key words: tree structure; data striping; efficient conflict coefficient; striping criterion

自20世纪80年代提出磁盘阵列以来,已诞生多种类似的新型存储系统,如AutoRAID[1], Massive Arrays of Idle Disks[2]和RAIDn等.上述几种存储系统虽然在"活跃"数据与"非活跃"数据处理、后台数据处理、数据纠错方式等方面有所创新,但系统结构与数据组织方式仍沿用传统磁盘阵列技术.有的存储系统只是在磁盘阵列基础上采用大量的缓存技术,其作用是加速读操作和缓存小写[3]或者在存储系统中添加虚拟层来实现缓存功能[4].目前流行的网络存储技术[5~7]在数据组织方式上还是沿用传统磁盘阵列数据分块技术[8,9].随着底层存储设备与各种存储技术的发展,存储系统并行性研究遇到了新的难题.为了解决底层设备通道受系统总线制约的问题,本研究试图采用树结构来构建新型存储系统,研究了树结构存储系统的数据分块方法及采用有效冲突系数来判断数据分块策略的优劣,并给予证明.

1 简单树结构存储系统结构及其数据分块方法

树结构存储系统是一种新型存储系统,它具有并行度高、系统扩展性强、兼容多种标准设备的

收稿日期: 2005-09-22.

作者简介: 周 可(1974-),男,副教授;武汉,华中科技大学计算机科学与技术学院(430074).

E-mail: raidkick@263.net

基金项目: 国家重大基础研究前期研究专项资助项目(2004CCA07400);国家自然科学基金资助项目(60503059);武汉市晨光计划资助项目(20055003059-1).

优点[10]. 为消除系统总线瓶颈, 在磁盘阵列技术基础上, 采用树型结构和单元控制器便可构成简单树结构存储系统(图 1). 简单树结构存储系统是指仅在最底层节点才挂接存储设备的树结构存储系统. 系统的根节点是主机, 通过两条 SCSI 通道连接下一层的两个分支, 并由虚拟设备驱动构成一个统一的存储空间. 系统的中间节点是单元控制器, 所有的单元控制器都具有一条向上连接的上通道和两条向下连接的下通道, 上通道驱动由 SCSI Slave 驱动程序完成, 下通道驱动由 SCSI Master 驱动程序完成. 磁盘构成系统的叶子节点. 简单树结构存储系统的数据组织应该充分发挥自身结构所带来的优势, 达到底层设备在数据传输时的最大并行度. 如图 1 所示, 数据块 1~ 8 的存放顺序依次为磁盘 1, 5, 3, 7, 2, 6, 4, 8. 这样做的目的是为了不论主机的数据请求有多大, 都能充分发挥多通道的并行传输能力.

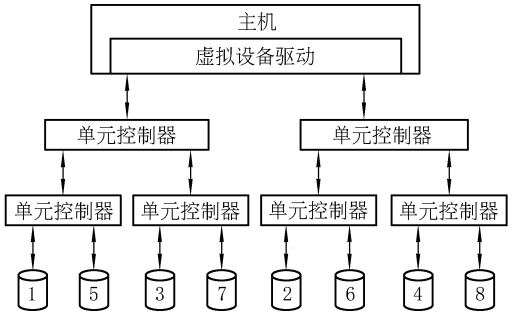

图 1 简单树结构存储系统结构及数据分块

2 相关定义及推论

为了研究简单树结构存储系统数据分块和它的访问路径, 对涉及的概念定义如下:

定义 1 从主机到存放某一数据块的物理位置所经过的所有通道号组成的集合, 称为该数据块的访问通道, 访问通道中的通道号称为该访问通道的成员. 如图 1 所示, 从主机到数据块 1 经过的所有通道号为 $0\vec{}00\vec{}000$, 则数据块 1 的访问通道是(0, 00, 000); 从主机到数据块 2 经过的所有通道号为 $1\vec{}10\vec{}100$, 则数据块 2 的访问通道是(1, 10, 100).

定义 2 访问通道的成员按访问顺序排列组成的二进制向量, 称为数据块的访问路径. 例如, 数据块 1 的访问路径是 000000, 数据块 2 的访问路径是 110100.

定义 3 若两个访问通道的交集不为空, 则访问通道冲突; 否则不冲突. 如访问通道(0, 00, 000)和(0, 01, 010)冲突, 访问通道(0, 00, 000)和(1, 10, 100)不冲突.

以下定义和推论只适合于二叉树.

定义 4 在从主机到某一数据块存放的物理位置经过的所有通道号中, 取最末一位组成的二进制向量, 称为该数据块的简单访问路径. 如图 1 所示, 数据块 1 的简单访问路径是 000, 数据块 2 的访问路径是 100.

定义 5 所有节点, 包括根节点、中间节点、叶节点的最大层次, 称为简单树结构存储系统的深度. 简单树结构存储系统的深度为 4(图 1).

推论 1 若简单树结构存储系统的深度为 n, 则访问路径位数为 $\sum_{i=1}^{n-1} i$, 简单访问路径位数为 $n-1$.

证明 从简单树结构存储系统结构不难得出, 深度为 n 的简单树结构存储系统的访问通道的成员数为 $n-1$, 成员位数由高到低分别为 1, 2, …, $n-1$. 因此, 根据定义 2 和定义 4, 访问路径位数为 $\sum_{i=1}^{n-1} i$, 简单访问路径位数为 $n-1$. 见图 1, 访问路径位数为 1+ 2+ 3= 6, 简单访问路径位数为 3.

定义 6 两个简单访问向量从高到低按位进行同或操作, 若结果为 1, 则继续下一位的计算; 若结果为 0, 则低位取 0. 这个过程称为简单访问向量的冲突计算, 计算的结果成为简单访问向量的冲突计算结果.

推论 2 若简单树结构存储系统的深度为 n, 则简单访问向量的冲突计算结果有 n 种.

证明 根据定义 6, 简单访问向量的冲突计算结果位数为 $n-1$ 位, 若高 m 位($0 \leqslant m \leqslant n$)为 1, 则低 $n-m$ 位为 0. 因此, 简单访问向量的冲突计算结果有 n 种. 如图 1 所示, 简单树结构存储系统深度为 4, 简单访问向量的冲突计算结果共 4 种, 分别是 000, 100, 110, 111.

推论 3 若简单访问向量的冲突计算结果为全 0, 则访问通道不冲突; 若简单访问向量的冲突计算结果为全 1, 则数据块处于同一叶子节点.

证明 若简单访问向量的冲突计算结果为全 0, 则访问通道成员无相同成员, 访问通道不冲突; 若简单访问向量的冲突计算结果为全 1, 则访问通道成员全部相同, 数据块处于同一叶子节点.

定义 7 简单访问向量的冲突计算结果中包含 1 的数目, 称为冲突系数.

推论 4 冲突系数为 0~ $n-1$ 间的自然数.

证明 简单访问向量的冲突计算结果位数为 $n-1$ 位, 因此, 根据冲突系数的定义, 冲突系数为 0~ $n-1$ 之间的自然数.

3　有效冲突系数证明

计算多个简单访问向量的有效冲突系数的步骤如下:

第 1 步　计算多个简单访问向量中所有的冲突系数, 选取最大的冲突系数.

第 2 步　选取得到最大冲突系数的任意一个简单访问向量作为标准简单访问向量, 计算其他简单访问向量与标准简单访问向量的冲突系数并求和, 即有效冲突系数.

推论 5　为计算方便, 假设简单树结构存储系统的根节点中虚拟设备驱动延时为 0, 通道的数据传输延时相等, 磁盘数据访问延时等于单个通道的数据传输延时, 单元控制器的处理延时为 0, 则分块数据的并行访问延时与分块数据的有效冲突系数成线性关系.

证明　采用归纳法证明.

a　假设简单树结构存储系统深度为 n, 一个通道的数据传输延时为 t_c, 则一个数据块的访问延时 $t_1 = nt_c$. 考虑两个数据块的并行传输过程, 若冲突系数为 x, 则表示两个数据块在 n 个通道的数据传输时发生冲突. 在每一个发生冲突的通道, 第二块数据必须等待延时 t_c, 即等待第一块数据传输完毕才能开始传输, 并行传输延时为 $2t_c$. 在不发生冲突时, 两块数据分别通过不同的通道传输, 并行传输延时为 t_c. 因此, 两块数据并行访问延时 $t_2 = t_1 + xt_c$.

b　若 m 块数据的并行访问延时为 t_m, 考虑 $m+1$ 块数据的并行访问延时, 假设 m 块数据的有效冲突系数为 A, $m+1$ 块数据的有效冲突系数为 $A+y$, 则 $m+1$ 块数据的并行访问延时 $t_{m+1} = t_m + yt_c$.

综上所述, 分块数据的并行访问延时随有效冲突系数的增长而线性增长. 证明完毕.

根据推论 5, 简单树结构存储系统数据分块的标准是尽量使分块数据的有效冲突系数最小.

4　有效冲突系数的应用

如图 1 所示的数据分块, 数据块 1 的简单访问路径为 000, 数据块 2 的访问路径为 100, 有效冲突系数为 0, 因此, 并行访问数据块 1 和数据块 2 的延时最小. 无论数据块 1~4 怎么放置, 其有效冲突系数最小为 1, 图 1 所示的分块方法中数据块 1~4 满足有效冲突系数最小. 再考虑数据块 1~8, 无论怎么放置, 其有效冲突系数最小为 4. 若分块数大于 8, 则有效冲突系数不小于 7. 可以看出, 图 1 所示的数据分块是一种最优的数据分块方式, 当然它不是惟一的方式.

参　考　文　献

[1] Wilkes J, Golding R. The HP AutoRAID hierarchical storage system [J]. ACM Trans on Computer Systems, 1996, 14(1): 108-136.

[2] Colarelli D, Grunwald D. Massive arrays of idle disks for storage archives [C] // Proceedings of the 15th High Performance Networking and Computing Conference. Los Alamitos: IEEE Computer Society Press, 2002: 1-11.

[3] 冯　丹, 熊建刚. 磁盘阵列 cache 数据一致性的研究与实现[J]. 华中科技大学学报: 自然科学版, 2005, 33(10): 70-72.

[4] 王　芳, 曾令仿, 冯　丹, 等. 可快速响应的虚拟三级存储系统 RAT[J]. 华中科技大学学报: 自然科学版, 2005, 33(增刊): 138-141.

[5] 李洁琼, 冯　丹. 一种基于网络磁盘阵列的高性能海量存储系统[J]. 小型微型计算机系统, 2006, 27(12): 2 327-2 330.

[6] 王　芳. 网络磁盘阵列系统的研究[D]. 武汉: 华中科技大学计算机科学与技术学院, 2001.

[7] 周　可, 冯　丹, 王　芳. 网络磁盘阵列 I/O 请求并行调度策略[J]. 华中科技大学学报: 自然科学版, 2006, 34(9): 1-3.

[8] 冼曙光, 冯　丹. 磁盘阵列控制软件设计及其实时响应分析[J]. 计算机工程与科学, 2006, 28(7): 109-111.

[9] 田　磊, 冯　丹. 存储区域网中磁盘阵列光纤通道接口的设计与实现[J]. 计算机工程与科学, 2005, 27(7): 106-108.

[10] Zhou Ke, Huang Yongfeng, Feng Dan. Disk tree: a case of parallel storage architecture to improve performance in random access pattern [J]. Chinese Journal of Electronics, 2005, 14(1): 39-44.