基于动态副本技术的云存储负载均衡研究*

董继光，陈卫卫，吴海佳，田浪军
（解放军理工大学 指挥自动化学院，南京 210007）

摘　要：以系统总响应时间最小化为目标，以文件热度为依据，提出了一种多时间窗负载均衡策略。在计算文件热度时，不仅考虑了访问的次数和大小，还将 I/O 访问时序引入到文件热度统计中，该方法能有效控制短时间突发性数据访问导致的不必要副本创建。在多时间窗负载均衡策略中，设置了三种不同大小的时间窗口，分别实现了存储节点负载均衡、文件副本的负载均衡以及低热度文件多余副本的删除工作。实验数据表明，多时间窗负载均衡策略能显著降低 I/O 访问响应时间。

关键词：云存储；负载均衡；副本；热点
中图分类号：TP302　　**文献标志码**：A　　**文章编号**：1001-3695(2012)09-3422-03
doi:10.3969/j.issn.1001-3695.2012.09.059

Load balancing study in cloud storage based on dynamic replica technology

DONG Ji-guang, CHEN Wei-wei, WU Hai-jia, TIAN Lang-jun
(*Institute of Command Automate, University of Science & Technology of PLA, Nanjing* 210007, *China*)

Abstract: For the load balancing of cloud storage system, this pape provided a multi-time windows load balancing strategy which had three time window with different size, this strategy could achieve the load balancing of storage nodes and replicas. It designed new method to statistic the heat of files, it associated the heat of a file with time sequence, which could effectively avoid the creation of replica because of burst of file access in a brief moment. The experiment indicates that multi-time windows load balancing strategy can effectively reduce the response time of I/O access.

Key words: cloud storage; load balancing; replica; hotspots

云存储系统轻松突破了 SAN、NAS 等在性能、容量、扩展性和成本等方面的限制，实现了性能与容量的线性扩展，因此，它也成为当前的研究热点。在云存储系统中，副本技术是提高数据可靠性和访问性能的关键技术。HDFS[1]、Dynamo[2] 等均采用副本技术来保证云存储系统的数据可靠性，然而它们并没有根据文件的访问热度和存储节点负载状况来动态调整副本数量和位置，当有较多的热门数据聚集在某些存储节点时，就会引起热点问题，降低系统的整体性能。

Rabinovich 等人[3]研究表明，存储系统中的数据被访问的概率呈类 Zipf 分布，当大量热门数据对象集中到某个或某些存储服务器上时，这些存储服务器就会成为热点，导致系统负载失衡，从而严重影响系统的整体访问性能。解决热点问题的一种可行方法是采用动态副本技术，为高热度文件创建较多副本，为低热度文件创建较少副本[4]，并将副本放置在负载较轻的存储节点上，从而实现云存储系统的负载均衡。

要利用动态副本技术实现云存储系统的负载均衡，首先要统计出存储节点中文件的热度。文献[5]研究表明，用户数据访问具有突发性，而且部分突发访问请求的维持时间很短，在这种情况下，如果在出现突发访问请求时，就认为文件的热度过高并创建新副本，这样不但不能均衡系统负载，反而会耗费大量的系统开销，降低系统性能。因此，用当前短时间内的访问量来度量文件的热度会造成不必要的副本创建。然而，如果用一个较长时间周期的总访问量来度量文件的热度，则会出现在文件热度很高时，由于并未到达文件的热度统计时间而不能触发负载均衡机制。因此，需要设计一种合理的文件热度统计方法，它应该既能避免短时间突发数据访问而造成的副本创建，又能反映文件近段时间的真实热度。

在数据网格和 P2P 存储领域，已有大量的文献[6~8]对动态副本技术作了深入研究，然而这些动态副本技术主要应用在跨地域的异构系统中，并不适用于紧耦合的基于大规模集群的云存储系统。

本文提出一种基于文件访问热度的副本数量调控策略，主要包括以下两方面内容：a）设计了一种基于时间序列的文件热度统计方法，该方法能有效避免短时间突发访问而导致的不必要副本创建，而且统计的文件热度与文件热度变化趋势一致；b）设计并实现了一个具有三个时间窗口的负载均衡策略，该方法既提高了存储节点的负载均衡，又实现了各文件副本的负载均衡。

1　基于时间序列的文件热度统计

合理地统计出云存储系统中的文件热度，是基于动态副本技术进行负载均衡的关键。图 1 中的（a）~（c）描述了文件 f

收稿日期：2012-01-16；**修回日期**：2012-02-27　　**基金项目**：国家自然科学基金资助项目（60603029）；国家“863”计划资助项目（2008AA01A309）

作者简介：董继光（1986-），男，河南周口人，硕士，主要研究方向为网络存储、云存储（pladjg@163.com）；陈卫卫（1976-），女，教授，，主要研究方向为计算机算法、软件工程、云计算；吴海佳（1986-），男，博士研究生，主要研究方向为网络存储、分布式文件系统；田浪军（1985-），男，硕士研究生，主要研究方向为分布式数据库.

在三种不同访问场景下被访问的频率变化情况。在相同的时间序列 t 内,三种访问场景中对文件 f 的总访问次数基本相同,但文件访问的时序却有很大差异。在图 1(a)中,用户对文件的访问主要集中在时间序列 t 的前一部分;在(b)中,用户对文件 f 的访问相对均匀地分布在整个时间序列 t 内;在(c)中,用户对文件 f 的访问主要集中在 t 的最近一个时间段内。根据数据访问的局部性原理可得,当前被频繁访问的数据在未来一段时间内可能还会有较高的访问频率,而当前很少被访问的数据在未来一段时间内可能仍然很少被访问。因此,这三种场景下的文件 f 应具有不同的热度,即图 1(a)中的热度最高,(b)中的热度次之,(c)中的热度最低。

因此,文件的热度取决于两方面:a)单位时间内该文件的访问量(包括访问频率和访问大小);b)单位时间内用户对该文件的访问时序。基于以上分析,本文提出了一种基于时间序列的文件热度统计方法。其计算方法初步设计为

$$H_0(f)=0 \tag{1}$$

$$H_i(f)=\lambda\times S_i+(1-\lambda)\times H_{i-1}(f),\lambda=\frac{\triangle t_i}{\triangle t_i+\triangle t_{i-1}} \tag{2}$$

其中:f 表示云存储系统中的某个文件;$H_i(f)$ 表示第 i 次访问到达后,文件 f 的更新热度值;S_i 表示第 i 次访问的字节数;λ 为热度调节因子,它与文件的访问时序有关;Δt_i 表示从文件 f 热度最后一次清零到第 i 次访问请求到达时的时间间隔。

当文件 f 的第 i 个请求到达时,系统会记录此次请求的到达时间,并通过式(2)计算出第 i 个请求后的文件热度 $H_i(f)$。在计算 $H_i(f)$ 时,通过 λ 将访问时序引入文件热度的度量当中,并综合考虑访问的字节数以及前一次的热度值 $H_{i-1}(f)$ 对文件热度的影响。从式(2)可以得出,随着时间的推移,距当前时间越久的访问对文件的当前热度影响越小,较近的访问对文件的热度影响较大,近期的访问量越大,文件的热度越高。

因为在用户数据访问的高峰时期,访问量可能会以爆发的形式出现,若此时仍在每次发生数据访问时计算文件热度,将会耗费大量的计算资源,反而降低系统性能,因此,对文件热度的统计不能在每次发生数据访问时进行,而应定义一个采样周期,统计采样周期内的文件平均访问量。设采样周期为 T,T 时间内对文件 f 访问请求数量为 n,第 t 次的访问请求大小为 S_t,可得 T 时间内发往文件 f 的总访问量 S',$S'=\sum_{j=1}^{n}S_j$。因此,基于时序的文件热度统计公式最终设计为

$$H_i(f)=\lambda\times S_i'+(1-\lambda)\times H_{i-1}(f)$$

其中:$\lambda=\frac{\Delta t_i}{\Delta t_i+\Delta t_{i-1}}$,$S_i'$表示第 i 个采样周期内文件总访问量。

2　多时间窗负载均衡策略

根据云存储系统的实际情况,为实现云存储系统各存储节点以及文件副本的负载均衡,本文提出一种多时间窗负载均衡策略。该策略设置三个时间窗口,分别为 T_1、T_2、T_3,其中($T_1<T_2<T_3$),T_1 时间窗负载均衡是为了解决由于文件分布不均衡而引起的单个存储节点负载不均衡问题;T_2 时间窗负载均衡是为不同热度的文件创建不同数量的副本,实现各个文件副本的负载均衡;T_3 时间窗负载均衡是为了删除低热度文件的多余副本。文件的热度采用第 1 章中基于时间序列的热度统计方法进行计算。

2.1　云存储架构介绍

根据云存储的特性,笔者自行设计了一个云存储系统 FFS(formicary file system)[9],通过自主开发的分布式文件系统将集群中大量廉价通用的存储设备协同起来,共同对上层应用提供海量存储服务。FFS 云存储系统由主控服务器、存储服务器和客户端三大模块组成。主控服务器模块对云存储客户端提供目录服务和元数据服务,并对存储服务器集群进行监控;存储服务器模块负责文件数据的具体存放;客户端模块负责对云存储客户机提供虚拟磁盘服务,将云存储客户机对虚拟磁盘的操作请求提交给主控服务器,并从存储服务器读取/写入数据。

2.2　存储节点负载状况统计

在进行云存储系统的负载均衡之前,必须收集各存储节点的负载信息。具体步骤如下:

a)存储节点定期向主控节点发送负载状况信息。设存储节点 i 的负载为 L_i。

b)主控节点根据各存储节点的负载,计算出整个系统的平均负载为 $L_{\text{average}}=\frac{1}{N}\sum_{i=1}^{N}L_i$,其中 N 为存储节点的数量。

c)主控服务器把存储节点分成三类,分别存放在 S_H、S_M、S_L 三个集合中。这里设置一个常量 $\delta(0<\delta<1)$,对于节点 N_i:

$$N_i\in\begin{cases}S_H & L_i\geqslant L_{\text{average}}(1+\delta)\\ S_M & L_{\text{average}}(1-\delta)<L_i<L_{\text{average}}(1+\delta)\\ S_L & L_i\leqslant L_{\text{average}}(1-\delta)\end{cases}$$

其中:S_H、S_M、S_L 分别表示负载重、负载适中、负载轻的存储节点集合。

2.3　T_1 时间窗负载均衡

设 $L_{\max}$ 表示存储节点负载上限。在 T_1 时间窗负载均衡阶段,每经过时间间隔 T_1,通过 2.2 节中的负载状况统计,主控节点可以发现系统中负载大于 $L_{\max}$ 的重载节点,之后在主控节点的控制下,重载节点为存储在自身的高热度文件创建新副本,创建副本的目的节点为轻载节点。负载均衡流程描述如图 2 所示。

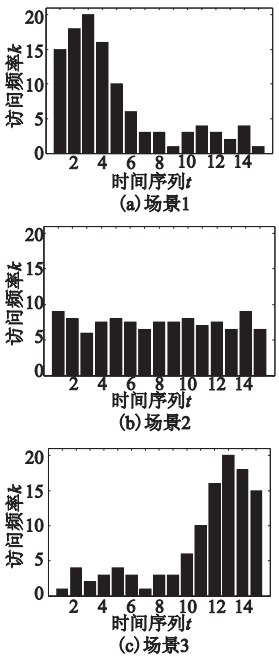

图1　访问场景3

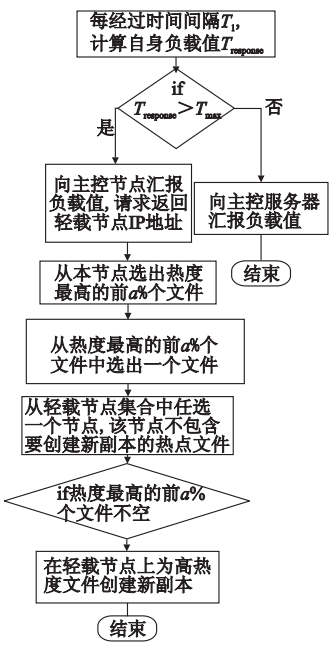

图2　T_1时间窗负载均衡流程

这里 T_1 的值一般设置得比较小,如几十分钟到几个小时,这样可以更加迅速地被发现因突发性的访问请求而导致的热

点问题,对突发性访问请求的敏感程度可通过调节 T_1 的值设置。对于要创建新副本的文件数量可通过 $a\%$ 的值进行调节,一般来讲,设置为 3% ~5% 比较合适,如果 $a\%$ 的值设置过小,将不能有效消除热点;如果 $a\%$ 的值设置过大,可能会因为创建新副本而消耗过多的系统资源。

2.4 T_2 时间窗负载均衡

T_2 时间窗负载均衡是根据文件热度为文件创建副本的阶段。每经过时间间隔 T_2,存储节点向主控服务器汇报自身所存储的所有副本的热度,主控服务器计算出每个文件的热度 H_{file} 以及云存储系统中所有文件的平均热度值 H_{average},即

$$H_{\text{file}} = \sum_{i=1}^{N_{\text{current}}} H_{\text{file}}(i)$$

$$H_{\text{average}} = \frac{1}{A}\sum_{A} H_{\text{file}}$$

其中:N_{current} 表示文件的当前副本数量;$H_{\text{file}}(i)$ 表示文件的第 i 个副本热度;A 为系统中文件总数。

为了保证云存储系统的数据可靠性,每个文件都必须有一个最小副本数量,设这个值为 C,则需要为每个文件创建新副本的数量为

$$N_{\text{add}} = \left\lceil \frac{C \times H_{\text{file}}}{H_{\text{average}}} \right\rceil - N_{\text{current}}$$

由此可知,文件的副本数量与文件热度成正比关系,当 $H_{\text{file}} \leqslant H_{\text{average}}$ 时,副本数量为最小值 C;当 $H_{\text{file}} > H_{\text{average}}$ 时,副本数量为 $\left\lceil \frac{C \times H_{\text{file}}}{H_{\text{average}}} \right\rceil$。

系统为单个文件创建副本流程如图 3 所示。

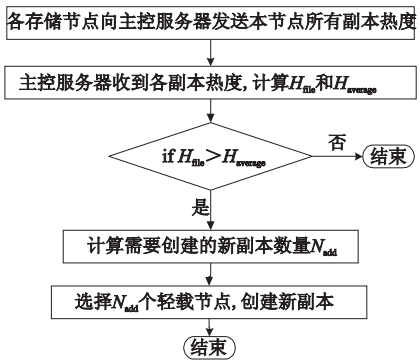

图3 T_2时间窗负载均衡为单个文件创建新副本流程

一般来讲,T_2 时间窗设置较大,如 1 天或几天。T_2 设置较大有两个原因:a)防止突发性数据访问造成的不必要副本创建;b)若 T_2 时间窗设置太小,会因为频繁负载均衡而造成较大的系统开销。T_2 时间窗负载均衡阶段根据文件的热度为文件创建新副本,使文件的副本数量与文件的热度成比例,使系统中所有副本的 I/O 负载相当,既可以提高文件的并发访问性能,又可以减小云存储系统产生热点的隐患,进一步提升云存储系统负载均衡能力。

2.5 T_3 时间窗负载均衡

在云存储系统中存储了很多曾经热度很高而现在却很少被访问的文件,根据多时间窗负载均衡策略,当文件热度很高时,在 T_1 和 T_2 阶段会为该文件创建较多副本,而现在文件的热度很低,过多的副本占用了大量宝贵的存储空间却很少被访问,并且会带来副本一致性维护的开销。针对这个问题,多时间窗负载均衡策略设置了 T_3 时间窗负载均衡。

在 T_3 时间窗负载均衡阶段,单文件热度 H_{file} 与系统中所有文件平均热度 H_{average} 的计算方法与 T_2 时间窗负载均衡阶段相同。当发现文件的副本数量大于最小副本因子(保证数据可靠性的最小副本数量),同时文件热度小于 H_{average} 时,直接将该文件的副本数量减小至最小副本因子。

T_3 时间窗一般设置很大,如一周或几周,这样既可以减小删除副本操作引起的系统开销,又可以减小短时间内文件再次热度升高而引起反复创建新副本的开销。

3 性能测试与实验结果分析

为验证基于时序的文件热度统计方法以及多时间窗负载均衡的有效性,本文在 FFS 云存储平台上实现了本文中的文件热度统计和负载均衡算法。FFS 云存储系统的硬件配置环境如下:主控服务器(Dell R710),CPU 2.93 GHz ×6,内存 64 GB,硬盘 SATA500 G/7200 转,数量为 1 台;存储服务器(普通 PC),CPU 2.33 GHz,内存 1 GB,硬盘 SATA500 GB/7200 转,数量为 20 台。测试中的所有 I/O 访问请求均为读访问;对于同一个文件的多个副本采用轮循方式为用户提供数据访问;T_1 = 1 小时,T_2 = 1 天,T_3 = 1 周,每个文件的最小副本数量为 3。

3.1 基于时序的文件热度计算公式测试与分析

在测试中,提取了某邮件服务器上 15 min 的 trace 数据,每 1 min 计算一次热度。考虑到单个文件 trace 数据不能代表大部分文件被访问的情况,本文从 trace 数据中任选出 10 个文件,对 10 个文件的平均访问频率和大小进行了分析。表 1 显示了所抽取的 10 个文件在 15 个时间段内的平均访问频率和大小。

表 1 样本文件在 15 个时间段的平均访问频率及大小

时间段	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
大小/KB	14	13	15	21	27	17	15	13	12	7	6	12	11	23	26
频率	30	45	23	34	36	50	39	40	82	69	75	46	52	32	29

图 4 显示了基于时间序列的热度统计值变化趋势和当前实际热度值变化趋势,这里的当前实际热度是通过计算 1 min 内的总热度值得到的。

从图 4 可以看出,当有突发的数据访问时,如时间段 5 和 9,基于时序的热度统计方法能够很好地削减短时间内突发数据访问导致的热度值飙升,这样可以无须为短时间内的突发访问创建不必要的文件副本。此外,基于时序的热度统计方法所计算出的热度值和真实的热度值具有相同的变化趋势,当热度值持续上升时,该统计方法能够真实地反映出热度变化趋势,为负载均衡决策提供了可靠的依据。

3.2 多时间窗负载均衡测试与分析

通过不断提高对云存储系统的访问压力,测试了云存储系统在有多时间窗负载均衡机制和无负载均衡机制的响应时间变化。如图 5 所示,随着读操作压力的增大,在无负载均衡情况下,系统的响应时间先是缓慢地增加,但是随着访问压力进一步增大,系统的响应时间迅速上升,当访问压力达到 8 千次以上时,系统的响应时间已经达到不能容忍的地步,系统几乎无法正常工作。在有多时间窗负载均衡策略情况下,随着访问压力的增大,系统响应时间略有上升,这是由于系统中存储节点的压力普遍有所上升,随着访问压力达到 5 千次时,系统的存储节点负载开始不均衡,部分节点开始成为热点,此时系统启动负载均衡机制,系统的响应时间维持在一个比较稳定的水平,随着访问压力的进一步增大,系统的平均响应时间也迅速上升,但响应时间远小于无负载均衡策略的情况。

(下转第 3436 页)

(上接第 3424 页)

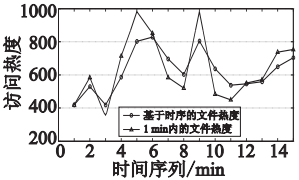

图4 基于时序的热度统计值变化趋势

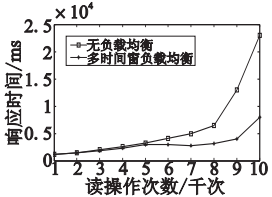

图5 系统响应时间对比

4 结束语

本文设计了一种基于时间序列的文件热度统计方法,该方法将文件的热度与 I/O 访问时序结合起来,能有效避免突发数据访问引起的不必要副本创建。本文提出的多时间窗负载均衡策略通过 T_1 时间窗负载均衡能有效解决突发数据访问以及数据分布不合理引起的热点问题,通过 T_2 时间窗负载均衡实现云存储系统中各副本的负载均衡,通过 T_3 时间窗负载均衡删除低热度文件的多余副本。

参考文献:

[1] SHVACHKO K, HAIRONG K, RADIA S, *et al*. The hadoop distributed file system[C]// Proc of the 26th IEEE Symposium on Mass Storage Systems and Technologies. 2010: 1-10.

[2] De CANDIA G, HASTORUN D, JAMPANI M, *et al*. Dynamo: amazon's highly available key-value store[C]//Proc of the 21st ACM SIGOPS Symposium on Operating System Principles. New York: ACM Press, 2007: 205-220.

[3] RABINOVICH M, RABINOVICH I, RAJARAMAN R, *et al*. A dynamic object replication and migration protocol for an Internet hosting service[C]//Proc of the 19th IEEE International Conference on Distributed Computing Systems. 1999: 101-113.

[4] 谭支鹏. 对象存储系统副本管理研究[D]. 武汉:华中科技大学, 2008.

[5] ANDERSON E. Capture, conversion, and analysis of an intense NFS workload[C]//Proc of the 7th Conference on File and Storage Technologies. Berkeley, CA: USENIX, 2009:139-152.

[6] ZHONG Hai, ZHANG Ze-hua, ZHANG Xue-jie. A dynamic relica management strategy based on data grid[C]//Proc of the 9th International Confe-rence on Grid and Cloud Computing. 2010: 18-23.

[7] QAISAR R, LI Jian-zhong, YANG Dong-hua. A load balancing replica placement strategy in data grid[C]//Proc of the 3th International Conference on Digital Information Management. 2008: 751-756.

[8] SASHI K, LECTURER S. A new replica creation and placement algorithm for data grid environment[C]//Proc of International Conference on Data Storage and Data Engineering. 2010: 265-269.

[9] 吴海佳,陈卫卫. FFS:一种基于网络的 PB 级云存储系统[J]. 通信学报,2011,32(9A): 18-23.