第44卷 第3期
2012年6月

西安建筑科技大学学报(自然科学版)
J. Xi'an Univ. of Arch. & Tech. (Natural Science Edition)

Vol. 44 No. 3
Jun. 2012

基于价值评估的海量数据迁移策略研究

边根庆[1],王艳云[1],邵必林[2],于泳浩[3]

(1.西安建筑科技大学信息与控制工程学院,陕西 西安 710055;2.西安建筑科技大学管理工程学院,陕西 西安 710055;
3.南京大学信息管理学院,江苏 南京 210046)

摘 要:根据分级存储管理HSM(hierarchical storage management)中的数据迁移任务,提出了一种基于价值评估的数据迁移模型.通过按比例计算数据的固有属性和预期价值,可以得到数据价值的精确表达,结合迁移过程控制策略,相应价值的数据将被分配到与其价值相适应的存储设备上.仿真实验结果表明,相对于LRU和LFU,该迁移算法能使绝大多数的访问命中于在线存储设备,并随着访问次数的增加,可以逐步提高算法的精度.

关键词:分级存储管理;数据迁移技术;固有属性;预期价值

中图分类号:TP309.3 **文献标志码**:A **文章编号**:1006-7930(2012)03-0441-05

数据信息的爆炸式增长给信息存储管理带来巨大的挑战,其增加了存储成本,使存储管理更加复杂,并且降低了访问速度[1-2].分级存储管理(HSM[3])技术的出现,为这一问题的解决提供了思路.如何对存储的数据进行迁移是分级存储管理中的一项关键技术[4].

目前分级存储普遍采用的两个数据迁移方法:基于存储空间的高低水位法[5]和基于数据访问率的Cache替换迁移算法[6].高低水位法中数据能否迁移,先决条件是磁盘剩余空间是否足够.需要对磁盘空间饱和度进行实时监控,判断磁盘空间饱和度的情况,以便在合适的时候启动迁移程序.此方法主要关注的是存储设备,并且基于它的存储状态迁移数据,没有考虑数据本身的特征,所以数据没有得到高效利用.Cache替换迁移算法采用一系列的方法移除无价值的数据,有许多的替换策略,包括FIFO、LRU、LFU、SIZE、LRV及Hybrid等.其中代表性的替换迁移算法为:LRU(Least-Recently-Used)[7]算法,将最近最少使用的数据移出磁盘,优点是实现简单,其缺点是只考虑数据的一个时间特性,没有考虑到数据访问的全局性,只是在访问时间上的局部优化,因此效率并不很高;LFU(Least-Frequently-Used)[8]算法,将访问次数最少的数据移出磁盘,LFU算法对磁盘中的每个数据设置一个引用计数器,在进行替换选择时,具有最低引用计数值的数据被替换.其优点是使用数据的访问频率,有利于数据的总体优化使用,其缺点是会存在磁盘污染,如果没有失效机制,可能使过时的数据永远留在磁盘里,而不会被其它数据所替换.

为了克服以上存在的问题,本文提出了基于数据价值的迁移模型MSDV(Migration Strategy Based on Data Value),此模型把数据的重要性和未来一段时间内被访问的可能性定义为存储系统中数据价值[9]DV(Data Value),它不仅考虑由数据大小(S)、访问时间(T)、数据的读写频率(F)和访问量(C)决定的固有属性IP(Inherent Property),而且考虑代表访问可能性的预期价值EV(Expected Value).综合考虑这两个因素,得到数据价值的精确表达,使重要数据和常用数据存储在高性能的存储设备中,而一些极少使用的数据则备份在廉价的海量存储设备中,在提高存储性能的同时降低了总所有成本TCO(Total cost of ownership).

收稿日期:2011-09-26 **修改稿日期**:2012-05-03

基金项目:国家自然科学基金资助项目(61073196);陕西省自然科学基础研究计划项目(2011JM8026);陕西省教育厅自然科学专项基金项目(11JK0982);

作者简介:边根庆(1968-),男,浙江浦江人,副教授,主要研究领域为海量信息处理、信息安全等.

1 MSDV 数据价值评估模型的设计

1.1 算法模型

MSDV 模型包括 *IP* 和 *EV*,如图 1 所示. *IP* 是数据迁移的先决条件,其代表数据的重要性,包括 *S*,*T*,*F*,*C*. 通过使用这些参数来计算数据的 *IP* 价值,*IP* 值越大,迁移的可能性越大. 为了判断数据的价值高低,以便进一步迁移,应该预见数据访问的可能性. *EV* 值越高,访问的可能性越大,反之亦然,*EV* 值的可以通过计算用户之间的相似性得到.

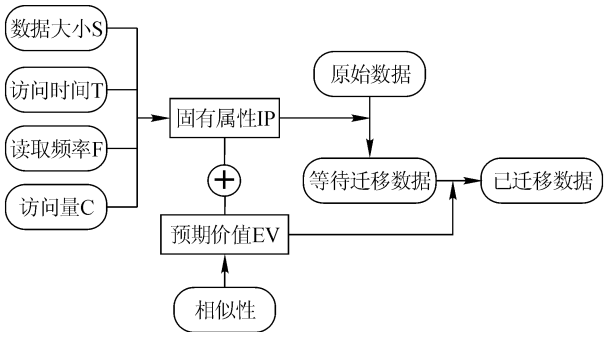

图 1 MSDV 模型

Fig. 1 MSDV model

数据 *X* 的 *IP* 值可以这样计算:

$$IP(X)=\frac{1}{S}+T+F+C. \tag{1}$$

1.2 参数设定

1) 数据大小 *S* 对于分级存储系统,高性能磁盘阵列的容量是有限的,如果较大的数据存储在高性能磁盘阵列,其会占据很大的存储空间,小而且重要的数据就无法及时被访问到,导致存储系统性能的降低,因此小数据的价值更高,应优先迁移小数据到高性能磁盘阵列中.

2) 访问时间 *T* 根据 *LRU* 数据迁移算法的思想,最新创建或最近使用过的数据较最近未使用的数据,前者被访问的可能性大,相应价值也就较高. 被访问过后未使用的时间越长,数据重要性就越低,数据被重新访问的可能性也相应降低,就需要迁移到较低性能的存储设备中.

定义 1.1 数据被创建以后,每次访问和修改的时间是集合$\{t_1,t_2,\cdots,t_n\}$,当前时间是 t,这些时间点距离当前时间的长度是 $t-t_1,t-t_2,\cdots,t-t_i,\cdots,t-t_n$,设以上时间长度为 $T_1,T_2\cdots,T_i\cdots,T_n$. 数据 *X* 的 *T* 值为:

$$T=\sum_{i=1}^{n}\frac{1}{T_i}=\frac{1}{T_1}+\frac{1}{T_2}+\cdots+\frac{1}{T_i}+\cdots+\frac{1}{T_n}. \tag{2}$$

3) 数据的读写频率 F[10] 数据被使用最直接的体现是读写操作,读写频率高的数据其价值更高.

定义 1.2 数据的读写频率用 *R* 和 *W* 表示,那么数据 *X* 的 *F* 值为:

$$F=\sum_{i=1}^{N}\{k_wW_i+k_rR_i\}. \tag{3}$$

其中是读补偿系数,是写补偿系数,它们表示存储设备在读写操作时间代价上的比值,W_i 和 R_i 表示数据 *X* 被创建以后每次访问的读写频率,*N* 表示访问次数.

4) 访问量 *C* 用户访问数量是数据价值高低的直接体现,数据被访问的用户数量越多,其价值就越高,如果一个数据被多个用户访问,那么它的改变和访问性能就会对更多的用户构成影响[9].

定义 1.3 数据 *X* 的访问量称作 *C*,每一次的访问记录用 *UD* 表示:

$$C=counter(UD) \tag{4}$$

$$UD=UD_1\cup UD_2 \tag{5}$$

在这里, $$UD_1=\{ud_1(i)\mid ud_1(i)\in \text{不同时间相同用户的访问记录}\}. \tag{6}$$

UD_1 越长,对用户来说越重要,*DV* 值越高.

$$UD_2=\{ud_2(i)\mid ud_2(i)\in \text{不同用户的访问记录}\}. \tag{7}$$

$counter(x)$ 函数,求数据访问记录的长度得到的是访问数据的用户数量.

5) 预期价值 *EV* 在 MSDV 模型中,数据 *X* 的 *EV* 值代表用户之间的相似性. 换句话说,通过计算用户访问 *X* 和没有访问 *X* 的相似性,有着高的相似性的用户被看做访问 *X* 的潜在用户,潜在用户数量,被称作 *X* 的 *EV* 值.

定义 1.4 如果用户 *u* 访问数据 *i*,则 $a_{u_i}=1$,否则 $a_{u_i}=0$. 用户 *u* 和 *v* 的相似性可以被表述成:

$$Sim(u,v)=\frac{\sum_{i=1}^{n}a_{u_i}a_{v_i}}{k(u)+k(v)}. \tag{8}$$

在这里,n 是两个用户访问数据的数量,$k(u)=\sum_{i=1}^{n}a_{u_i}$,$k(v)=\sum_{i=1}^{n}a_{v_i}$ 表示用户 u 和 v 的访问程度.

具体步骤如下:

a) 假设访问 X 的用户用 U 来表示,其它用 V 来表示,则有:

$$U=\{u_1,u_2,\cdots,u_n\}. \tag{9}$$

$$V=\{v_1,v_2,\cdots,v_m\}. \tag{10}$$

在这里,m 和 n 是用户总数.

b) 计算 U 和 V 元素之间的相似性,可以得到一个相似矩阵.

$$Sim(U_i,V_j)=\begin{bmatrix}S_{11} & S_{12} & \cdots & S_{1m}\\ S_{21} & S_{22} & \cdots & S_{2m}\\ S_{n1} & S_{n2} & \cdots & S_{nm}\end{bmatrix}. \tag{11}$$

在这里,$i\in(1,2,\cdots,n)$,$j\in(1,2,\cdots,m)$.

c) 根据大小,查看相似矩阵(11) 中每一行中的元素. 其值大于 1/2(这个参数可以调整) 的元素中的用户被看做潜在用户,用 V' 来表示,则有:

$$V'=[V'_1\quad V'_2\quad\cdots\quad V'_n]. \tag{12}$$

在这里,V'_i 为与一致的用户. V'_i 的总数是用户访问 X 的 EV 值,用 $counter(x)$ 函数计算.

d) 每个子 EV 的和是 X 的 EV 值的总数.

$$EV=\sum_{i=1}^{n}counter(V'_i). \tag{13}$$

e) 在集合 V' 里有这么一种情况,如果 $V'_1=\{v_1,v_2,v_3,v_4\}$,$V'_2=\{v_3,v_4,v_5,v_6\}$,我们发现 v_3 和 v_4 被重复计算,重复的用户应只计算一次,X 的 EV 值可以这样计算:

$$EV(X)=counter(V'_1\cup V'_2\cup\cdots V'_n). \tag{14}$$

综合考虑 IP 和 EV 来决定 DV,DV 是数据迁移的基础. 数据 X 的 DV 值:

$$DV(X)=\alpha IP(X)+(1-\alpha)EV(X). \tag{15}$$

这里是一个调节参数,$0\leqslant\alpha\leqslant1$.

约束条件:

$$\sum_{i=1}^{n}S_i\leqslant Q. \tag{16}$$

在这里 S_i 是数据 i 的大小,n 是迁移数据量,Q 是存储容量.

2　基于 MSDV 的数据迁移过程

2.1　分级存储结构

在分级存储结构中数据分为在线(On-Line)存储、近线(Near-Line)存储、离线(Off-Line)存储. 其中在线存储采用高端存储产品(如 FC 磁阵)存储一些需要经常和快速访问的数据;离线存储采用一些低端存储产品(如磁带库)存储一些很少被访问或不被访问的数据;近线存储采用存取速度和价格介于磁盘与磁带之间的低端存储设备(如光盘库)存储并不是经常用到或者周期性访问的数据. 如何迁移不同存储层的数据是分级存储的关键所在.

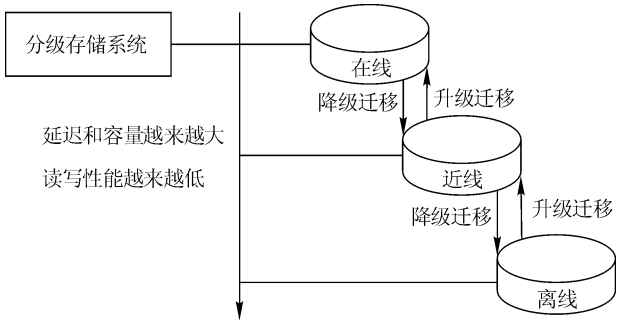

图 2　分级存储数据迁移方向

Fig. 2　The direction of hierarchical storage data migration

分级存储系统数据的迁移包括两部分:降级迁移和升级迁移,如图 2 所示. 所谓降级迁移就是将非热点数据从高端存储设备向低端存储设备迁移的过程,迁移总体方向为“在线——近线,近线——离

线”.升级迁移恰好相反是将热点数据从低端存储设备迁移到高端存储设备的过程,迁移总体方向为“离线——近线,近线——在线”.不管是降级迁移还是升级迁移都需要以数据的价值作为依据,选择合适的数据组建任务队列.

2.2 基于 MSDV 的数据迁移执行过程

有了数据价值评估模型,就可以在分级存储平台上以此指导存储系统对数据的迁移.具体的迁移步骤如下:

步骤 1:在分级存储平台上,按照分级存储的要求对数据进行存储,一般新创建的数据都存储在在线存储设备;

步骤 2:创建迁移线程和双候选迁移任务队列,包括升级迁移的任务队列和降级迁移的任务队列;

步骤 3:根据 MSDV 总公式(15)对在线,近线,离线存储设备中所有数据的价值进行计算,然后按价值高低进行排序,并将其反馈给迁移调度模块,迁移调度模块根据数据价值的排序调整迁移任务队列,转入迁移过程控制;

步骤 4:迁移完成后对数据的相关属性信息进行修改,定位数据到迁移后的位置,置原来的数据存储位置为无效,是否终止数据迁移过程,若否,则转入步骤 2,若是,则结束迁移.

整个迁移过程如图 3 所示.

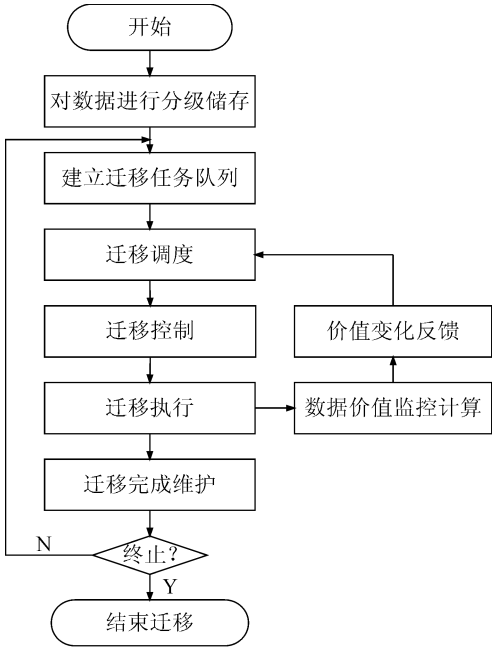

图 3 基于 MSDV 的数据迁移流程

Fig. 3 Data migration process based on MSDV

3 性能测试与分析

本文采用命中率来表示存储系统的性能,当用户访问磁盘中的数据时,如果该数据存在于在线存储设备称之为命中,反之称为不命中.

$$在线存储命中率=\frac{在线存储命中次数}{I/O\ 总请求次数} \quad (17)$$

实验采用卡内基梅陇大学开发的 Disksim 模拟器[11]来仿真存储系统,比较了 LRU,LFU 和 MSDV 数据迁移算法.Disksim 是局部模拟器,采用踪迹驱动(Trace-driven)方法进行测试.在测试过程中,采用 HP Lab 的 Cello92[12]负载,其是一个很经典的 I/O Trace,是由 Cello 系统于 92.4.8—92.6.20 取得的负载,记录了 60 多天的时间内,在某多用户环境中,日常文件读写相对应的磁盘操作情况,数据访问的对象为 8 块磁盘,每天的平均读写命中数大约在 40 万左右.实验进行了两类测试,第一类测试了在 Cello92 负载驱动下,当在线存储设备容量变化时,三种迁移算法的命中率,如图 4 所示.第二类测试了在 Cello92 负载驱动下,当在线存储设备容量固定,访问次数总量变化时,三种迁移算法的命中率,如图 5 所示.

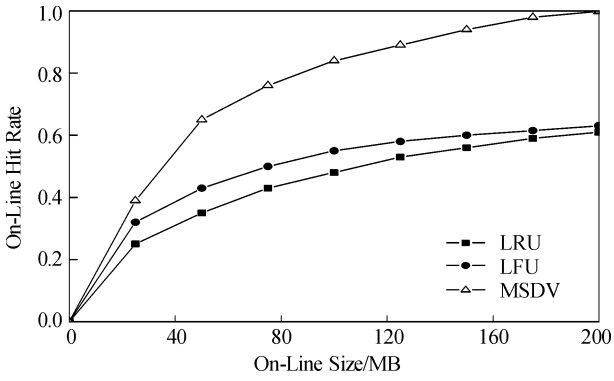

图 4 Cello92 负载下各迁移算法在不同容量在线存储设备中的命中率

Fig. 4 The hit rate of the migration algorithm in different capacity On-Line storage devices Under Cello92 load

从图 4,图 5 可以得到如下结论:

①在第一类测试中,不同容量的在线存储设备上,访问率较高的数据几乎均命中于在线存储设备,验证了本文设计的 MSDV 数据价值评估模型应用于分级存储系统中后,能使绝大多数访问命中于在线存储设备,从而极大地提高了系统的访问性能;在第二类测试中,当在线存储的容量固定,访问次数总量变化时,验证了随着访问次数的增加可以逐步提高对数据价值分析的准确度.

②以上实验数据表明:MSDV 迁移算法要比传统的基于数据访问频率的 Cache 替换迁移算法

LRU、LFU 更能高效地使绝大多数访问命中于在线存储设备,达到了预期的效果.

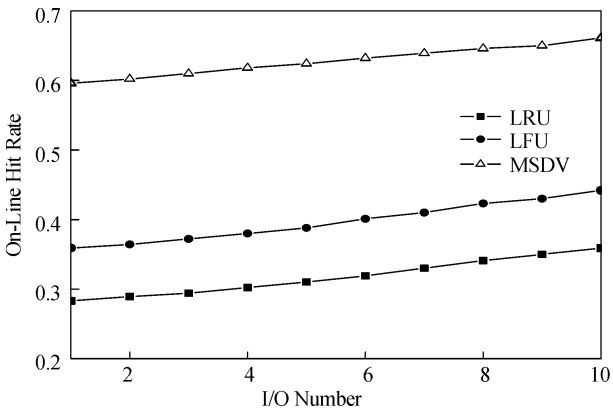

图 5　Cello92 负载下各迁移算法在不同访问次数总量($\times 10^5$)下的命中率(在线存储容量设定为 50 MB)

Fig. 5　The hit rate of the migration algorithm in different I/O number ($\times 10^5$), the On-Line storage capacity is set to 50MB Under Cello92 load

4　结　语

数据迁移技术在容灾恢复、负载均衡、前置数据、应用开发、数据仓库等方面越来越有着广泛的应用,其要求保证在不中断系统运行的情况下,将数据进行流动并且始终保持一致,同时确保数据的立即可用,从而改善存储系统性能. 本文在分级存储平台的基础上设计了 MSDV 数据价值评定模型和迁移过程控制策略,实现对数据价值的精确判定,在尽量减小对系统访问性能影响的基础上,实现数据在三级存储设备间无人值守的迁移,与传统的数据迁移算法相比较,更加合理、高效,在保证存储系统性能的基础上大大降低了存储成本. 密切关注各种高性能存储设备的研究进展,调整已经设计好的数据迁移策略,使之能够实时的适应复杂环境的要求,将是下一步工作需要研究的重要课题.

参考文献　References

[1] BRINKMANN A, SALZWEDEL K, SCHEIDELER C, et al. Storage management as means to cope with exponential information growth[C]// L'Aquila. Proceedings of SSGRR, 2003: 73-79.

[2] 邵必林,吴宝江,边根庆. 基于 ISCSI 技术的 SAN 应用研究[J]. 西安建筑科技大学学报:自然科学版,2009,41(1):112-116.
SHAO Bi-lin, WU Bao-jiang, BIAN Gen-qing. Application study of SAN based on ISCSI technology[J]. J. Xi′an UNiv. of Arch. &Tech.: Natural Science Edition, 2009, 41(1): 112-116.

[3] ZHAO Xiao-nan, LI Zhan-huai, ZENG Lei-jie. A Hierarchical storage strategy based on block-level data valuation [C]// The proceeding of the 4th Internatimal conference on Net-worked Computing and Advanced Information Management, 2008: 36-41.

[4] 冯　泳,张延园. 数据迁移在 SAN 中性能优化的研究和应用[J]. 计算机工程,2005(4):43-45.
FENG Yong, ZHANG Yan-yuan. Research and application of data migration in storage area network performance optimiza-tion [J]. Computer Engineering, 2005(4): 43-45.

[5] 唐　竟. 基于信息生命周期管理数据迁移技术研究[D]. 长沙:湖南大学,2009.
TANG Jing. Research of Date Migration's technology based on the Information Lifecycle Management[D]. Changsha: Hunan University, 2009.

[6] JEONG J, DUBOIS M. Cost-sensitive Cache Replacement Algorithms[C]// Proceedings of the Symposium on High-Performance Computer Architecture (HPCA), 2003(1): 327-337.

[7] Smitha and A. L. N. Reddy. LRU-RED: An active queue management scheme to contain high band-width flows at congested routers[J]. In Proceeding of Global Telecommunications Conference, 2001.

[8] B. Reed and D. D. Long. Analysis of caching algorithms for distributed file systems[J]. Operating Systems Review, 1996, 30(3): 12-21.

[9] 吕　帅,刘光明,徐　凯,等. 海量信息分级存储数据迁移策略研究[J]. 计算机工程与科学,2009,31(A1):163-167.
LÜ Shuai, LIU Guang-ming, XU Kai, et al. Research on the data migration strategy of hierarchical mass storage system[J]. Computer Engineering&Science, 2009, 31(A1): 163-167.

[10] 江　菲,汤小春,张　晓,等. 基于价值评估的数据迁移策略研究[J]. 电子设计工程,2011,19(7):11-13.
JIANG Fei, TANG Xiao-chun, ZHANG Xiao, et al. Research of data migration strategy based on data valuation[J]. Electronic design engineering, 2011, 19(7): 11-13.

(下转第 451 页)

LRU、LFU 更能高效地使绝大多数访问命中于在线存储设备,达到了预期的效果.

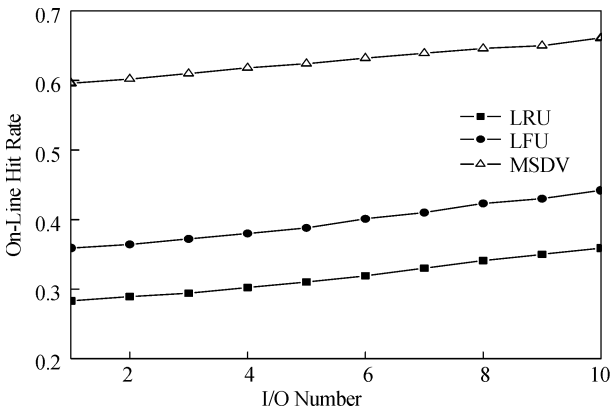

图 5　Cello92 负载下各迁移算法在不同访问次数总量($\times 10^5$)下的命中率(在线存储容量设定为 50 MB)

Fig. 5　The hit rate of the migration algorithm in different I/O number ($\times 10^5$), the On-Line storage capacity is set to 50MB Under Cello92 load

4　结　语

数据迁移技术在容灾恢复、负载均衡、前置数据、应用开发、数据仓库等方面越来越有着广泛的应用,其要求保证在不中断系统运行的情况下,将数据进行流动并且始终保持一致,同时确保数据的立即可用,从而改善存储系统性能. 本文在分级存储平台的基础上设计了 MSDV 数据价值评定模型和迁移过程控制策略,实现对数据价值的精确判定,在尽量减小对系统访问性能影响的基础上,实现数据在三级存储设备间无人值守的迁移,与传统的数据迁移算法相比较,更加合理、高效,在保证存储系统性能的基础上大大降低了存储成本. 密切关注各种高性能存储设备的研究进展,调整已经设计好的数据迁移策略,使之能够实时的适应复杂环境的要求,将是下一步工作需要研究的重要课题.

参考文献　References

[1] BRINKMANN A, SALZWEDEL K, SCHEIDELER C, et al. Storage management as means to cope with exponential information growth[C]// L'Aquila. Proceedings of SSGRR, 2003: 73-79.

[2] 邵必林,吴宝江,边根庆. 基于 ISCSI 技术的 SAN 应用研究[J]. 西安建筑科技大学学报:自然科学版,2009,41(1):112-116.
SHAO Bi-lin, WU Bao-jiang, BIAN Gen-qing. Application study of SAN based on ISCSI technology[J]. J. Xi′an UNiv. of Arch. &Tech.: Natural Science Edition, 2009, 41(1): 112-116.

[3] ZHAO Xiao-nan, LI Zhan-huai, ZENG Lei-jie. A Hierarchical storage strategy based on block-level data valuation [C]// The proceeding of the 4th Internatimal conference on Net-worked Computing and Advanced Information Management, 2008: 36-41.

[4] 冯　泳,张延园. 数据迁移在 SAN 中性能优化的研究和应用[J]. 计算机工程,2005(4):43-45.
FENG Yong, ZHANG Yan-yuan. Research and application of data migration in storage area network performance optimiza-tion [J]. Computer Engineering, 2005(4): 43-45.

[5] 唐　竟. 基于信息生命周期管理数据迁移技术研究[D]. 长沙:湖南大学,2009.
TANG Jing. Research of Date Migration's technology based on the Information Lifecycle Management[D]. Changsha: Hunan University, 2009.

[6] JEONG J, DUBOIS M. Cost-sensitive Cache Replacement Algorithms[C]// Proceedings of the Symposium on High-Performance Computer Architecture (HPCA), 2003(1): 327-337.

[7] Smitha and A. L. N. Reddy. LRU-RED: An active queue management scheme to contain high band-width flows at congested routers[J]. In Proceeding of Global Telecommunications Conference, 2001.

[8] B. Reed and D. D. Long. Analysis of caching algorithms for distributed file systems[J]. Operating Systems Review, 1996, 30(3): 12-21.

[9] 吕　帅,刘光明,徐　凯,等. 海量信息分级存储数据迁移策略研究[J]. 计算机工程与科学,2009,31(A1):163-167.
LÜ Shuai, LIU Guang-ming, XU Kai, et al. Research on the data migration strategy of hierarchical mass storage system[J]. Computer Engineering&Science, 2009, 31(A1): 163-167.

[10] 江　菲,汤小春,张　晓,等. 基于价值评估的数据迁移策略研究[J]. 电子设计工程,2011,19(7):11-13.
JIANG Fei, TANG Xiao-chun, ZHANG Xiao, et al. Research of data migration strategy based on data valuation[J]. Electronic design engineering, 2011, 19(7): 11-13.

(下转第 451 页)