

低能耗磁光混合归档系统的设计与实现

缪嘉嘉¹, 付印金¹, 余沛毅¹, 毛捍东²

(1. 解放军理工大学 指挥信息系统学院, 江苏 南京 210007;

2. 北京普世时代科技有限公司, 北京 100192)

摘 要: 层次型混合归档系统是数据存储领域的研究热点, 在工业界也被广泛接受, 小到个人存储大到数据中心都在使用混合存储系统。针对大数据中心的能耗问题, 引入更为廉价低能耗的光介质存储, 建立阵列、在线光盘库、离线光盘库构成的混合归档系统, 采用低能耗磁光混合的存储架构, 在牺牲陈旧文件读取速度的情况下, 大幅降低了存储能耗; 在研究分析数据的一致性保证机制以及多级存储系统的弹性设计的基础上, 针对光介质的读写特性, 重点研究了磁光混合归档系统的文件缓存和预取机制。采用基于整体访问频率的数据迁移策略解决了热文件的访问效率不受影响的问题, 采用基于 I/O 特征预测模型的预取算法, 提升了多级存储结构的命中准确度。实验结果表明, 所构建的系统能够有效节省能源并可维持数据检索查询的时效性。

关键词: 低能耗; 光盘库; 磁光混合; 多级存储系统; 文件预取; 文件缓存

中图分类号: TP302

文献标识码: A

文章编号: 1673-629X(2017)08-0052-05

doi: 10.3969/j.issn.1673-629X.2017.08.011

Design and Realization of Energy-efficient Hybrid Magneto-optical Filing System

MIAO Jia-jia¹, FU Yin-jin¹, YU Pei-yi¹, MAO Han-dong²

(1. Institute of Command Automation, PLA University of Science and Technology, Nanjing 210007, China;

2. Pushtime Technology Inc., Beijing 100192, China)

Abstract: Hierarchical hybrid archiving system is a research hotspot in the field of data storage and is also widely accepted in the industry. The hybrid storage systems have been used by not only personal storage but also the data center. In order to solve the problem of energy consumption in large data center, a hybrid archiving system composed of array, online optical disk library and off-line optical disk library with low cost and low energy consumption has been introduced and low energy consumption magneto-optic hybrid storage architecture has been adopted. Based on the study of data consistency guarantee mechanism and the elastic design of multi-level storage system, the optical read/write characteristics of optical media has been investigated as well as the characteristics of magneto-optical hybrid. The file cache and prefetch mechanism of the archiving system has been adopted and the data migration strategy based on the overall access frequency is adopted to solve the problem that the access efficiency of the thermal file is not affected. The prefetching algorithm based on the I/O characteristic prediction model has also been adopted and the multi-level storage structure has been improved. The experimental results show that the proposed system can effectively save energy and maintain the timeliness of data retrieval query.

Key words: energy-efficient; optical disk library; hybrid magneto-optical; multilevel storage system; file prefetching; file caching

1 概述

随着数据量的增长以及人们对于数据价值的深刻认知, 归档系统的高并行性、高可靠性、高性价比变得越来越重要。然而在建、在用的数据中心, 电力的消耗越来越严重, 数据中心的能耗成本还在不断增加。早在 2006 年, Jonathan 等^[1]认为美国数据中心能耗占到

了该国总能耗的 1.2%, 且其增长速度大约为 5 年翻一番; William 等^[2]估算的数据中心能耗密度范围为 1 076 ~ 2 150 W/m²。国内数据中心规模呈快速增长趋势, 数据中心能耗也随之快速增加。2009 年, 国内数据中心总耗电量约 364 亿 kWh, 占当年全国总耗电的 1%。未来, 国内数据中心仍将快速发展, 如果维持

收稿日期: 2016-09-07

修回日期: 2016-12-22

网络出版时间: 2017-06-05

基金项目: 国家自然科学基金资助项目(61402518); 总装预研基金(9140A15070414JB25224)

作者简介: 缪嘉嘉(1980-), 男, 博士, 高级工程师, 研究方向为数据处理、数据安全。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20170605.1510.080.html>

当前的低能效水平,到2015年,仅全国的数据中心就将消耗掉三峡电站1年的发电量^[3]。

因此,加强数据中心节能、提高数据中心能效是必要和紧迫的。北京、上海、南京等地均有相关的实际数据采集^[4-5],分析后发现,IT及网络通信设备的能耗占51%,空调制冷系统的能耗占24%,空调通风加湿系统的能耗占11%,照明能耗占2.5%,其他能耗占11.5%,因此数据中心机房的节能重点是IT及网络通信设备和机房空调。国内研究者也进行了相关的能耗分析及节能措施,主要手段是调整机房的物理结构,采用低能耗设备等等^[6-7]。

一方面,为节约数据中心能耗,干福熹院士携手国内外20多位院士联合署名,倡议为迎接大数据的挑战,应该开展安全、节能和长寿命的光存储技术研发和应用。另一方面,据行业调查分析显示,归档系统中无论何时都有70%~80%的数据是静止不动的^[8]。数据不同时期有其存在的不同意义:数据刚生成时,访问频率最高;随着时间的推移,访问频率降低,低访问频率的数据量远远超过高访问频率的数据量。将这部分数据称为“冷”数据。“冷”数据由于访问频率降低,如果在归档系统中依然在线存储,这是对能耗的极大浪费。综上所述,节约能耗是数据中心规划建设运行过程中不可忽视的重要一环,采用低速的光介质设备能够降低能耗,但带来了访问效率低下的问题。为此,提出了一种磁光混合归档系统,采用高速介质缓存方法,能够在降低能耗的情况下保障数据的访问速度在可接受范围。

2 相关技术

混合存储系统(Hybrid Storage System)通常是指在闪存技术飞速发展的背景下出现的一种集固态硬盘和磁盘驱动器技术于一体,以大容量、高性能和低成本为目标的异构性非易失归档系统。其设计思想在于使性能好、价格高的SSD在归档系统中发挥杠杆作用,发挥SSD和HDD的各自优势并弥补对方的短处,让系统以接近磁盘的价格提供近似固态硬盘的性能。

所提出的磁光混合归档系统在缓存技术、预取技术方面借鉴了现有混合存储系统中的现有研究成果,针对光盘本身的I/O特性进行了相应调整。

2.1 缓存技术

Cache技术被广泛地运用于多层存储体系结构中,通过程序局部性原理将I/O集中于高性能存储层,从而弥补不同层次存储器之间性能和价格的差异,实现以低购置成本得到高性能的设计目标。

已有的缓存算法研究多基于磁盘存储和DRAM Cache,并针对磁盘的内部特征进行了大量优化,比如

尽量以顺序方式访问磁盘、让磁盘空闲时段延长等。近期,缓存技术被移植到基于Flash、磁盘的混合归档系统,针对Flash介质的独特特性,如有限的擦写(Program/Erase,P/E)次数、不对称的读写性能(Asymmetric Read and Write)等问题,也有研究跟进。

此外,在以往基于磁盘的DRAM Cache中,命中率(Hit Rate)是最主要的Cache性能指标。而在混合归档系统中,无论是Flash层之上的DRAM Cache,还是磁盘层之上的Flash Cache,缓存算法的评价指标都将变得更为复杂。Intel公司的Matthews等^[9]指出,仅当一个请求完全命中Flash Cache(Full Hit)时,才能减少磁盘访问,若请求部分命中(Partial Hit)Flash Cache并不意味着系统性能的必然提升。CFLRU^[10]指出,对于Flash存储之上的DRAM Cache,脏页(Dirty Page)替换的代价要高于干净页(Clean Page)替换。在DRAM Cache中数据替换的代价可以忽略不计,然而OP-FCL^[11]指出,在Flash Cache中数据替换的代价很高,必须要将数据在Flash Cache中的写入时间和被替换数据的垃圾回收时间考虑进去。

因此,磁光混合归档系统的缓存管理技术的设计,必须针对光盘的内部特征重新量化Cache的成本收益(Cost-benefit)模型,建立Cache插入策略。

2.2 预取技术

预取技术应用的领域非常广泛,包括处理器、Web系统结构、数据库、文件系统、存储控制器等。在归档系统中,应用最广泛的预取技术是顺序预取(Sequential Prefetching),即通过顺序流侦测来预测未来的请求模式。顺序预取之所以被普遍采用源于其所需语义简单、预取精度高,且I/O成本低^[12]。现有的顺序预取方案主要分为三大类,即持续预取(Prefetch Always, PA)、缺失预取(Prefetch On Miss, POM)和命中预取(Prefetch On Hit, POH)^[13]。PA型预取并不需要预测模块,对每一个请求它都会预取与之连续的数据。

Gill等提出了AMP^[14]预取算法,通过渐进性的启发式策略来不断调整预取的强度和触发器(trigger)位置,从而获取最高的聚合吞吐量。一些研究建议把预取、缓存和调度权限交给应用程序来控制。还有一些研究提出不去修改应用程序的代码,而是通过特殊的方式执行应用程序来分析该预取哪些数据。这些方法都涉及到I/O接口的修改、应用的重构和一些复杂计算。

在混合存储系统中,异构介质的存储设备构成了多层缓存系统。多层缓存系统有其不同于传统缓存的特点。伊利诺斯大学的Zhou等^[15]指出,在第一层Cache中往往使用基于局部性的Least Recently Used(LRU)替换算法,因而访问第二层Buffer Cache的访

问体现出较第一层相对更弱的时间局部性。此外,FAST^[16]使用基于固态盘的预取策略来加快个人电脑中程序的启动速度。该系统是将 SSD 中的数据预取到 DRAM 缓存中,并非将磁盘数据预取到 SSD 中。

2.3 蓝光相关技术指标

以硬盘和磁带为代表的磁存储技术,由于存储速度快、存储量大和使用方便,成为当今主流的存储技术,被广泛应用于数据中心乃至企业中。现有主流的存储技术难以满足大数据时代对海量数据长期、安全、高效存储的要求。蓝光盘利用波长较短的蓝色激光读取和写入数据,极大地提高了光盘的存储容量。光存储的主要优势有三个:一是基盘由坚固、耐久的材料制成;二是光存储的非易失性;三是可长期保存。光存储技术发展至今,其安全、能耗低、寿命长和单介质数据容量增加快的特点,使之在大数据时代满足对数据长期、安全、高效存储需求上具有独特的优势。

不同存储介质的特性对比见表 1。

表 1 不同存储介质的特性对比

参数/类型	磁盘阵列	磁带库	蓝光光盘库
访问速度(在线时)/ms	10	7 000	1 000
数据传输速度/(MB/s)	400	140	216
数据读写方式	接触	接触	非接触
保存时间(介质)	3 年	5 年	50 年以上
数据信赖性	有数据丢失的风险	保存条件差时,数据无法读取	不怕电磁干扰,没有误删除风险,可防病毒软件破坏
环境要求	温度 14 ~ 24 ℃ 湿度 20% ~ 40%	温度 16 ~ 25 ℃ 湿度 20% ~ 50%	温度 10 ~ 40 ℃ 湿度 20% ~ 80%
介质保管时的空调	必要	必要	不需要
耗电量 (30 年)/kWh 保存	100 TB 108 000	3 500	3 200
维护成本	高	高	低

3 磁光混合归档系统

磁光混合归档系统主要由离线盘柜、光盘库、存储阵列、服务器组成,根据数据的访问速度将存储阵列中的称为在线数据,光盘库中的为近线数据,离线盘柜的称为离线数据,服务器中的元数据服务存储元数据组织信息,具体如图 1 所示。

上述存储架构中客户端主动或被动将数据移动至归档服务器的阵列中,通过 API 接口或 Web 接口可以完成对已归档数据的使用,若阵列中数据已满或有部分数据长久不被访问,那么逐步迁移至近线存储即光盘库中。其中离线设备需要通过人工干预才能进行数据访问,因此不在讨论范畴之内。

图 2 是磁光混合多级存储的体系结构,对用户端

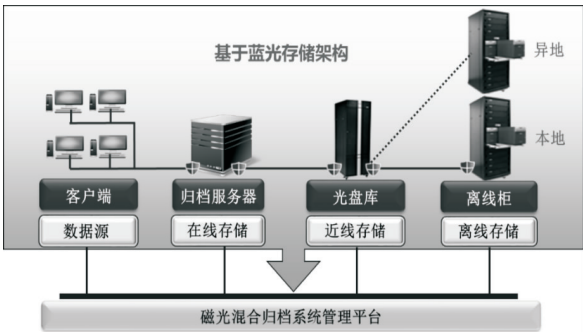


图 1 基于蓝光存储架构

系统支持客户端 API,允许以服务方式提供数据的查询、访问,对数据源接口采用归档计划、任务方式,设置定期的归档时间,自动或手动方式完成数据归档,非结构化数据通过数据预处理进入元数据服务器,结构化数据通过 ETL 工具完成关系型数据到面向对象数据结构的转换,并加载到元数据集群中。

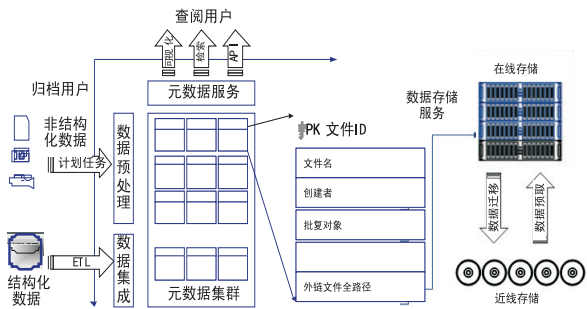


图 2 磁光混合多级归档的体系结构

元数据集群中存储面向对象的文件属性结构,以值对方式记录对象的属性,存在外部文件链接,指向对象实体,对象实体可以是文本、图片、视频等富媒体方式文件。元数据库采用 ES 架构,一方面利用 ES 本身的易扩展性、高可靠性等特点,能够纵向或横向进行节点扩展,ES 本身也能够很好地支持全文检索。

通过上层的元数据服务,可以支持查阅用户进行全文检索、数据可视化展示,也允许用户利用 API 接口与其他应用程序对接。

文档数据进入归档服务器采用光盘存储结构按照蓝光盘片大小进行组织目录,允许一个蓝光盘片中存储多个任务数据,也允许一个任务数据横跨多个蓝光盘片。归档数据采用新的组织结构的原因在于,便于与近线存储进行迁移,并且归档系统中不再关注文档的物理路径存放,可以通过元数据的再组织,形成逻辑视图供用户查阅。

数据迁移:在数据量不超过在线存储容量的情况下,所有数据以光盘大小划分组织目录,对外提供在线的数据检索,随着数据容量的增大,在线容量不能满足归档需求时,系统将访问量较小的数据内容,开始向光盘库进行迁移,访问量的统计单位是每个光盘上数据

的访问统计,而不是以单个文件的访问来进行核算。

数据预取:通过较长时间的使用,归档系统的数据根据使用情况产生了不同情况的分布,基本可以确定的是在线存储基本处于 80% 使用状态,如果发生用户访问到近线存储的文件,需要调度光盘库将光盘内容写入在线存储,普遍想法是将用户指定的读取文件写入即可,在这种访问速度上,访问时间从秒级下降到分钟级。

3.1 数据一致性机制

数据归档系统属于分布式架构,必然存在一致性保证问题。该系统有两处隐患,一是元数据存储,元数据底层采用分布式架构,允许多台设备存储冗余存储元数据,使得系统能够负载均衡和容错;二是文件副本可以分布在在线存储的缓冲区,也可以存储于近线存储的光盘介质中^[17]。

系统的元数据集群可采用横向扩展,通过增加节点来传播负载和增加可靠性,如图 3 所示,其中外围方框标识节点,带星号的为主节点,小正方形表示分片。节点是运行的元数据实例。一个集群是一组具有相同节点的集合,节点间协同工作、共享数据并提供故障转移和扩展功能,当加入新节点或者删除节点时,集群就会感知到并自动平衡数据。集群中一个节点会被选举为主节点,用来管理集群中的一些变更,例如新建或删除索引、增加或移除节点等。任何一个节点互知道数据存在于哪个节点上,可以转发请求到外部需要数据所在的节点上,主节点负责收集各节点返回的数据,最后一起返回给客户端。当元数据集群扩容或缩小,系统将会自动在节点间迁移分片,以使集群保持平衡。



图 3 元数据横向扩展架构

对于第二点,该混合归档系统不支持数据文件本身改变,在进入近线存储,即进行光盘刻录后,不支持数据文件的改写,因此不涉及文件副本的不一致问题。

3.2 系统弹性设计

归档系统的元数据和文件数据分离存储。在数据一致性机制中提及元数据的存储采用易于扩展的 ES 架构,元数据中包括全文索引数据可能会大于原数据文件,但是通过增加处理节点,一方面可以增加实际容量,另一方面也可以提升并发能力。从元数据角度来看,系统具备较好的扩展性。从实际数据文件存储上来看,保持在线存储和近线存储的容量比例不变,同比扩充增加存储容量,不会导致数据迁移或数据预取的性能损耗,因此在数据文件的存储上,系统也具备较大弹性。

4 磁光混合归档系统关键技术

该混合归档系统采用蓝光存储作为二级存储介质,采用阵列作为一级存储介质,将元数据信息存放在一级存储介质上,确保信息检索速度,在数据量超过一级存储容量时会产生数据迁移和数据预取需求。

4.1 基于整体访问频率的数据迁移策略

文件访问频率是文件迁移的重要检索条件。文件迁移的最小单位为光盘,在线存储中已经按照光盘盘片大小划分了存储组织,这样能够保证文件在迁移到近线存储时不改变文件的组织结构。访问频率低的光盘数据,会被迁移到近线存储。如可能设定如下策略:将一年内访问次数少于 10 次的文件从一级存储迁移到二级存储^[18]。

定义 BD 表示某个盘片数据的被访问次数,盘片数据中存在 n 个文件, A_n 为第 n 个文件的被访问次数,则有:

$$BD = \alpha * \sum_{i=1}^n A_i + \beta * \text{MAX}\{A_n\}$$

其中, α 表示访问和的权重, β 表示最大访问次数的权重,两者取值范围均为 0 ~ 1。如果用户读取归档数据时侧重于突发性读取,那么 $\alpha < \beta$,如用户突发读取后,基本会采用顺序读取获取周围的数据文件,那么 $\alpha > \beta$ 。

IBM 在 STEPS 架构中提出了 Policy Cache 的概念, Policy Cache 可以看作为一个三元组的表,其中包含策略号 (Rule Number)、策略预期执行时间 (Time)、文件 iNode 唯一对应的文件对象号 (file object ID)。

在磁光混合归档系统中借鉴 Policy Cache 的思想,将记录下整个文件系统的文件完整路径名,数据类型,数据创建时间,最后修改时间以及文件访问频率信息记录在 Policy Metadata Container (PMC) 中,根据 (R, D, T) 从 PMC 中查询得到属于该策略的数据分类文件的应用导向和程序导向的元数据,作为该策略的元数据库,即相应策略的 Policy Cache。

系统采用过滤驱动技术记录文档的访问次数、访问时间。

4.2 基于 I/O 特征预测模型的预取算法

文件预取技术中,如何提高文件预取的命中率和适用度一直是研究的焦点。尤其是在面对大批量数据读取时,如何提高预取命中率对系统的性能提升有着至关重要的影响。提出了识别 I/O 特征的预测模型,该模型通过记录文件的历史访问信息获得 I/O 特征,再分析这些 I/O 访问模式,设计一个简单高效的特征符号表来表示这些模式。此预测模型可有效地识别出顺序读、固定点读、逆序读、跳读、多步跳读等多种模式。同时,该模型添加应用程序的信息,可有效地对不

同程序之间的交叉读做出预测,有很高的预测命中率。

踪迹模块捕获应用程序的外存数据 I/O 操作,构建 I/O 访问信息流,提供特定 I/O 访问操作的查询功能;模式识别模块根据踪迹模块捕获的应用程序 I/O 访问信息流识别应用程序的 I/O 访问模式。可以支持顺序读、固定点读、逆序读、单步跳读、多步跳读等模式;数据预取模块提供一些预取库函数,完成顺序读、固定点读、逆序读、单步跳读、多步跳读等模式数据块的预取工作。文件预取框架图如图 4 所示。

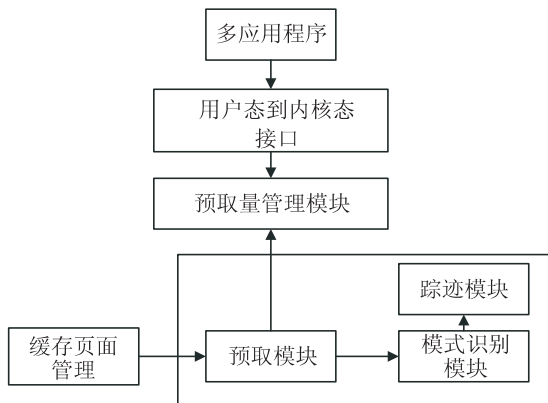


图 4 文件预取框架图

当有读线程的时候,先判断数据是否在缓存中。如果在,则直接从缓存中取数据;否则向系统发出读磁盘的请求,此时,判断是否在 stable 状态,如果在,则根据 I/O 特征表的一些信息预测下次读请求的 offset 和 size 并预取到缓存中。

5 结束语

针对不断增多的数据中心建设,关注度持续走高的能耗问题,系统设计多级存储架构,拟在牺牲数据访问效能的基础上大幅降低数据中心能耗。该系统引入更为廉价低能耗的光介质存储,建立了由阵列、在线光盘库、离线光盘库构成的混合归档系统,理论上当在线存储与近线存储容量为 1:9 时,能够节省 90% 的能量损耗,而在辅以文件缓存和预取机制的基础上,文件的突发读写在 20% 情况下会造成分钟级等待,但大部分情况或是顺序读取时,系统能够恢复在线查询效率。实验证明,该系统能够有效节省能源并维持数据检索查询的时效性。

参考文献:

[1] Koomey J. Estimating total power consumption by servers in the u. s. and the world[R]. Berkeley: Lawrence Berkeley National Laboratory, 2007.

[2] Tschudi W, Xu Tengfang, Sartor D, et al. Energy efficient data centers[R]. Berkeley: Lawrence Berkeley National Laboratory, 2003.

[3] 谷立静,周伏秋,孟辉.我国数据中心能耗及能效水平研究[J].中国能源,2010,32(11):42-45.

[4] 黄森,潘毅群.上海某数据中心能效调研分析[J].制冷与空调,2011,25(2):208-211.

[5] 林明,刘振安,李彤.北京电信 IDC 机房网络机柜的节能分析[J].邮电设计技术,2012(5):75-79.

[6] 柳运昌,杨二瑞,许建霞.面向云数据中心的能耗管理[J].电信科学,2012,28(12):96-102.

[7] 田宝华,蒋句平,李宝峰,等.基于统一资源管理的超级计算机系统节能方案[J].计算机应用,2012,32(3):835-838.

[8] He Mei, Xing Ling, Li Guo. A data migration strategy for HSM based on data value[J]. Journal of Information & Computational Science, 2011, 8(2): 312-317.

[9] Matthews J, Trika S, Hensgen D, et al. Intel turbo memory: nonvolatile disk caches in the storage hierarchy of mainstream computer systems[J]. ACM Transactions on Storage, 2008, 4(2): 1-24.

[10] Park S Y, Jung D, Kang J, et al. CFLRU: a replacement algorithm for flash memory[C]//Proceedings of the 2006 international conference on compilers, architecture and synthesis for embedded systems. Seoul, Korea: ACM, 2006: 234-241.

[11] Oh Y, Choi J, Lee D, et al. Caching less for better performance: balancing cache size and update cost of flash memory cache in hybrid storage systems[C]//Proceedings of the 10th USENIX conference on file and storage technologies. San Jose, CA: USENIX, 2012: 25.

[12] Yang L, Feng W. SoAP: a strip-oriented asynchronous prefetching for improving the performance of parallel disk systems[C]//Proceedings of the high performance computing and communication. [s. l.]: [s. n.], 2012: 96-103.

[13] Li M, Varki E, Bhatia S, et al. TaP: table-based prefetching for storage caches[C]//Proceedings of the 6th USENIX conference on file and storage technologies. San Jose, CA: USENIX, 2008: 1-16.

[14] Gill B S, Bathen L A D. AMP: adaptive multi-stream prefetching in a shared cache[C]//Proceedings of the 5th USENIX conference on file and storage technologies. San Jose, CA: USENIX, 2007: 26.

[15] Zhou Y, Chen Z, Li K. Second-level buffer cache management[J]. IEEE Transactions on Parallel and Distributed System, 2004, 15(6): 505-519.

[16] Joo Y, Ryu J, Park S, et al. FAST: quick application launch on solid-state drives[C]//Proceedings of the 9th USENIX conference on file and storage technologies. San Jose, CA: USENIX, 2011: 19-39.

[17] 丁海骏,卢菁.云环境下元数据弹性分级一致性保障机制研究[J].计算机应用研究,2016,33(7):2039-2042.

[18] 周斌,汪浪,张莹,等.基于数据块级迁移策略的设计与实现[J].计算机工程与设计,2016,37(7):1822-1826.