

一种面向移动终端的云存储方法研究

吴永发, 曾 碧

(广东工业大学 计算机学院, 广东 广州 510006)

摘 要: 在分析 HDFS 和 Dynamo 架构的基础上, 引入数据库技术以及动态一致性哈希算法分类处理不同对象组的方法. 本方法可根据各个存储节点存储代价, 进行综合调度, 对外提供最优存储节点. 通过算法描述和仿真结果分析表明, 此方法可以使系统达到负载均衡, 并且从整体上提高了系统的效率, 在数据查找定位上也可以满足要求.

关键词: 云存储; 副本策略; 动态节点; 一致性哈希算法; 负载均衡

中图分类号: TP311.133.1; TP302

文献标识码: A

文章编号: 1000-7180(2013)06-0048-04

Resesrch on the Method of Cloud Storage Oriented to the Mobile Terminal

WU Yong-fa, ZENG Bi

(School of Computers, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: Based on the analysis of HDFS and Dynamo architecture, a strategy that the database technology and the dynamic consistent hashing algorithm are adopted to deal with different target groups, is presented in this paper. The strategy can schedule integratively according to the storage cost of each storage node, provide the optimal storage node to external. Through the theoretical analysis and the simulation results, this strategy can make the system achieve load balance. Moreover, it can improve overall efficiency of the system and also meet the requirements the data lookup positioning.

Key words: cloud storage; replication policy; dynamic node; consistent hashing algorithm; load balance

1 引言

目前, 物联网技术的迅猛发展, 给我们的生活带来了许多便捷. 移动智能手持终端^[1]在许多领域等到了应用, 针对不同的应用环境, 传统的云存储系统并不能很好地适应, 存储的密度不大, 综合效率比较低等.

当前云存储架构中比较成熟的方案有 HDFS^[2]和 Dynamo^[3].

经过以上对 HDFS 和 Dynamo 的分析, 本文引入了数据库技术以及基于一致性哈希布局算法, 分组处理不同大小的数据对象. 针对小对象文件采用关系数据库进行存储索引, 大对象文件采用类 Dy-

namo 系统模型处理, 整体上提高系统效率, 同时保证系统的负载均衡. 通过实验, 验证了本方案的可行性.

2 云存储系统结构模型

云存储系统架构^[4]共划分为 4 个层次, 如图 1 所示. 由上到下依次为: 用户访问层、应用接口层、基础管理层以及数据存储层. 从用户的角度来看, 该架构可以只划分为两层, 用户的客户端程序和云状的后台程序. 在本文中重点关注数据备份以及系统运行状态监测.

图 1 中, 基础管理层涉及的功能是整个系统的重点和难点, 需要实现系统的用户权限控制、数据的

收稿日期: 2012-07-20; 修回日期: 2012-09-27

基金项目: 国家自然科学基金(61173046); 广东省部产学研结合项目(2010B090400183); 广州市黄埔区科技计划项目(1130)

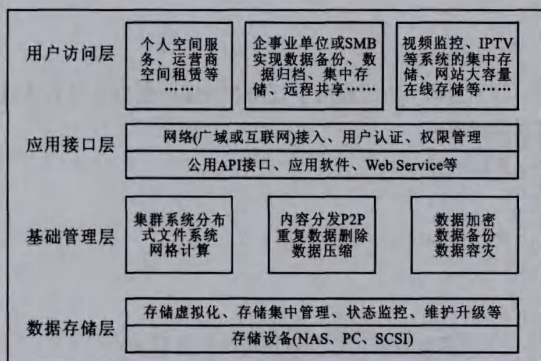


图1 云存储系统架构

备份和存储策略。针对不同大小的数据对象采用不同的存储策略和备份方案。而数据存储层突出的是系统运行状态的监控。整个系统对外服务时,为了提高系统的性能,存储节点之间要达到负载均衡的效果,就要知道系统中的各个节点的运行状况,综合调度各个节点。

为了应对可以存储不同大小的文件,本系统的整体架构如图2所示。

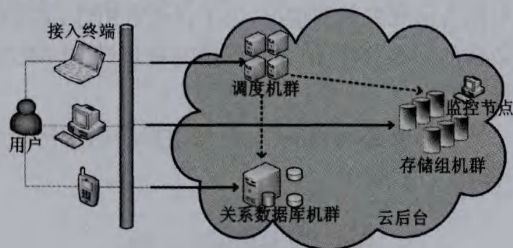


图2 云存储系统整体架构

用户通过接入终端连接云后台^[5-6],接入终端先与调度机群中负载最低的机器通信,然后根据文件大小把控制权交给关系数据库机群或存储组机群。而这两个机群根据自身的负载情况,选择负载低的机器对外服务。小文件交由关系数据库机群处理,而大文件就交给存储组机群处理。系统中采用的都是廉价的PC机,成本底下,但软硬件的稳定可靠性也低,为了提高系统的可靠性、增强系统的容灾性,云存储系统中引入了副本机制。数据的多副本机制还可以提高系统的效率,移动终端请求服务时,可以选择最近的副本区域。而传统的云存储系统都是先选择主副本,只有主副本错误了才请求备用副本。这种方式没有考虑地域性,影响了系统的速度。假设云存储系统的可靠性为 R ,存储组机群中的PC机的可靠性为 R_j ,关系数据库机群中PC机的可靠性为 R_k ,副本数为 r ,则系统的可靠性模型为:

$$R = [1 - (1 - R_j)^r][1 - (1 - R_k)^r]$$

其中 $R_j = R_k = 0.9, r = 3$,则 $R = 0.998$ 。

以上系统中关系数据库机群和存储组机群分别采用不同的存储策略和备份方案,将对此展开描述。

3 算法描述

HDFS默认的副本数为3,其放置的位置为同一个机架以及另一个机架。数据的放置过程中会考虑存储节点的容量大小,节点容量越小被选中的概率就越小,所以当存储节点的容量相差比较大时,系统中容量大的节点就会出现负载过重的情况。

本文的副本机制可以提高系统的可用性和整体性能,并且其中的副本策略可以分为静态副本策略和动态副本策略。静态副本策略指的是副本个数、放置的位置从创建开始到数据失效期间都固定不变的,而动态副本策略指的是系统运行中会因为系统性能要求,负载情况等因素影响,调整副本的个数、副本的放置位置等。对于存储的可扩展性,系统还需要考虑存储节点的扩容,包括节点的加入以及节点的退出等,下面分别针对小文件和大文件的副本策略算法进行描述。

3.1 小文件副本策略

这里定义的小文件指的是大小不超过10 MB的文件。采用的关系数据库是MySQL数据库,经过调度机群判断文件大小,数据库机群获取控制权后,直接把文件等相关属性数据保存到数据库的表中。

在数据库机群中有一个主服务器,用于判断数据库机群中哪个机器负载比较低,然后需要存储的数据直接与这个机器交互,这样可以做到动态选择最优节点。为了保证数据的可靠性,需要保存一份数据的副本到备服务器上。同时主服务器还需要保存文件存放位置的机器IP,这样查询文件的时候需要先从主服务器得到保存文件的机器IP,然后再跟数据库服务器交互。其结构如图3所示。

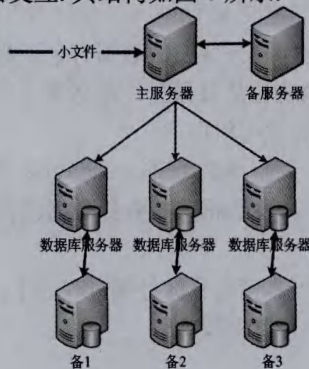


图3 小文件处理结构图

图中,主服务器选取负载最低的数据库服务器用于存储文件内容,备服务器用于对主服务器的备份.备 1 到备 3 分别是对数据库服务器的备份.主服务器有如表 1 所示的结构,用于保存文件名和对应实际保存位置的 IP 地址.当数据库服务器的数据量到达一定量时,就会对这些数据打包,保存到存储组机群中.

表 1 主服务器文件表

字段	含义	类型	长度
ID	序号	int	
filename	文件名	varchar	255
IP	IP 地址	varchar	15

数据库服务器是实际存储文件内容的地方,它有如表 2 所示的结构.

表 2 数据库服务器文件表

字段	含义	类型	长度
ID	序号	int	
filename	文件名	varchar	255
filesize	文件大小	int	
createtime	创建时间	datetime	
context	文件内容	mediumtext	

表中,context 保存了文件的二进制文本内容.

主服务器和数据库服务器对各自的数据库表建立了索引,可以加快检索速度.

对于存储的可扩展性,涉及副本的调整过程:需要分一部分数据到新加入的数据库服务器,同时更新主服务器上的文件表.系统实现主要有以下函数:

(1)主服务器调度函数 MainSchedule(). 该函数的功能是当有小文件请求时,通过计算当前环境下数据库服务器的负载情况,选择负载最低的服务器,直接返回该服务器的 IP 地址,然后外界的请求就可以直接与之交互了.

(2)数据库服务器中的数据打包函数 PackData(). 该函数的功能是对数据库服务器上的数据进行打包保存到存储组机群中.

(3)副本调整函数 ChangeCopy(). 该函数的功能是,添加和减少数据库服务器时,数据的调整迁移处理.

小文件副本策略具体处理过程如下:

输入:小文件请求(读或写)

处理:

HandleSmallFile()

```
{
    new Thread()
    /* 启动一个线程用于监控数据库中数据量是否达到
    打包的标准 */
    While(1)
    {
        PackData(); /* 打包数据 */
    }
}
/* 该函数返回当前负载最低的机器 IP */
char * strIP = MainSchedule();
.....
/* 当出现增加或减少服务器时,就会调用调整副本的
函数 */
ChangeCopy();
}
```

输出:读写成功标志

3.2 大文件副本策略

在图 2 系统整体架构中,存储组机群采用普通 PC 机为存储介质.由于普通 PC 机的可靠性不高,有可能会出现失效的情况,为了保证可用性,需要保存数据的副本.这里的大文件指的是超过 10 MB 的文件.

从网络互连的角度看,存储组机群采用网状结构,如图 4 所示.

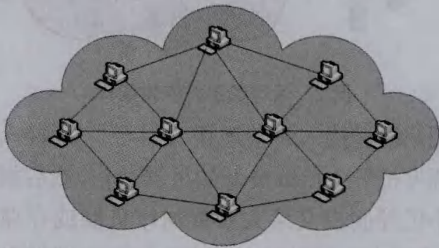


图 4 存储组机群网络拓扑图

图中,每台 PC 机都保存了相邻节点的存储信息,通过查询节点中的存储信息可以快速定位到对应的存储节点.

整个存储组机群的存储空间统一编址,我们把这个存储空间看作是环状结构的,但由于不同 PC 机之间的实际性能是存在差异的,为了屏蔽这些差异,在实际的存储空间上虚拟了一层连续的存储空间.系统采用 MD5 函数作为地址映射的哈希函数,将实际物理地址通过哈希处理,得到一个 32 比特的信息串,即可映射到虚拟连续地址中,这样就屏蔽了底层 PC 机之间的差异,同时该信息串也用于数据的校验,如图 5 所示.

每当有数据进入系统,以该数据的内容作为输

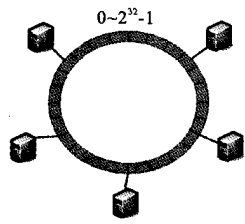


图 5 虚拟存储空间图

入,经过同一个哈希函数处理,也映射到了虚拟存储空间环中,然后顺时针查找到第一个已经映射的 PC 机,最终数据就被保存到该机器上,同时保存两份数的副本到相邻的 PC 机上. 根据哈希函数性质,当数据量越大,最终数据空间分布就会越均匀.

PC 机在保存数据的时候,都要更新对应 PC 机以及相邻的 PC 机上的路由信息. 路由表结构如表 3 所示.

表 3 路由信息表

字段	含义	类型	长度
ID	序号	int	
filename	文件名	varchar	255
filesize	文件大小	int	
IP	文件所在 PC 机的 IP	varchar	15

表中的 IP 字段包含副本所在 PC 的 IP,该路由信息只记录与之相邻的节点信息,并且当相邻节点出现故障后,与之相邻的节点可以捕获到,并做相应的处理,如修改路由信息、重新备份该 PC 机上的存储信息.

当有 PC 机新加入存储机群网络时,根据一致性哈希算法的原理,哈希环上已经映射的空间不会发生变动,只是新增加的 PC 机的存储空间被映射到新的空间中去了. 这样可以有效避免空间震荡的问题. 同时也要更新与之相邻的 PC 机上的路由信息.

新增 PC 机的处理过程:

```
AddPC()
{
    /* 通过广播消息,获取与之相邻的 PC 的 IP 信息 */
    char * arrIP = Broadcast();
    /* 根据一致性哈希算法映射虚拟地址 */
    HashMap();
    /* 更新自身的路由信息 */
    UpdateNode();
    /* 更新相邻节点的路由信息 */
    UpdateNode(arrIP);
}
```

```
.....
}
```

减少 PC 机的处理过程与增加 PC 机的处理过程是类似的,不同点在于,假如顺时针看三台 PC 机的顺序是 A、B、C,则 B 退出时,A 和 C 都能检测到,这时 B 中保存的 A 的副本转移到 C,B 中保存的 C 中的副本转移到 A. 而 B 中本身为主备份的数据则转移到 C 的下一个映射的 PC 机上.

4 实验与结果分析

本文采用两类实验验证本方案的可行性. 一是采用小规模的真实环境,实际部署系统,统计各个数据. 二是采用 CloudSim 仿真^[7],大规模模拟一定数量的存储节点,仿真真实环境,验证系统的可行性.

5 结束语

本文提出的面向移动终端的云存储系统副本策略,分别对大小不同粒度的文件对象采用不同的存储策略,综合考虑系统的效率,提高系统的负载均衡效果,同时也提高了系统的稳定性和扩展性. 在接下来的关注重点是在保证系统高稳定,高可靠的前提下,减少副本数量,进一步提高系统的存储效率.

参考文献:

[1] 张迪,朱立谷,侯振宇,等. 基于 WEB 的移动端云存储技术研究[J]. 计算机工程与应用,2010,46(36):66—69.

[2] Dhruba Borthakur. The hadoop distributed file system: architecture and design[EB/OL]. [2012-04-30]. http://hadoop.apache.org/common/docs/r0.18.0/hdfs_design.pdf.

[3] Giuseppe Decandia, Deniz Hastorun, Madan Jampani, et al. Dynamo: amazon's highly available key-value store [C] //Proceedings of the 21st ACM Symposium on Operating Systems Principles. New York: ACM, 2007:205—220.

[4] 徐婧,杨寿保,王淑玲,等. CDRS: 云存储中一种代价驱动的自适应副本策略[J]. 中国科学院研究生院学报,2011,28(6):759—767.

[5] Zeng Wenying, Zhao Yuelong, Ou Kairui, et al. Research on cloud storage architecture and key technologies [C] // ICIS'09: Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human. New York: ACM, 2009:1044—1048.

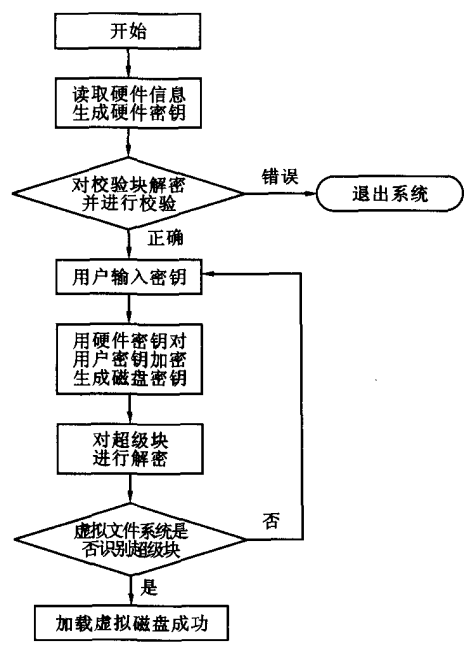


图 5 虚拟磁盘认证的流程

于对移动存储设备中机密文件进行保护,因此 SVFS 具有更高的实用性.

参考文献:

[1] 李清俊,甘萌. 基于虚拟磁盘的文件加密方法[J]. 计算机工程与设计,2006,27(15):2835-2838.
[2] 朱征宇,周发贵,于春雷. 一种基于计算机指纹的 U 盘文档监控与保护方法[J]. 重庆理工大学学报:自然科学,2011,25(3):62-68.
[3] 包怀忠. EXT2 文件系统的分析与研究[J]. 计算机工程与设计,2005,26(4):1022-1024.
[4] 陈莉君,张琼声,张宏伟,等译. 深入理解 Linux 内核[M]. 3 版. 北京:中国电力出版社,2007.
[5] 魏丕会,卿斯汉,刘海峰. 基于安全操作系统的透明加密文件系统的设计[J]. 计算机科学,2003,30(7):132-135.
[6] 王俊华,译. 操作系统设计与实现[M]. 北京:清华大学出版社,2008.

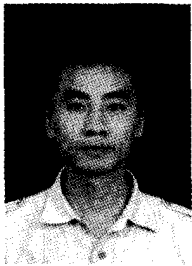
作者简介:

崔 奇 (1987-),硕士. 研究方向为计算机应用技术.

(上接第 51 页)

[6] Li W, Yang Y, Yuan D. A novel cost-effective dynamic data replication strategy for reliability in cloud datacentres[C] // International Conference on Cloud and Green Computing. Sydney,2011:496-502.
[7] Calheiros R N, Ranjan R. CloudSim;a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms[J]. Software: Practice and Experience, 2011, 41(1):23-50.

作者简介:



吴永发 (1987-),硕士研究生. 研究方向为分布式网络存储、云存储、嵌入式软件技术.

曾 碧 女,(1963-),教授. 研究方向为移动计算与物联网技术、嵌入式系统.