一种基于效益-代价均衡的磁带库调度算法*

石 晶, 邢春晓, 周立柱

(清华大学 计算机科学与技术系,北京 100084)

E-mail: shijing@mails.tsinghua.edu.cn; xingcx@tsinghua.edu.cn; dcszlz@tsinghua.edu.cn

http://dbgroup.cs.tsinghua.edu.cn

摘要: 诸如数字图书馆等规模在 10^{12} 字节以上的大型数据库需要在线存取大容量磁带库中的海量数据.由于这些需求都是对海量数据的随机存取,而磁带库的随机存取性能很差,所以,研究有效的磁带库随机 I/O 调度策略和算法是改善磁带库系统性能的重要课题.提出并研究了一种基于效益-代价均衡的调度算法,给出一种有效的效益-代价加权比的估算方法.该算法根据系统的工作负载特点,动态调节调度的效益和代价的加权比,从而改善了磁带库系统在各种负载下的系统性能.研究解决了已有磁带库调度算法的对工作负载敏感的问题,极大改善了调度算法在重负载下的有效性.

关 键 词: 在线磁带库系统;磁带随机 I/O 调度;热数据复制;效益-代价估算

中图法分类号: TP316 文献标识码: A

随着科学技术的发展,各领域的数据呈现爆炸式增长[1],出现了规模 10^{12} 字节(terabyte)以上的电信通话记录数据库、大型数字图书馆,地理、空间及环境数据库和视频音频归档数据库,对这样的海量数据的存储和操作完全依赖磁盘存储系统是不切实际的.一个重要的发展趋势是把诸如磁带这样的大容量存储介质直接用于数据的存储和查询操作,使之成为大型数据库系统的存储结构中的"主动存储层次"[2].为此,首先需要解决数据库系统对磁带库中数据的在线随机存取问题.

由于磁带库是典型的顺序存取设备,其随机存取的性能很差,所以,研究有效的随机 I/O 调度策略和算法是改善磁带库的在线存取效率的关键之一.在这方面已经看到了一些研究成果[3~6].其中,文献[3]讨论了 DLT 驱动器的搜寻定位模型以及基于该模型的非连续磁带数据查询的调度算法,这些研究只涉及磁带库调度中对单条磁带上请求的调度.文献[4]则给出了基于层次的磁带库静态随机调度算法.文献[5]详细讨论了磁带库的随机 I/O 调度问题,给出了一种基于热数据复制的动态调度算法.文献[6]对用于卫星图像数据库的可伸缩磁带归档器(由自行开发的磁带迁移单元连接的多个磁带库构成)进行了性能分析,提出使用热分解和热复制两个模式改善由多个磁带库组成的系统的性能.上述成果给出了改善磁带库系统性能的一些有效方法,主要包括热数据复制、热负载平衡和动态调度等.但这些方法的使用都有一定的局限性.热数据复制和动态调度都属于贪心调度策略,在重负载条件下会使很多不幸的请求处于无限的等待,从而使系统总体性能急剧恶化.热负载平衡方法是基于磁带迁移单元在多个磁带库之间迁移磁带实现负载平衡来改善系统性能的,因磁带迁移单元是自行设计,而非市场产品,所以很难通用.

本文侧重研究单个磁带库在重负载下的调度问题.由于磁带库系统的负载特点与磁盘系统不同,它主要有两类负载条件:(1) 日常存取,这类负载属于较轻负载,通常是存取最近生成的新数据,存取的重复性很强;(2) 突

* 收稿日期: 2001-04-20; 修改日期: 2001-09-05

基金项目: 国家重点基础研究发展规划 973 资助项目(G1999032704)

作者简介: 石晶(1969－),女,辽宁辽阳人,博士生,主要研究领域为海量信息处理,第三级存储设备操作与管理;邢春晓(1967－),男,河南南阳人,博士,副教授,主要研究领域为海量信息处理及其在数字图书馆中的应用;周立柱(1947－),男,江苏连云港人,教授,博士生导师,主要研究领域为数据库,海量信息处理,Web 技术.

发存取,这类负载属于重负载,通常是存取相关的大量数据,存取的重复性较弱.对于第一类负载,采用大磁盘缓存的方法是很有效的.而对于后一种负载,磁盘缓存的作用并不明显,更多地依赖好的在线调度算法来提高系统的性能.本文的研究是针对第二种负载类型的.我们提出一种基于效益-代价均衡的调度算法,该算法在负载较轻时,使系统性能接近于已有的调度算法,而在重负载下则明显好于已有算法.

本文的研究以 Eliant 820 磁带驱动器和 Exabyte 220 磁带库的性能参数为基础,使用磁带库仿真器对算法进行模拟研究.第 1 节详细描述了调度算法的设计思想.第 2 节则重点介绍效益-代价的估算方法.第 3 节介绍了算法的实现步骤.第 4 节给出了算法比较的仿真结果.第 5 节给出小结.

1 算法设计

基于效益-代价均衡的调度算法在设计时主要考虑了如下优化策略:

(1) 尽可能减少磁带的交换次数.由于磁带交换包括磁带反绕、旧磁带卸载、机械手交换新旧磁带、新磁带加载等多个顺序执行的机械运动,其消耗时间在分钟的数量级,所以减少磁带交换次数可以改善系统的性能.

(2) 尽可能顺序存取磁带.由于磁带驱动器的随机定位时间比磁盘高 3~4 个数量级,其随机存取的性能很差.所以,把针对同一磁带的请求排序后进行顺序存取,可以减少搜寻定位距离和定位时间.

(3) 在重负载下考虑所有磁带的公平服务.由于调度算法采用的优化策略基本上都是不公平服务策略,而在重负载情况下就可能累积出越来越多的处于无限等待的请求不能得到及时的服务,使系统的整体性能下降.所以,在优化调度时也需要考虑磁带的均衡服务策略,以保证系统的总体性能和服务质量的改善.

很显然,策略(1)和(2)是贪心策略,追求效益的最大化,而策略(3)是公平策略,侧重考虑调度策略的代价.表面上看,它们在优化目标上是相互矛盾的,但实际上,由于它们在不同的负载条件下有各自的优势,所以,综合考虑这些优化策略会使系统获得较优的性能.基于效益-代价均衡的调度算法就是基于这种思想设计的.

首先,我们根据突发性、重负载的磁带库请求特点,采用相关数据热复制策略,增强每次磁带调度的服务率.由于这类请求不具有很强的偏斜性(低于 60/40,即 60%的请求针对 40%的数据,而数据库界公认负载特点是 80/20 规则),因而磁盘缓存的作用不明显;但这类请求负载具有一个突出的特点,就是所存取的大量数据具有较强的相关性,因此,我们提出相关数据热复制策略,对相关的数据进行热复制并选用适当放置策略,从而增加每次调度包含的请求数,减少磁带库交换磁带的次数.为此,我们综合考虑了磁带的顺序存取特点和复制的空间代价,选择在磁带的尾部复制相关数据(在我们的实验中使用 20%的带长存放相关数据的拷贝),每个相关数据只复制一份,且数据本身与数据的拷贝不在同一条磁带上.实验表明,相关数据热复制策略极大地增加了调度的服务率,在一般的负载条件下会使系统获得很好的性能,但它服务的不公平性导致重负载下系统性能的恶化.

为此,我们在算法中引入了效益-代价均衡的思想,从磁带选择入手,均衡考虑算法的效益和代价.这其中的核心问题是确定磁带选择策略,即当有驱动器空闲时应该如何选择合适的磁带进行调度的策略.由于磁带库系统中磁带数远远多于驱动器数,所以磁带选择策略的好坏直接影响系统的性能.常用的磁带选择策略包括循环(round robin)选择、最老请求选择、最多请求选择、最大有效带宽选择等.其中,最多请求选择和最大有效带宽选择属于贪心策略,循环选择和最老请求选择属于公平策略.基于贪心策略和公平策略各自的优点,我们在磁带选择估算中引入效益-代价加权的概念.其中,“效益”是指磁带执行的有效带宽(即一次调度所传输的数据总量(MB)除以该调度执行所花费的时间(秒)所得到的值),该值越大,调度的效益越高,系统的性能越好;“代价”是指磁带上请求的平均等待时间,该值越大,调度的代价越高,系统的性能也越差.算法根据负载的变化,通过调节效益和代价的权值来获得最佳的磁带选择策略.下面是磁带选择算法中磁带的效益代价计算公式(简式):

$$E_t = W_{bw} * \frac{S_t}{S_{\text{avg}}} + W_{wt} * \frac{T_t}{T_{\text{avg}}}. \tag{1}$$

其中,E_t 是磁带 t 的效益代价估算值,调度算法总是选择 E_t 最大的磁带作为下一个服务的磁带.S_t 是待计算的磁带 t 的有效带宽,T_t 是待计算的磁带 t 上的当前所有请求的平均等待时间,这两个参数是每个磁带的自有参数.S_{avg} 是系统当前平均有效带宽,T_{avg} 是系统当前已完成的所有请求的平均等待时间,两个参数值都是随着系统

的执行而不断变化的,反映系统当前的平均性能.W_{bw} 是有效带宽的效益加权,W_{wt} 是平均等待时间的代价加权,两者值的相对变化将直接影响磁带选择策略的倾向,我们将在下一节详细论述之.

基于效益-代价均衡的调度算法是通过磁带选择策略对调度进行优化的,它使系统在各种负载条件下都能获得很好的性能.

2 效益-代价估算

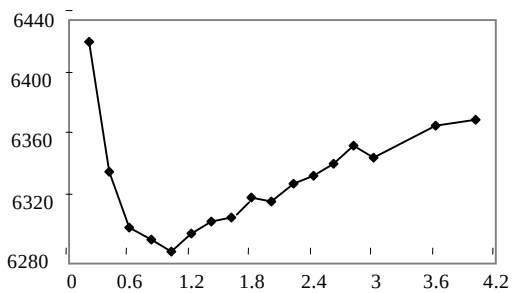
由第 1 节可知,基于效益-代价均衡的调度算法的关键问题是确定效益加权 W_{bw} 和代价加权 W_{wt}.由于系统状态错综复杂,且动态多变,所以,理论推导出不同负载下的最佳权值是非常困难的.本文给出了一种实用而有效的估算方法,称之为负载估算法.下面先介绍推出负载估算法的 3 个重要结论.

2.1 3个重要的实验结论

为了获得效益-代价加权的估算方法,我们做了大量仿真实验,仿真环境参见本文第 4 节.从实验中,我们发现了效益-代价加权与系统性能、系统负载之间的 3 个重要规律:

(1) 磁带库系统性能只与代价-效益加权的比值 $R_{w/b}$($R_{w/b}=\frac{W_{wt}}{W_{bw}}$)有关.这个结论是很显然的,因为该比值反映了效益和代价在磁带选择中的权重,不同的大小决定了磁带选择的不同倾向,或者有效带宽占主导(比值很小时),或者平均等待时间占主导(比值很大时),或者综合考虑两方面因素.利用这一规律,可以把效益-代价加权的估算简化为对 $R_{w/b}$ 的估算.

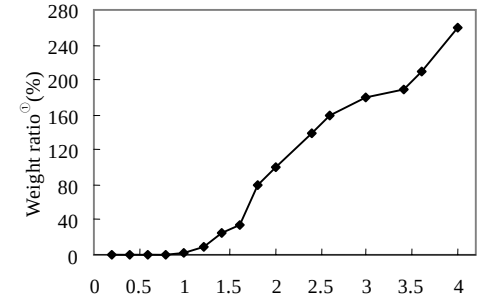
(2) 在每一个系统负载条件下,$R_{w/b}$ 都存在且只存在一个使系统性能最好的最佳值区间.这结论说明在某一负载下只要找到最佳值区间中的某一值就可以获得接近最好的系统性能.图 1 给出了请求到达率为 2.0 时的系统性能与加权比的关系.它表明,确实存在一个加权比值的最佳值区间(本实验中为 0.8~1.2),在该区域内取值,可以使系统性能得到极大的改善.



①平均响应时间,②等待时间和有效带宽加权比.

Fig.1 Impact of weight ratio on response

图 1 加权比对响应时间的影响



①加权比,②请求数/分钟.

Fig.2 Relationship of Weight ratio and workload

图 2 加权比与请求负载的关系

(3) $R_{w/b}$ 的最佳值区间内的值是随着系统负载的加重而增大的.这表明系统负载的加重使 $R_{w/b}$ 的最佳值有增大的趋势.这个结论也很直观,因为负载的加重促使系统更多的考虑调度的代价.图 2 给出了加权比值的最佳值与请求负载之间的关系.在请求负载较轻时,加权比值极小,表明此种情况下最大有效带宽策略占主导.随着负载的加重,加权比值开始增加,说明重负载条件下需要均衡最大有效带宽的效益和请求长时间等待的代价,才能获得好的系统性能.

2.2 负载估算法

上述 3 个结论是我们对效益-代价进行近似估算的重要依据.为此,我们把式(1)变换成如下形式:

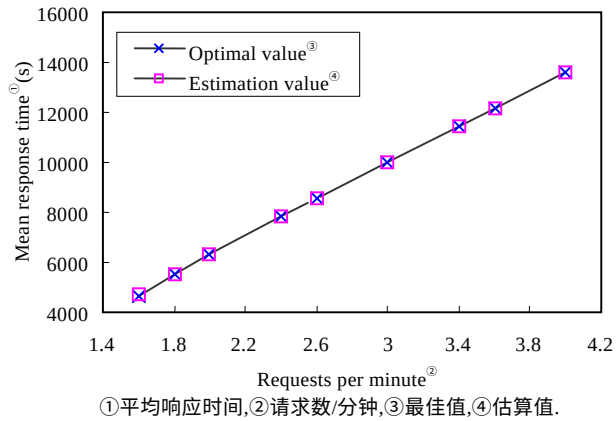
$$E'_t = 1 * \frac{S_t}{S_{\text{avg}}} + R_{w/b} * \frac{T_t}{T_{\text{avg}}}. \quad (2)$$

这样,只要估算出 $R_{w/b}$ 的值,磁带选择策略就确定了.下面是我们对 $R_{w/b}$ 的估算公式:

$$R_{w/b} = \begin{cases} 0 & \frac{\lambda * S}{60 * TR * N} < 0.9 \\ \frac{\lambda * S}{60 * TR * N} & \frac{\lambda * S}{60 * TR * N} \geq 0.9 \end{cases}. \quad (3)$$

其中,λ表示请求的平均到达率,即每分钟平均到达λ个请求;S 表示请求的平均大小(MB);TR 表示磁带驱动器的传输率(MB/秒),N 表示磁带库的驱动器个数,60 是时间换算因子.事实上,$\frac{\lambda * S}{60 * TR * N}$ 是系统负载的粗略估算,用它来估算效益-代价加权比正体现了效益-代价估算与负载的关系,所以我们把这种方法称为负载估算法.

仿真实验表明,当 $\frac{\lambda * S}{60 * TR * N}$ <0.9 时,$R_{w/b}$ 的估算偏差超过 4%,而 $R_{w/b}$ 的最佳值对系统性能的改善率小于 1%,所以,在这种情况下,$R_{w/b}$ 取 0,对系统性能的影响小于 1%.当 $\frac{\lambda * S}{60 * TR * N} \geq 0.9$ 时,$R_{w/b}$ 估算值对系统性能的改善与最佳值相比偏差小于 1%.通过仿真实验对这一结果做了大量比较,结果偏差几乎都小于 1%.图 3 给出了利用 $R_{w/b}$ 估算值得到的系统性能与利用 $R_{w/b}$ 最佳值(通过大量仿真实验获得)得到的系统性能的比较,其中请求大小为 64M,请求倾斜度为 70/10.图 3 表明 $R_{w/b}$ 估算公式在偏差允许范围内是可以接受的.



①平均响应时间,②请求数/分钟,③最佳值,④估算值.

Fig.3 Performance comparison for estimation value and optimal value of $R_{w/b}$

图 3 $R_{w/b}$ 估算值与最佳值的系统性能比较

3 算法实现

本节将给出基于效益-代价均衡的磁带库调度算法的一个近似优化的实现方法.由于该算法引入了相关数据热复制策略,使相关数据和它的拷贝同时存在,从而增加了磁带选择调度的复杂性.我们的实现策略是分界法.该方法把所有请求分为两类:U 类请求和 R 类请求,U 类请求指请求所涉及的数据无复制的情况,R 类请求指请求所涉及的数据有复制的情况.算法通过 U 类请求确定每个磁带的请求分界点,它标明这些磁带执行必须定位的最远距离.对于 R 类请求,则根据请求的位置(在分界点以内或以外)分别进行处理.下面给出算法的详细描述:

(1) 当新请求到达时,首先进入请求等待队列中等待;当有驱动器空闲时,新一次调度开始.

(2) 顺序扫描请求等待队列,所有请求按照磁带分组,把 U 类请求直接按地址升序*插入相应磁带调度列表中,并把各磁带调度列表中的最后一个请求的起始地址记为该磁带的分界点;把 R 类请求插入一个临时等待队列 TQ.

(3) 顺序扫描 TQ,把请求插入相应磁带的调度列表中.对于请求数据地址大于所插入磁带调度列表的分界点的请求,还需要为该请求数据的每个拷贝复制一个存取其数据拷贝的请求,并把这些请求插入到拷贝所在的磁带调度列表中.

(4) 利用式(3)计算加权比 $R_{w/b}$;计算系统当前的平均有效带宽 S_{avg} 和已完成请求的平均等待时间 T_{avg};计算每一个磁带调度列表的有效带宽 S_t 和请求的平均等待时间 T_t,利用式(2)估算每一个磁带调度列表的效益-代价值.

* 这里使用的 Eliant 820 磁带驱动器是螺旋扫描方式,其顺序扫描的方式是按地址线性递增扫描的,所以,本算法中请求按地址顺序排序;如使用蛇型扫描方式的 DTL 驱动器,则请求顺序应按文献[3]中定位模型排序.

(5) 把具有最大效益-代价估算值的磁带调度列表 L_{max} 提交给空闲的驱动器执行服务.如果提交的磁带刚好存在于其他驱动器中,则把它插入该驱动器等待队列中等待,然后在其余的磁带调度列表中找到具有次大估算值的磁带列表 L_{next} 进行调度,在估算之前,需要把 L_{max} 中涉及的有复制的数据的相应拷贝从其他列表中清除,因为这些数据的请求已在 L_{max} 中得到服务.

上述步骤完成一次调度选择过程,其中,步骤(2)的目的是通过 U 类请求确定必选的所有磁带,并使用分界点标明这些磁带执行必须定位的最远距离.步骤(3)把 R 类请求中的在分界点以内的请求直接插入,这样可以减少定位时间,而不考虑它们的复制数据,从而降低了算法的复杂性;对分界点外的请求则考虑其拷贝对选择磁带的作用,增加每次调度包含的请求数.

4 仿真结果

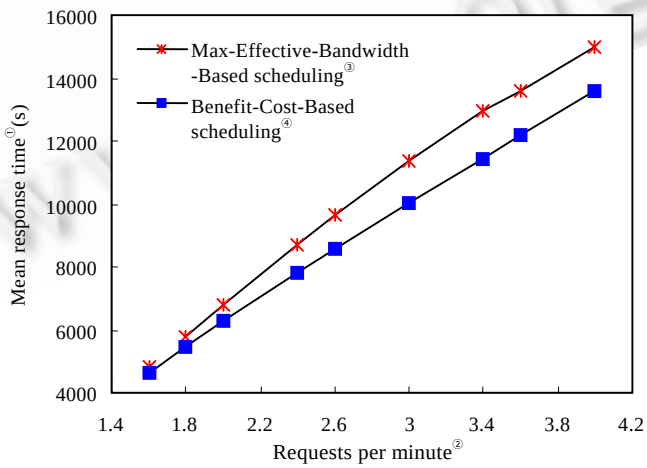
本文对基于效益-代价均衡的磁带库调度算法进行了仿真研究.磁带库仿真系统以 Exabyte 220 磁带库、Eliant 820 磁带驱动器、EXABTYE 8mm 磁带的特性为基础,实验使用的性能参数除磁带容量外都是通过实际测量得到,参数见表 1.该磁带库由 2 个驱动器、1 个机械手和 20 盘带组成.

Table 1 Measurement of tape library Exabyte 220

表 1 Exabyte 220 磁带库的性能参数

TAPE (1) (20)	Capacity (4)	7GB(uncompressed(14))
TAPE DRIVE (2) (2)	Data transfer rate(5)	0.92MB/s (uncompressed)
	Mean drive load time(6)	25s
	Mean drive eject time(7)	23s
	Rewind startup time(8)	7s
	Rewind rate(9)	22MB/s
	Seek startup time(10)	8.5s
	Seek rate(11)	30.2MB/s (uncompressed)
TAPE ROBOT (3) (1)	Mean tape unload time(12)	7.2s
	Mean tape load time(13)	7.42s

(1)磁带,(2)磁带驱动器,(3)磁带机械手,(4)容量,(5)数据传输率,(6)平均驱动器加载时间,(7)平均驱动器卸载时间,(8)反绕启动时间,(9)反绕速率,(10)搜寻启动时间,(11)搜寻速率,(12)平均磁带卸载时间,(13)平均磁带加载时间,(14)未压缩.



①平均响应时间,②请求数分钟,③基于最大有效带宽的调度,④基于效益-代价均衡的调度.

Fig.4 Benefit-Cost-Based scheduling

图 4 基于效益-代价均衡的调度

图 4 给出了基于效益-代价均衡调度算法与基于最大有效带宽调度算法的性能比较.从图 4 可以看出,在重

负载条件下（到达率>1.5）,效益-代价均衡调度显示明显的优势,且随着负载的加重,其优势越明显.而在到达率≤1.5 时,如第 2 节所讨论,因效益-代价估算对系统性能的改善在 1%以内,所以,我们设 $R_{w/b}$ 为 0,即磁带选择策略为最大有效带宽策略,此种条件下的系统性能与基于最大有效带宽的调度算法相同,故图 2 中未画出此种情况.

5 结 论

本文提出并研究了一种基于效益-代价均衡的调度算法,该算法侧重改善重负载条件下的系统性能.针对系统的效益-代价难以估算的问题,本文给出一个简单、实用而又有效的估算方法,从而解决了该算法的实用问题.进一步的工作是研究磁带库动态调度算法的效益-代价估算问题.

References:

[1] Cariño, F., Kaufmann, A., Kostamaa, P. Are you ready for Yottabytes? In: Kobler, B., ed. Proceedings of the 17th IEEE Symposium on Mass Storage Systems in Cooperation with the 8th NASA GSFC Conference on Mass Storage Systems and Technologies. Los Alamitors, CA: IEEE Computer Society Press, 2000. 476~485.

[2] Cariño, F., Burgess, J., O'Connell, W., *et al*. Active storage hierarchy, database systems and applications——socratic exegesis. In: Malcolm, P.A., Maria, E.O., *et al*., eds. Proceedings of the 25th International Conference on Very Large Data Bases. Edinburgh: Morgan Kaufmann Publishers, Inc., 1999. 611~614.

[3] Hillyer, B.K., Silberschatz, A. Random I/O scheduling in online tertiary storage systems. In: Jagadish, H.V., Mumick, I.S., eds. Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data. Quebec: ACM Press, 1996. 195~204.

[4] Triantafillou, P., Georgiadis, I. Hierarchical scheduling algorithms for near-line tape libraries. In: Cammelli, A., Wagner, R. R., eds. Proceedings of the 10th International Conference and Workshop on Database and Expert Systems Applications. Florence: IEEE Computer Society Press, 1999. 50~54.

[5] Hillyer, B. K., Rastogi, R., Silberschatz, A. Scheduling and data replication to improve tape jukebox performance. In: Papazoglou, M., Pu, C., Kitsuregawa, M., eds. Proceeding of the 15th International Conference on Data Engineering. Sydney: IEEE Computer Society Press, 1999. 532~541.

[6] Nemoto, T., Kitsuegawa, M. Scalable tape archiver for satellite image database and its performance analysis with access logs——hot declustering and hot replication. In: Miller, E., ed. Proceedings of the 16th IEEE Symposium on Mass Storage Systems in Cooperation with the 7th NASA GSFC Conference on Mass Storage Systems and Technologies. San Diego: IEEE Computer Society Press, 1999. 59~71.

A Cost-Benefit-Based Scheduling Algorithm of Online Tape Library*

SHI Jing , XING Chun-xiao , ZHOU Li-zhu

(*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084, *China*)

E-mail: shijing@mails.tsinghua.edu.cn; xingcx@tsinghua.edu.cn; dcszlz@tsinghua.edu.cn

http://dbgroup.cs.tsinghua.edu.cn

Abstract: The terabyte-level (10^{12} bytes) database systems such as digital libraries need to use tape library as an online device to store and retrieve their massive data. Since the access of a tape library is often random and the tape library has low random access performance, thus it is critical to study the random I/O scheduling strategies and algorithms in order to improve the performance of tape library. In this paper we study a cost-benefit-based scheduling algorithm, and as well as give an effective estimating method of cost-benefit weight ratio. This algorithm improves the performance of tape library system under different workloads by dynamically tuning the cost-benefit weight ratio of scheduling policies according to workloads. This algorithm particularly overcomes the problem of workload-sensitive of existing scheduling algorithms, and is significantly effective under heavy workload.

Key words: online tape library system; random I/O scheduling of tapes; hot data replication; cost-benefit estimating

* Received April 20, 2001; accepted September 5, 2001
Supported by the National Grand Fundamental Research 973 Program of China under Grant No.G1999032704