第 39 卷 第 4 期
Vol.39 No.4

计 算 机 工 程
Computer Engineering

2013 年 4 月
April 2013

• 先进计算与数据处理 • 文章编号：1000—3428(2013)04—0032—04 文献标识码：A 中图分类号：TP399

基于固态硬盘的云存储分布式缓存策略

李东阳，刘 鹏，丁 科，田浪军

(解放军理工大学指挥自动化学院，南京 210007)

摘 要：为满足海量数据存储的需求，提出一种基于低功耗、高性能固态硬盘的云存储系统分布式缓存策略。该策略对不同存储介质的硬盘虚拟化，将热点访问数据的缓存与存储相结合，实现在不同存储介质之间的热点数据迁移，解决热点元数据的访问一致性与存储服务器的动态负载均衡问题。工作负载压力测试结果表明，该策略可使云存储系统的读峰值速率最高提升约 86%，并且能提高存储服务器的吞吐量。

关键词：云存储；固态硬盘；分布式缓存；热点数据；数据迁移；负载均衡

Distributed Cache Strategy in Cloud Storage Based on Solid State Disk

LI Dong-yang, LIU Peng, DING Ke, TIAN Lang-jun

(Institute of Command Automation, University of Science and Technology of PLA, Nanjing 210007, China)

【**Abstract**】In order to meet the needs of mass data storage, this paper proposes a distributed cache strategy in cloud storage based on Solid State Disk(SSD) and low power consumption. This strategy achieves the virtualization among different storage mediums, combines the cache and storage of the hot spot data together, and migrates the hot spot data among different storage mediums. The strategy also resolves the problems of the consistency of the hot spot metadata and the dynamic load balancing among the storage servers. Load pressure test indicates that the distributed cache strategy in cloud storage system based on SSD can improve the read peak rate by about 86%, and improves the system throughput effectively.

【**Key words**】cloud storage; Solid State Disk(SSD); distributed cache; hot spot data; data migration; load balancing

DOI: 10.3969/j.issn.1000-3428.2013.04.008

1 概述

为提高存储系统的容量、提升 I/O 速度、削减存储成本、降低读写时延、应对海量存储的需求，云存储[1]架构的思想应运而生。在针对海量数据共享的数据密集型应用的管理过程中，对海量数据的读操作是大量的、经常的，而且实时性要求更高，虽然围绕内存的随机存储器(Random Access Memory, RAM)缓存技术在一定程度上可以缓解硬盘读写操作速率不高带来的压力，但受限于缓存空间的容量和处理页面置换导致的开销瓶颈的制约，硬盘响应性能也受到一定程度的影响。

传统的串行高级技术附件(Serial Advanced Technology Attachment, SATA)硬盘存储空间大、价格低，一直以来作为主流的存储介质被大量使用，虽然已经出现了一系列的优化算法，但由于其机械臂的寻址方式，制约了读写性能的提升。近年来，出现以固态硬盘(Solid State Disk, SSD)为代表的新型存储介质，随着生产工艺的进步与生产成本的降低，SSD 无机械磁头读写，并发随机访问响应速度快，有利于最大程度提升服务器的负载能力，而且具有发热小、功耗低、无噪音等诸多优势，越来越得到推广应用。文献[2]提出一种在文件系统级实现的 SSD 多级存储的启发式优化算法，开机与程序启动时间与单独使用 SSD 相似，同时又有容量大、价格低的硬盘驱动器(Hard Disk Drive, HDD)可供使用，但它是针对单机使用的。文献[3-4]采用 SSD 硬盘缓存数据，但实际可用存储空间仍然是磁性存储介质的容量。文献[5-6]针对 SDD 硬盘在存储上的突出性能，结合混用 HDD 与 SSD 的优势，分别提出 hybridFS、Hystor 文件系统。文献[7-9]分别研究了把 SSD 闪存作为非易失性缓存进行应用的不同方法，其目的是把待访问数据块缓存在 SSD 闪存中，从而降低磁盘转速、减少功耗。从目前应用

基金项目：工业和信息化部电子信息产业发展基金资助项目“云安全存储系统关键技术研究”(KYHXRJ021101)

作者简介：李东阳(1985－)，男，硕士、CCF 会员，主研方向：分布式系统，网格计算；刘 鹏，教授、博士生导师；丁 科，讲师、博士；田浪军，硕士

收稿日期：2012-06-11 **修回日期**：2012-07-03 **E-mail**：li_ndows@126.com

情况看，SSD 主要应用在单机上提高读写性能，应用范围小。而在云存储的应用背景下，对其研究相对较少，考虑到容量与价格因素的制约，不可能全部采用 SSD 硬盘的云存储解决方案。例如，文献[10]通过在软件层面对面向存储服务的缓存管理模型进行优化，从而达到有效减少系统访问延迟、增加吞吐量、提高分布式环境下系统性能的目的。

本文考虑到 SSD 应用的前景与优势，在兼顾云存储平台提供商的运营成本与服务质量的前提下，设计并实现一种基于 SSD 的云存储分布式缓存策略，在满足数据可靠性的前提下，实现对不同存储介质的虚拟化，并研究在分布式的云存储平台下热点数据分布式缓存的迁移与动态负载均衡问题。

2　分布式缓存技术的设计与实现

2.1　硬盘虚拟化

用户空间文件系统(Filesystem in Userspace, FUSE)是 Linux 中用于挂载某些网络空间到本地文件系统的模块，是完全在用户态实现的文件系统，提供实现 FUSE 的用户态目录与文件操作的接口。在云存储系统中，通过 Linux 提供的 FUSE，采用硬盘虚拟化技术，既对不同的存储介质进行有效管理，又屏蔽了客户应用对基于 SSD 硬盘分布式缓存的云存储系统的访问细节，如图 1 所示。

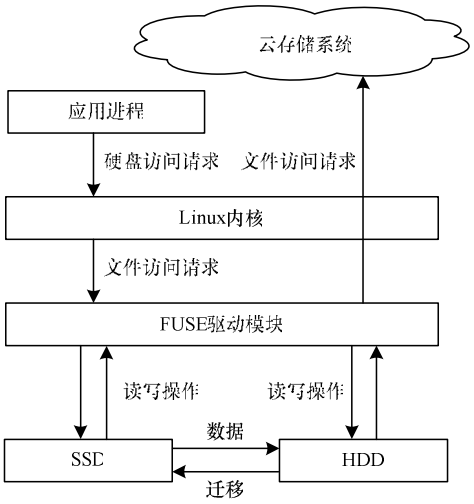

图 1　硬盘虚拟化

SSD 既作为 RAM 内存与 HDD 之间的缓存区，用于保存从 HDD 上置换的缓存区热点数据，又作为存储服务器的存储介质，在操作过程中存储热点数据，从而充分利用其读写性能与容量空间带来的性能提升，避免 SSD 单纯作为缓存使用造成的成本压力与空间浪费。

2.2　HDD 热点数据监控与元数据一致性管理

在数据服务器端，考虑到云存储系统是在数据块级对数据进行管理而非在文件级进行数据管理，在内存中对 HDD 与 SSD 各自的数据块元数据分别进行管理，而内存中的元数据链表保存 2 种不同硬盘介质各自对应数据块的状态信息，数据服务器根据这些状态信息对元数据进行管理。另外，为提高元数据在用户访问过程中与对应数据块的映射速度及命中率，元数据链表采用 hash 函数进行管理。

在对热点数据进行监控时，为减少云存储系统的开销，采用具体方法是：元数据中设有一个监控变量 accessfrequency，用于对数据块的访问行为进行监控，并记录数据的冷热程度。对于 HDD，每当数据块被成功访问，则更新其对应元数据链表 hddhashtable 上相应数据块的状态信息，并将该数据块的 accessfrequency 变量加 1。对于 SSD，在内存中管理一个待迁出元数据链表 migratehashtable，设存储服务器上所有数据块的元数据的集合为 Ω，hddhashtable 上的元数据集合为 Ωh，migratehashtable 上的元数据集合为 Ωm，SSD 硬盘元数据链表 ssdhashtable 上的元数据集合为 Ωs，则 $\Omega h \cup \Omega h = \Omega$、$\Omega m \subseteq \Omega s$。当系统初始化时，所有的 hash 表需要加载到内存。为了解决元数据一致性的问题，数据服务器周期性地遍历 SSD 元数据链表，并采用最近最少使用(Last Recently Used, LRU)算法进行判断，出现冷数据，则将其元数据加入到待迁出元数据链表 migratehashtable，但此时既未删除数据块，也未真正删除在 ssdhashtable 上对应数据块的元数据，只有当 SSD 可用空间容量不满足最小缓存需求的临界条件时，将数据回迁至 HDD 上，并做相应的修改。以图 2 元数据读操作为例，简要描述元数据管理的流程。

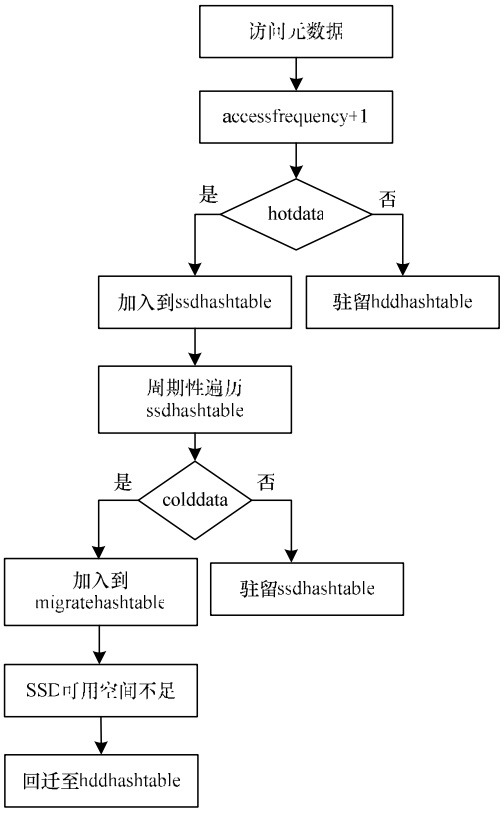

图 2　元数据读操作

2.3　分布式缓存的动态负载均衡

在云存储系统中，对数据进行冗余备份处理的方式一般有 2 种：文件副本，纠删码。针对数据密集型应用，为

了降低处理纠删码数据所带来的读写性能影响，采用文件副本的机制，在此前提下，为解决因数据分布过于集中造成的存储压力，进而避免读写操作过热造成数据服务器端的访问压力，不仅需要综合考虑容量方面的均衡，同时还要兼顾访问均衡的需求，而容量方面的均衡既包括 HDD 的容量均衡，又涉及到空间相对有限的 SSD 容量均衡。同时，当 SSD 硬盘可用空间不足 20%时，还要释放一部分存储空间，作为应对突发集中访问情况下的数据服务器端的分布式缓存区。

由于副本保存在不同数据服务器上，因此管理服务器依靠数据服务器定期发送的主动式心跳报文进行管理，管理服务器根据加权的数据服务器的可用容量信息选择用户访问的数据服务器节点，进而对容量负载进行平衡。

对于存储服务器节点 Si 的负载能力 $L(Si)$，考察目标有：HDD 有效空间大小 $HS(Si)$，SSD 有效空间大小 $SS(Si)$，HDD I/O 速率 $HV(Si)$，SSD I/O 速率 $SV(Si)$，内存容量 $M(Si)$，计算如下：

$$L(Si)=\Sigma(\omega 1\times HS(Si),\omega 2\times SS(Si),\omega 3\times HV(Si),\\ \omega 4\times SV(Si),\omega 5\times M(Si))$$

其中，ωi 为负载能力影响因子，且 $\sum\omega i=1$，存储服务器的负载匀衡情况与 SS、SV 呈正比，与 HS、HV 呈反比。负载能力影响因子的各个元素值分别取决于对应指标在不同应用环境下对 $L(Si)$的影响程度。

主服务器的访问负载均衡算法采用权重轮询均衡算法(Weight Round-Robin Scheduling)，根据存储服务器的不同负载能力，给每个存储服务器分配不同的权值，根据不同的权值，选择在分布式环境下可以接收服务请求的若干台存储服务器以响应用户需求，达到负载访问均衡的目的。

3 性能测试与分析

为检验基于 SSD 的云存储分布式缓存的性能，对其进行工作负载压力测试，测试平台包括 1 台管理服务器、5 台存储服务器、16 台客户端。各主机均通过千兆网卡连接到华为 3Com(H3C)SOHO-S1224R 24 口千兆机架交换机上，存储服务器配置如表 1 所示，测试工具采用 I/OMeter。

表 1 云存储服务器配置

存储介质	容量	理论 I/O 速率/(MB·s^{-1})
WD10EARS	1 TB	108/107
Kingston V+200	120 GB	535/480
KVR1333D3N9	2×2 GB	8 118/6 528

图 3 是用 IOMeter 测得的基于 HDD 硬盘的云存储系统在不同客户端并发访问时的读写带宽情况。图 4 是使用 SSD 分布式缓存技术的云存储系统在不同客户端并发访问时的读写带宽情况。由于 SSD 容量空间受限，目前仍无法直接存储海量数据，因此没有进行基于 SSD 的云存储系统测试。

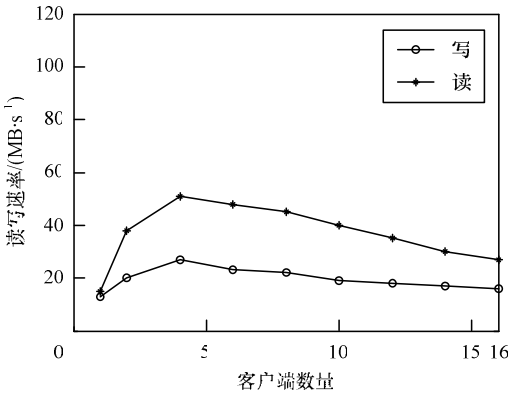

图 3 HDD 云存储系统读写带宽

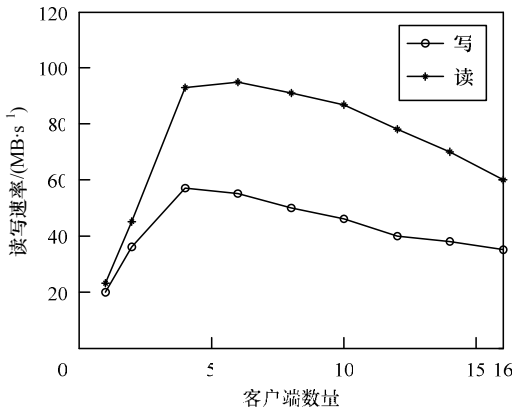

图 4 SSD 分布式缓存云存储系统读写带宽

从图 3 可以看出，随着用户访问的增多，特别是在用户大规模并发访问的情况下，无论是读速率还是写速率，云存储系统的性能下降幅度都表现的比较明显，性能恶化比较严重。分析结果表明，这主要是由于内存空间有限，无法缓存更多数据，以及受到 HDD 的物理特性影响所决定的，多个用户集中访问时，对同一个存储服务器并行读写数据，导致存储服务器磁盘频繁调度，从而影响系统整体性能。云存储系统的吞吐量受到较大的影响。

从图 4 可以看出，在采用 SSD 的分布式缓存云存储系统中，通过对比可知，系统性能得到显著提升，不仅在用户较少时对用户访问有很好的支持，随着用户访问的增多，系统对大规模访问也能进行很好支持，由于采用了分布式缓存技术，特别是存储服务器的读峰值速率，最高可以提升约 86%，系统对海量数据的写支持也有很好改善，最高写速率目前可以达到 55.2 MB/s。一方面是由于 SSD 寻址响应时间短、读写速率快，在硬件方面可以对系统进行加速；另一方面，SSD 对热点数据的缓存缓解了用户访问时读写速率的制约，提高了存储服务器的吞吐量。

千兆机架交换机 1 Gb/s 的链路带宽是固定的，因此，作为缓存的 SSD 容量大小在某种程度上决定了云存储系统的读写性能提升的空间，而考虑到在实际应用过程中成本因素的影响，SSD 的容量不能无限扩展。因此，对于实际应用的云存储系统，在考虑性价比最优化的前提下，必须合理选用 SSD，根据应用需求设置负载能力影响因子。

4 结束语

本文设计了一种基于 SSD 的云存储分布式缓存策略，在数据密集型应用中，在云存储平台的存储服务器端，利用该策略将热点数据的缓存与实际存储结合起来，能有效提高云存储平台下用户访问的响应需求，控制存储成本。受到 SSD 空间大小的限制，尽管数据密集型应用对写操作的要求不高，目前的副本冗余机制还不能使基于 SSD 硬盘的分布式缓存云存储系统对大量数据的写操作进行有效的支持。因为，为满足不同客户的需要，在副本冗余机制下，如何有效提升对海量写数据的性能支持是下一步的研究方向。

参考文献

[1] 武永卫，黄小猛. 云存储[J]. 中国计算机学会通讯, 2009, 5(6): 44-52.

[2] Payer H, Sanvido M A, Bandic Z Z, et al. Combo Drive: Optimizing Cost and Performance in a Heterogeneous Storage Device[C]//Proc. of the 1st Workshop on Integrating Solid-state Memory into the Storage Hierarchy. Washington D. C., USA: [s. n.], 2009.

[3] Trika S, Hensgen D. Intel Turbo Memory: Nonvolatile Disk Caches in the Storage Hierarchy of Main Stream Computer Systems[J]. ACM Transactions on Storage, 2008, 4(2): 17-25.

[4] Panabaker R. Hybrid Hard Disk and ReadyDrive™ Technology: Improving Performance and Power for Windows Vista Mobile PCs[Z]. 2006.

[5] Jinsun S, Jaechun N. HybridFS: Integrating NAND Flash-based SSD and HDD for Hybrid File System[C]//Proc. of the 10th WSEAS International Conference on Systems Theory and Scientific Computation. Taipei, China: [s. n.], 2010.

[6] Chen Feng, Koufaty D, Zhang Xiaodong. Hystor: Making the Best Use of Solid State Drives in High Performance Storage Systems[C]//Proc. of the 25th ACM International Conference on Supercomputing. Tucson, USA: ACM Press, 2011.

[7] Bisson T, Brandt S A. Reducing Energy Consumption with a Non-volatile Storage Cache[C]//Proc. of IWSSPS'05. San Francisco, USA: IEEE Press, 2005.

[8] Bisson T, Brandt S A, Long D D. A Hybrid Disk-aware Spin-down Algorithm with I/O Subsystem Support[C]// Proc. of IPCCC'07. [S. l.]: IEEE Press, 2007.

[9] Chen Feng, Jiang Song, Zhang Xiaodong. SmartSaver: Turning Flash Drive into a Disk Energy Saver for Mobile Computers[C]//Proc. of ISLPED'06. [S. l.]: ACM Press, 2006.

[10] 王 侃，陈志奎. 面向存储服务的分布式缓存系统研究[J]. 计算机工程, 2010, 36(15): 80-82.

编辑 陆燕菲

(上接第 26 页)

检测速度及准确性都有所提高。本文还找出了模型参数在识别不同对象时的变化规律，可应用于木材单板缺陷图像的多目标识别中。下一步将在该模型的基础上，通过引入新的数学模型项及参数，实现对单板蓝变、腐朽、裂纹等缺陷图像的检测研究，以提高该模型对各种单板缺陷图像的有效识别。

参考文献

[1] 唐利明. 基于 GAC 模型的自适应图像分割算法[J]. 小型微型计算机系统, 2010, 31(6): 1223-1225.

[2] 谢 敏. 基于偏微分方程的彩色火焰图像分割与跟踪[D]. 南昌: 南昌大学, 2009.

[3] 何 宁, 张 鹏. 基于边缘和区域信息相结合的变分水平集图像分割方法[J]. 电子学报, 2009, 37(10): 2215-2218.

[4] 涂 虬, 许毅平, 周曼丽. 基于全局最小化活动轮廓的多目标检测跟踪[J]. 计算机应用研究, 2010, 27(2): 794-797.

[5] Chan T F, Vese L A. Active Contours Without Edges[J]. IEEE Trans. on Image Processing, 2001, 10(2): 266-277.

[6] Caselles V, Morel J M, Sapiro G. Geodesic Active Contours[J]. International Journal of Computer Vision, 1997, 22(1): 61-79.

[7] Rudin L, Osher S, Fatemi E. Nonlinear Total Variation Based Noise Removal Algorithms[J]. Nonlinear Phenomena, 1992, 60(1-4): 259-268.

[8] Aujol J F, Chambolle A. Dual Norms and Image Decomposition Models[J]. International Journal of Computer Vision, 2005, 63(1): 85-104.

[9] Chambolle A. An Algorithm for Total Variation Minimization and Applications[J]. Journal of Mathematical Imaging and Vision, 2004, 20(1-2): 89-97.

[10] Ekeland I, Temam R. Convex Analysis and Variational Problems[M]. Amsterdam, Holland: North-Holland Publishing Co., 1976.

编辑 顾姣健