基于对象存储系统中的热点数据平衡策略

周功业 雷 伟 陈进才

(华中科技大学 计算机科学与技术学院，湖北 武汉 430074)

摘要：针对基于对象存储系统(OBS)中的热点数据问题，提出了一种基于预测和数据副本迁移的均衡策略.对基于对象存储结点(OSD)的热度进行预测，元数据服务器(MDS)据此对存储结点进行均衡协调，存储结点则响应协调规则并采用贪心法选出尽量小而且访问客户多的对象进行副本迁移.此外，采用增量备份的方法解决副本迁移可能造成的数据不一致问题.实验结果表明策略的预测误差在8%以内.

关 键 词：基于对象存储系统；基于对象存储设备；热点数据；预测；副本迁移

中图分类号：TP302.7 **文献标识码**：A **文章编号**：1671-4512(2007)12-0028-04

Hotspot data balancing in OBS

Zhou Gongye Lei Wei Chen Jincai

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

Abstract: A balancing strategy based on prediction and data copy's migration was proposed in allusion to the hotspot data in OBS (object based storage). At first, it predicted the temperature of OSD (object based storage device), and then MDS (meta data service) coordinates them according to the states in prediction; OSDs executed the coordination, and migrate the copies of the object which is as small in size and more in accounts of related clients as possible, the object was chosen by the algorithm of Greedy. Moreover, the way of incremental backup was used to solve the problem in data inconsistency because of the copy migration. The experiment results show that the method successfully controls the error below 8 %.

Key words: object based storage (OBS); object based storage device (OSD); hotspot data; prediction; copy migration

基于对象存储系统(OBS)[1~3]中，多客户对同一文件的集中访问、对象分配不合理、存储结点的硬件性能存在较大差异等，都会导致系统中存储结点访问频率悬殊，从而引起严重的负载不平衡.存储结点中被频繁访问的数据就是热点数据.

OBS具有存储结点多的特点，数据以对象为单位存储在基于对象的存储设备(OSD)[4]中.文献[5~8]中的方法在OBS中效果不甚理想，一方面，OSD达到忙碌状态后才请求仲裁服务器元数据服务器(MDS)调节，这将会使OSD在从提交请求到执行仲裁的时间内处于超负荷状态，并延长执行仲裁的时间；另一方面，频繁的迁移影响客户对系统的正常读写[9]，也会造成迁移过程中元数据与实际数据之间的映射失效.基于以上问题，结合OBS中对象属性的动态特性，本文提出了一种基于节点预测和副本迁移的热点数据平衡策略.

收稿日期：2006-09-18.

作者简介：周功业(1954-)，男，教授；武汉，华中科技大学计算机科学与技术学院 (430074).

E-mail：smilinglw@126.com

基金项目：国家重点基础研究发展计划资助项目 (2004CB318201)；湖北省自然科学基金资助项目 (2005ABA257).

1　热点数据平衡策略

1.1　平衡系统结构

温度. 用来衡量 OSD 需要降低热度的紧急程度，采用 OSD 当前的热度相对于自身能够承受负载的比例来衡量. 温度越高表示 OSD 越接近热点数据引起的超负荷状态.

沸点/冰点. OSD 根据 OBS 不同的应用和 OSD 的硬件和网络状态，它自身所能承受的最大负载上限温度和可以接受其他负载的上限温度分别称为沸点和冰点. 对于因将要达到沸点和冰点而向 MDS 提交“降温”或“升温”请求的 OSD，分别称为沸点 OSD(B_{osd})和冰点 OSD(F_{osd})，其他未提交请求的 OSD 称为常温 OSD.

平衡对. MDS 对沸点 OSD 和冰点 OSD 进行热度平衡，结果以平衡对(B_{osd}，F_{osd1}，F_{osd2}，…，F_{osdz})发送到沸点 OSD 中，表示将沸点 OSD 中的热点数据副本同时迁移到 z 个冰点 OSD 中.

如图 1 所示，整个系统由 OSD 中的温度报警、对象选择、对象副本迁移 3 个模块和 MDS 中的温度平衡模块组成. 首先，OSD 预测温度，如即将达到沸点或冰点，发送“升/降温”请求到 MDS 中；然后，温度平衡模块对所有沸点和冰点 OSD 进行调节，并发送平衡对到各沸点 OSD 中；最后，OSD 根据平衡对执行对象副本迁移.

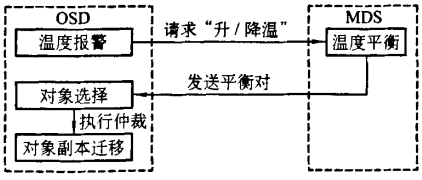

图 1　平衡系统结构图

1.2　温度报警

每隔周期 T ($T>\sum_{i=1}^{N}T_i/N$，T_i 为 MDS 访问 OSD_i 的时间)，温度报警模块进行自我检测，并预测 OSD 在下 1 个周期的温度. 本策略通过预测温度增长率的变化来预测温度. 某时刻的预测增长率主要由全局增长率(该时刻前 m_f 个周期的平均增长率)和局部增长率(该时刻前 m_l 个周期的平均增长率)两个因素决定，其中 $m_f>m_l$.

在时刻 i 以前 mT 时间内的平均增长率

$$\lambda(m,i)=\frac{1}{m}\sum_{j=i-m+1}^{i}[f(t_j)-f(t_{j-1})]/f(t_{j-1}),$$

式中：$i>0$；$t_{i+1}-t_i=T$.

全局增长率为 $\Delta(m_f,\ i)$，局部增长率为 $\gamma(m_l,\ i)$. 此时刻预测温度相对实测温度的误差为 α_i，根据应用不同而设置的容忍误差为 β. 则预测增长率

$$f(i+1)=\begin{cases}\gamma(m_l,i)+(\alpha_i/\beta)\Delta(m_f,i) & (\alpha_i>\beta);\\ \gamma(m_l,i) & (\alpha_i\leqslant\beta).\end{cases}$$

对未达到沸点或冰点的 OSD，若预测此 OSD 温度在时间 t ($0\leqslant t\leqslant 2T$)后达到沸点或冰点，则发送警报信号和时间 t 至 MDS；为统一量化 OSD 距离到达沸点或冰点的时间，对于已经达到沸点或冰点的 OSD，则发送警报信号和负数 t(t 为已达到沸点/冰点时刻与检测时刻之间的时间)至 MDS.

1.3　温度平衡

温度平衡模块对即将或已经达到沸点/冰点的 OSD 进行调节，以达到所有 OSD 温度平衡，并将平衡对发送至源 OSD 中. 调节原则：

a. 将最近要达到沸点的 OSD 与最近要达到冰点的 OSD 组成平衡对.

b. 若冰点 OSD 数目足够多，为达到更好的“降温”效果，减少沸点 OSD 的客户端访问数，则对沸点 OSD 可分配多个冰点 OSD 组成平衡对.

MDS 中有 3 个队列：存放沸点 OSD 编号的沸点队列，存放冰点 OSD 编号的冰点队列和存放未提交“升/降温”请求的常温 OSD 编号的常温队列. OBS 初始状态时，所有 OSD 编号均存放于常温队列中. 温度平衡模块接收来自各个 OSD 的警报信号，并将其按时间 t 从小到大排列存放在沸点队列或冰点队列中. 为保证时间 t 的实时性，每隔周期 T 对 3 个队列中的 t 进行更新.

沸点队列、冰点队列和常温队列中当前 OSD 个数分别为 n_b，n_f 和 n_s. 由于 3 个队列包含了所有 OSD 的状态，因此在分配平衡对过程中，始终满足 $N=n_b+n_f+n_s$. 平衡对的分配分为以下 4 种情况讨论：

a. $n_b\neq 0$，$n_f\neq 0$，$n_b\geqslant n_f$. 沸点 OSD 多于冰点 OSD，分别从沸点队列和冰点队列中取队首的结点组成平衡对(B_{osd}，F_{osd})，此循环操作至 $n_f=0$ 时，转向 **d** 或全是沸点 OSD 的情况.

b. $n_b\neq 0$，$n_f\neq 0$，$n_b\leqslant n_f$. 沸点 OSD 少于冰点 OSD，分别从沸点队列和冰点队列中取队首的结点组成平衡对(B_{osd}，F_{osd1}，F_{osd2}，…，F_{osdz})($z=[n_b/n_f]$)，循环操作直至结束.

c. $n_b=0$，$n_f\neq 0$，$n_s\neq 0$. 此时 OBS 中无沸点 OSD，只有冰点 OSD 和常温 OSD. 将常温 OSD 当作沸点 OSD，按 **a**，**b** 进行操作.

d. $n_b \neq 0, n_f = 0, n_s \neq 0$. 此时 OBS 中无冰点 OSD,只有沸点 OSD 和常温 OSD. 将常温 OSD 当成冰点 OSD,按 **a**,**b** 进行操作.

此外,还有两种发生几率较小的情况:系统中全是冰点或沸点 OSD. 在前种情况下,OBS 系统负载较轻,不需要进行热点数据迁移;在后种情况下,OBS 负载较重,每个 OSD 中均有热点数据,热点数据迁移性价比(迁移性能和迁移代价比值)较低.

1.4 对象选择

每个 OSD 中都有一个对象选择模块,由 MDS 发送平衡对的消息触发运行. 此模块选择现有访问客户数最大、容量最小的对象进行迁移.

假设此 OSD 连接客户数为 L,其中含有对象数为 n,对象连接的客户数和大小分别为 L_j 和 M_j,则有 $L=\sum_{j=1}^{n} L_j$. 按照以上标准,引入对象迁移队列,在 OSD 当前所有对象中,被添加进队列的对象,每单位大小有最大的客户连接数. 为方便描述,引进对象 j 的权重 $P_j = L_j/M_j$,用于权衡对象 j 单位大小被访问的客户数.

对被多个客户同时访问的对象描述如下:数组 $L[n]$和 $M[n]$分别含有对象 j 的连接客户数 L_j 和大小 M_j,且两数组中均按 $W_j \geqslant W_{j+1}$ 排序. 这是个简单的 0-1 背包问题[10]. 用贪心算法[10]计算出最终迁移对象的最优解,使得迁移对象尽量容量小且连接客户数多.

2 对象一致性问题

本文采用对象副本迁移来平衡 OBS 热点数据,存在对象一致性的问题. 采用存储备份中增量备份的方法,即仅向副本所在 OSD 发送操作和操作数据来减少网络流量,避免对象在各 OSD 中的不一致.

如图 2 所示,OSD_i 中的对象 i 在 OSD_{i+1} 至

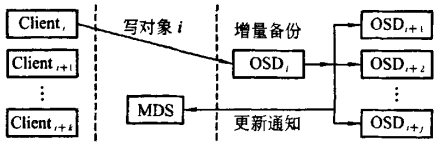

图 2 数据一致性策略

OSD_{i+j}上都有副本,在一段时间内,$Client_i$,$Client_{i+1}$,…,$Client_{i+k}$都对对象 i 进行访问,并由 MDS 中的温度平衡模块控制分布访问 OSD_{i+1},OSD_{i+2},…,OSD_{i+j}. $Client_i$ 对 OSD_i 的对象进行写操作. 首先,OSD_i 记录此操作,并通知 MDS,使得对继续访问对象 i 的客户端给予正确的数据访问导向;然后,OSD_i 将记录的操作和操作数据转发给对象 i 副本所在的所有 OSD;最后,这些对象 i 副本所在的 OSD 完成对象 i 的数据更新,并通知 MDS 更新完成,已可接收客户端的访问.

3 实验结果与分析

在本策略中,温度预测十分重要,预测的准确性直接影响着副本迁移对性能提升的程度. 以每秒输入输出比 α 为指标分别采用 DSwap[8] 和本策略对 OBS 作了测试.

OBS 由 20 个 OSD 结点和 10 个客户端组成. 温度预测测试中,在随机抽取的时间点分别提取 3 种 OSD(接近沸点的 OSD,保持常温的 OSD 和接近冰点的 OSD),并在从此时间点开始后的第 5 000 s到 9 000 s 的时间段内,每隔 500 s 提取预测温度和实际温度数据.

图 3 表明对沸点 OSD 的预测误差在时刻 7 500 s时误差较大,此时沸点 OSD 正进行副本迁移,预测温度稍稍高于实际温度. 当沸点 OSD 逐渐趋于常温时,预测温度越来越接近实际温度. 沸点 OSD 的测量中最大温度预测误差在 8 %以内.

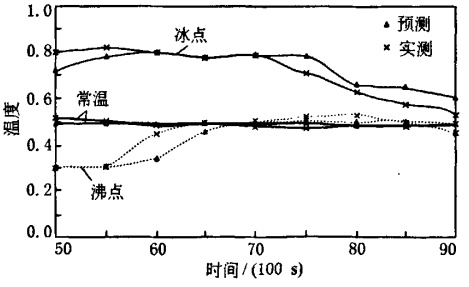

图 3 预测与实测比较

图 4 中,在时间点 6 000 s 时刻,在 20 个客户端同时发起读命令,使得 OBS 系统 IOPS 突然增大,DSwap 的系统中 IOPS 值波动较大,而本策略所在系统中 IOPS 更加趋于平衡.

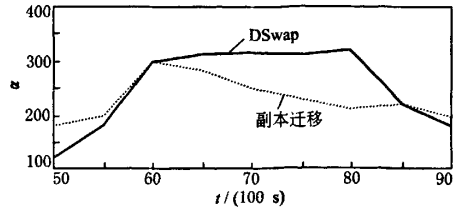

图 4 吞吐量

参　考　文　献

[1] Mesnier M, Ganger G R, Riedel E. Object-Based Storage [J]. Communications Magazine, 2003, 41(8): 84-90.

[2] Azagury A, Dreizin V, Factor M, et al. Towards an object store[C]// Proc. of 20th IEEE/11th NASA Goddard Conference on Mass Storage Systems and Technologies. New York: IEEE, 2003: 165-176.

[3] Mesnier M, Ganger G R, Riedel E. Object-based storage: pushing more functionality into storage[J]. Potentials, 2005, 24(2): 31-34.

[4] Technical Committee T10. Object-Based Storage Device Commands-2[EB/OL]. [1999-12-08]. http://www.t10.org.

[5] Hu Liang, Meng Faner. A Dynamic Load Balancing System Based on Data Migration[C]// Proc. of 8th International Conference on Computer Supported Cooperative Work in Design. Piscataway: IEEE, 2004: 493-499.

[6] 庞丽萍，许　俊，徐　婕，等. PVFS 数据访问的负载平衡[J]. 华中科技大学：自然科学版，2004，32(7)：19-20.

[7] Zhu F, Sun X W, Salzberg B, et al. Supporting Load Balancing and Efficient Reorganization During System Scaling[C]// Proc. of 19th IEEE International Parallel and Distributed Processing Symposium. Piscataway: IEEE Computer Society, 2005: 49a.

[8] Sundaram V, Wood T, Shenoy P. Efficient Data Migration in Self-managing Storage Systems[C]// Proc. of 3rd International Conference on Autonomic Computing. Piscataway: IEEE, 2006: 297-300.

[9] Dasgupta K, Ghosal S, Jain R, et al. QoSMig: Adaptive Rate-Controlled Migration of Bulk Data in Storage Systems[C]// Proc of 21st International Conference on Data Engineering. Piscataway: IEEE Computer Society, 2005: 816-827.

[10] 余祥宣，崔国华. 计算机算法基础[M]. 武汉：华中科技大学出版社，2001.

EMC 公司与我校签署战略合作协议

10 月 26 日，作为全球最大的存储厂商美国 EMC 公司与国内唯一具有自主知识产权的存储企业华工科技海恒公司签署战略合作协议，双方携手共同推动中国存储产业的发展. 仪式上，EMC 公司还与华中科技大学、武汉大学分别签署了一个全球非营利性公益教育项目——EMC 学院联盟计划合作备忘录，加强信息技术学科建设及信息科技人才的培养.

武汉海恒信息存储有限责任公司是华工科技联合华中科技大学，以武汉光电国家实验室、外存储国家专业实验室、华中科技大学信息存储系统教育部重点实验室为技术支撑于 2006 年 7 月共同投资设立的，以专业化信息存储为主业、是目前国内唯一一家拥有自主知识产权的网络存储企业. 公司核心技术包含了国家 973 重点基础研究发展计划、国家重大基础研究前期研究专项、国家自然科学基金、国家重点攻关项目等研究开发的多项成果.

此次海恒存储与 EMC 战略合作协议的签订，标志着海恒存储与 EMC 战略合作伙伴关系的确立，代表全球最大的存储厂商与中国唯一具有自主知识产权的存储企业的联手. 通过合作，双方将共同推动中国存储产业的发展，促进湖北武汉乃至国内的广大用户在信息存储和信息管理与应用等方面得到更多更好的技术支持与服务.

据介绍，EMC2006 年总收入高达 112 亿美元，已连续第 7 年获得在全球存储软件供应商排名第一. 在 2007 年《财富》杂志评选的美国财富 500 强中，EMC 位列第 224 名，同时排名全球第六大软件公司.

基于对象存储系统中的热点数据平衡策略

作者： 周功业， 雷伟， 陈进才， Zhou Gongye， Lei Wei， Chen Jincai
作者单位： 华中科技大学,计算机科学与技术学院,湖北,武汉,430074
刊名： 华中科技大学学报（自然科学版） ISTIC EI PKU
英文刊名： JOURNAL OF HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY(NATURE SCIENCE)
年，卷(期)： 2007，35(12)
被引用次数： 0次

参考文献(10条)

1. Mesnier M. Ganger G R. Riedel E Object-Based Storage 2003(08)

2. Azagury A. Dreizin V. Factor M Towards an object store 2003

3. Mesnier M. Ganger G R. Riedel E Object-based storage:pushing more functionality into storage 2005(02)

4. Technical Committee T10 Object-Based Storage Device Commands-2 1999

5. Hu Liang. Meng Faner A Dynamic Load Balancing System Based on Data Migration 2004

6. 庞丽萍. 许俊. 徐婕 PVFS数据访问的负载平衡[期刊论文]-华中科技大学学报(自然科学版) 2004(07)

7. Zhu F. Sun X W. Salzberg B Supporting Load Balancing and Efficient Reorganization During System Scaling 2005

8. Sundaram V. Wood T. Shenoy P Efficient Data Migration in Self-managing Storage Systems 2006

9. Dasgupta K. Ghosal S. Jain R QoSMig:Adaptive Rate-Controlled Migration of Bulk Data in Storage Systems 2005

10. 余祥宣. 崔国华 计算机算法基础 2001

相似文献(6条)

1. 期刊论文 宋健. 刘川意. 鞠大鹏. 汪东升. SONG Jian. LIU Chuan-yi. JU Da-peng. WANG Dong-sheng 基于对象存储系统中多维服务质量保证的设计与实现 -计算机工程与设计2008,29(3)

存储系统的服务质量(QoS)保证对于满足上层应用的需求至关重要.基于对象存储(OBS)使用支持丰富语义的对象级接口,能够更好地实现QoS保证.在研究了基于对象存储系统中QoS交互机制的基础上,提出了一个基于对象存储设备(OSD)的多维QoS框架,设计并实现了一个应用于OSD的多维QoS算法.实验显示,该框架和算法能有效满足多个客户端的不同维度QoS要求.

2. 期刊论文 谢黎明. 冯丹. 覃灵军. Xie Liming. Feng Dan. Qin Lingjun 对象存储系统中属性管理方法研究与实现 -计算机研究与发展2007,44(z1)

在对象存储系统中,对象属性用于描述数据特征以及安全策略.系统性能的改善可以通过对象属性的合理管理来实现.目前,属性存放在固定大小的数据结构中,这与属性的可扩展性相违背.针对现有对象属性放置与管理方法的不足,提出了对象属性基于扩展Hash的管理方法,以及属性的放置策略、缓冲策略.通过分析仿真测试的结果,发现在处理大量属性操作时,对象文件系统性能退化很严重.但是,用扩展Hash方法进行属性管理时,系统性能不退化,比对象文件系统更具优势.

3. 期刊论文 赵水清. 冯丹. ZHAO Shui-Qing. FENG Dan 基于对象存储设备上的服务质量研究 -计算机科学 2006,33(9)

基于对象存储(Object-based Storage,OBS)作为下一代互联网存储协议标准逐渐被人们所接受,基于对象存储设备(Object-based Storage Device,OSD)所具有的可扩展性和智能性能很好地支持应用程序的服务质量(Quality of Service,QoS)需求.本文分析了基于对象存储系统的QoS需求,讨论了如何把应用客户的qoS需求转化为QoS属性,扩展了OSD SCSI协议集标准,以支持QoS,接着采取量化分析方法对QoS进行了分析,详细分析应用客户和OSD之间QoS信息交互的工作过程,最后,紧密结合OSD的特性,给出OSD上的QoS三层模型Q-Model和一优化算法BRP.

4. 期刊论文 冯丹. 史伟. 覃灵军. 关卿. Feng Dan. Shi Wei. Qin Lingjun. Guan Qing 基于对象存储系统的对象文件系统设计 -华中科技大学学报（自然科学版）2006,34(12)

基于对象文件系统是建立在对象存储系统上的一种应用,它利用对象的特点对数据进行存储和管理.基于对象文件系统由客户端、元数据服务器和基于对象存储设备组成,通过对系统的各部分进行分析,设计并实现了基于对象文件系统.通过对不同文件大小、不同文件分块大小条件下文件系统的性能进行测试分析,找出系统的数据传输瓶颈,使用以对象属性为导向的缓存和预取技术以及聚合读写的方法对文件系统进行了优化.优化后读写吞吐率分别提高了60 Mbyte/s和40 Mbyte/s.

5. 期刊论文 李忠民. 喻占武. LI Zhong-min. YU Zhan-wu 地形数据对象存储组织方式及其分布策略研究 -测绘学报 2008,37(4)

结合基于对象存储技术,将地形数据按平面格网划分成若干个瓦片,每个瓦片映射为一个对象存储在基于对象存储系统中.根据地形瓦片之间的位置相

关性,提出一种地形数据矢量基对分布策略,将多个瓦片对象分布在多个基于对象存储设备中,充分发挥基于对象存储系统的并行性.并采用分布模板的思想来实现矢量基对分布策略,减小运算的复杂度.该策略保证查询窗口在一定范围内每台基于对象存储设备最多只被访问一次,且响应时间恒定.

6.学位论文 陕振 OSD文件系统研究与设计 2005

基于对象存储体系结构是一种新的网络存储体系结构，具有高性能、高可扩展性、良好的安全性。基于对象存储设备(OSD)是基于对象存储系统中的基本存储设备，主要承担对象属性和数据的管理、设备安全以及同外界的网络通信功能，其中对象的属性、数据管理功能要通过一个设备内置的文件系统——OSD文件系统来实现。本文研究的目的是要结合OSD本身的特征设计一个适用于管理小型对象的OSD文件系统。

本文深入地分析了OSD本身的特征，并在此基础上研究了面向小型对象的OSD文件系统。在OSD文件系统的设计中，一方面借鉴了现有通用文件系统中的优秀设计思想，另一方面考虑使文件系统充分发挥OSD本身的优势。本文重点研究了对象ID到对象数据的映射机制、存储空间的分配与回收策略以及存储空间的动态优化策略，提出了一种基于“盘区记录表”的对象空间映射方式、一种基于大块分配和基于盘区分配相结合的空间分配策略和一种主动的对象空间动态调整策略。

最后依据本文的设计方案在Linus系统下实现了一个OSD文件系统的仿真系统，再将仿真的OSD文件系统置入一个基于对象存储系统中，替换了原系统中以Ext3文件系统实现的OSD文件系统，并使用标准的测试工具postmark对采用不同OSD文件系统的新、老存储系统的整体性能进行了对比测试。测试结果表明，本文设计的 OSD 文件系统在以小型对象为主的负载下能提高系统15%到60%的整体性能，并且文件系统性能可以随系统使用时间的延长保持稳定。