文章编号：2096-1472(2018)-03-31-03
DOI:10.19644/j.cnki.issn2096-1472.2018.03.008

基于伯努利大数定律的云存储数据方法研究

陈维华，何彩虹
(河北软件职业技术学院，河北 保定 071000)

摘 要：随着科技水平的提高，对云存储服务的可靠性、安全性和稳定性都有了更高的要求。面对云存储服务，如何优化资源配置，进而提高用户的体验质量，本文提出了基于大数定律的云存储方法，具体方法是用伯努利大数定律按照存储频率，根据历史信息确定用户对资源的需求，然后进行再分配，从而减少了数据访问延迟。

关键词：大数定律；云存储；数据分配
中图分类号：TP399 **文献标识码**：A

Research on the Method for Cloud Storage Based on Bernoulli LLN

CHEN Weihua,HE Caihong
(*Hebei Software Institute,Baoding* 071000,*China*)

Abstract:with the improvement of science and technology,there are higher requirements for the reliability,security and stability of cloud storage service.In terms of cloud storage service,how to optimize the allocation of resources and to improve the quality of user experience. This paper proposes a method of cloud storage based on the Law of Large Numbers,in which the concrete way is to determine and redistribute users' demand for resources based on the frequency of storage and historical information by means of Bernoulli LLN,thereby reducing the data access delay.

Keywords: Law of Large Numbers (LLN);cloud storage;data distribution

1 引言(Introduction)

云计算的发展在现如今的信息技术中应用越来越广泛，云存储服务[1]也凭借着它的高扩展性、高可靠性、成本低、方便数据管理的诸多优势受到人们的青睐，与云服务相关的产品也越来越受到用户的喜爱。云存储系统[2-4]是一个以数据存储和管理为核心提供云计算能力的高性能计算系统。它可是实现对于海量数据的采集、管理和保护等功能。用户可以通过云存储实现不同区域，不同时间的资源共享和互动，并且通过应用权限的控制、传输加密、信息加密和数据隔离技术这些技术保证数据的安全性。

在享受大量数据在云存储服务中的便利的同时，其对于数据的访问速度也有了更高的要求。由于在云端人们不断的把数据上传保存上去，在海量的存储空间中对数据信息进行存储。随着存储的数据量越来越大，对于数据的访问时间也就越来越长。云存储系统的DBAS结构为B/S三层体系结构，分别是用户层、应用层、数据存储层。在数据存储层存储大量的数据信息和数据逻辑，所有与数据有关的安全、完整性控制、数据的一致性、并发操作都是在这层完成的。B/S结构的特点是分布性强，维护方便、开发简单并且共享性强、总体拥有成本低等特点。但是数据安全性问题，以及数据传输速度慢等的缺点也显而易见。由于系统期望用户可以从云存储中及时的获得想要的数据，因此减少用户在对数据进行访问的时候造成访问延迟成为数据资源分配有待解决的问题。

目前针对云存储中的数据资源分配问题，研究者们提出了各种不同的云数据存储方案。祁志阳[5]从经济学原理的角度分析了云计算的经济学模型，以用户对资源评价的相似度为约束条件进行资源调度管理，结合经济学的超边际分配方法对资源进行分配的。由于在建模过程中数据是静态的，而在物理环境下数据是动态形式，会造成数据不准备等缺点。Siva ThejaMaguluri[6]根据一个随机的过程，如果作业到达时请求访问资源，采用加入最短的队列算法和MaxWeight调度选择

基金项目：2017年河北省教育厅青年基金项目《基于移动智能终端的医疗系统设计与实现》(项目编号：QN2017340，主持人：陈维华).

算法，建立了一个负载均衡，以便在资源利用的过程中提高吞吐量。Fabien Hermenier[7]提出了一种通过减少虚拟机迁移和分配到主机的云计算时间的方法，以提升资源利用率。基于此，本文的主要工作要工作有：

(1)通过历史访问痕迹对用户进行分类。由于不同类型用户对于宽带、延迟等的要求不同，所以需要提供的服务也不一样。不同类型用户有不同的需求，可以根据历史信息计算各类型用户需求量。

(2)应用伯努利大数定律确定各类用户对数据的需求，根据需求分配存储资源。

2 数据存储(Data storage)

2.1 云存储数据的资源配置

在云存储数据资源配置的过程中，涉及数据的采集、数据维护、数据的存储方式等多个要素，它们彼此之间的相互协作构成了云存储数据的资源配置过程。然而，不同用户对云存储的资源进行上传和下载的过程中，对资源的需求也不同，对需要云服务数据的配置也会有不同的要求，这样就加大了资源配置的复杂性。从用户的角度来说，对需求的服务造成延时，会降低用户的使用效率，同时，对云存储服务器也会造成负载不均衡的状态。

因此，本文通过访问痕迹对于每一类的数据需求进行分类，在流量约定设置的优先级里，一些特定的网络数据流也需要定义服务质量。比如多媒体数据流要求有保障的通过量；IP电话则需要严格的抖动和延迟限制；在远程外科的手术中则要求有可靠保证的可用性。除了这些特定的数据服务外，对于一般的普通用户而言也需要要有针对性的数据服务。可以通过访问痕迹对每一组数据的内容进行分析，对数据内容可以按兴趣爱好、年龄、性别、工作性质等大致分成几类。对于这些访问的数据用伯努利大数定律计算出每一类的需求量，从而确定服务器存储数据内容的优先级。

根据用QoS服务分析每个传送的报文内容，将这些报文归类到以CoS(分类的标准)值来表示的各个数据流中，对它进行标注。

由于云计算环境具有虚拟化的特点，把硬件物理资源虚拟化为虚拟服务资源，这样可以对虚拟服务资源利用软件程序进行重新配置，并通过配置子程序实现不同用户的需求。

2.2 根据伯努利大数定律计算量建立存储结构

在伯努利大数定律计算的数据量建立的存储结构中，按照计算的数值，云计算服务系统被划分为无数个资源池。这些资源池不只是存储资源，还要对已有的资源池进行管理。每一个资源池里，信息管理系统对资源中分类的数据信息进行统一管理。

如图1所示：客户端和服务端两部分构成了伯努利大数定律计算数值建立的存储结构。

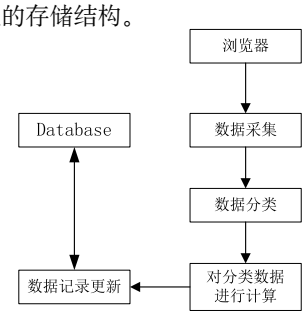

图1 存储结构

Fig.1 Storage structure

下面给出伯努利大数定律的计算公式：

设定 μ 是n次独立试验中事件A发生的次数，且事件A在每次试验中发生的概率为P，则对任意正数 $\varepsilon>0$，则成立。

其中公式中的n是访问的数据量，μ 指的是在n次访问的数据量中某一类型在固定的时间内访问数据的次数，且A是每次实验中发生的概率为P。

相应的根据伯努利大数定律确定的数据量的值确定的存储结构的步骤：

Step1：对参数进行初始化，设定任务的数量和属性；

Step2：根据浏览器的访问记录计算当前的访问数据值；

Step3：根据兴趣爱好，性别、年龄将数据资源划分为几类；

Step4：通过伯努利大数定律计算每一种类型的所占比；

Step5：依据每一种类型的所占总访问量资源的比重重新对云存储数据的方式进行设置。

3 实验与分析(Experiment and analysis)

3.1 实验环境

针对本文所提出的根据伯努利大数定律的云存储数据方法，本文在QoS服务的基础上，通过浏览器的访问数据痕迹进行了仿真实验，并计算了几种不同类型的用户对数据访问量的数值。在实验过程中，主机的内存为8GB，硬盘为520GB，操作系统为windows 10。本文提出的基于伯努利大数定律的云存储模型与文献中[5]的信誉度约束下的超边际约束的云存储资源分配模型与文献[6]最短队列算法进行对比，通过在执行任务的平均时间t和负载均衡度 σ 作为指标来衡量QoS服务质量的性能。因此有如下定义：

定义1：t表示执行任务的平均时间，在资源上执行任务n所用的整体的时间T，$\max\{t_{ci}\}$表示在任务集T中完成最后一个任务的时间，$\min\{t_{cj}\}$表示在任务集T的第一个任务开始的时间，则有如下公式：

$$t = \frac{\max\{t_{Ci}\} - \min\{t_{Cj}\}}{n}$$

定义2：σ表示负载均衡度的大小，即云服务资源处理任务所需时间的方差L与带处理任务数n的比值。其中vl_i表示虚拟机负载量，m表示虚拟资源的数量，$avlc$表示虚拟服务资源的平均负载值，则：

$$L = \sqrt{\frac{\sum_{i=1}^{m}(vl_i - avlc)^2}{m}}$$

$$\sigma = \frac{L}{n}$$

σ的值越小，说明负载均衡的性能越好。反之，性能越差。

3.2 性能分析

本文选定了50的虚拟服务系统，通过比较伯努利大数定律的云存储数据的方法与信誉度约束超边际分析云存储数据方法和采用对短队列算法进行比较。选取50的虚拟服务器，将任务数量从50到1000个独立任务构成的任务集，对执行任务的平均时间t和负载均衡度σ的进行分析得出以下量表数据信息，如图2和图3所示。

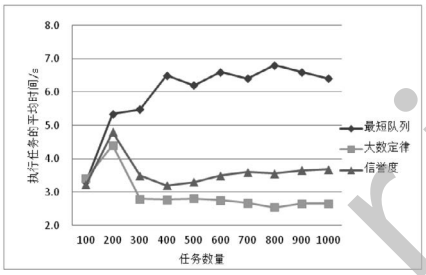

图2 三种方法执行任务的平均时间

Fig.2 The average time of three methods perform tasks

图2可以看出：用伯努利大数定律算法比信誉度约束算法和最短队列算法执行任务的平均时间越来越少，曲线也更平稳。

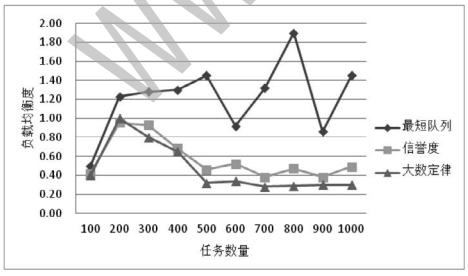

图3 三种方法的负载均衡比较

Fig.3 Load balancing comparison of three methods

在图3中，最短队列算法的负载均衡度较大，变化幅度也大，不稳定；在信誉度算法和大数定律的算法中两者的负载均衡度差距不是很大，但是大数定律的曲线更平稳。因此，用伯努利大数定律算法能更好的提高资源利用率，是云服务系统的负载均衡达到更好的效果。

仿真结果表明，对于云服务数据存储的资源配置问题，依据伯努利大数定律的数值结果进行分配，使云计算资源节点的使用率达到最佳，减少了延迟，并提升了任务完成时间。

4 结论(Conclusion)

本文分析云服务数据存储方式在资源配置过程存在负载不均衡问题进行了研究，提出了一种基于大数定律的云存储数据方法，给出了思路和过程。通过对历史访问数据的分类和用伯努利大数定律的计算，计算了不同类型的数据访问量的数值，并加以分析。实验表明本方法对数据访问量的计算，可以以此为依据对数据资源进行合理分配。提高了资源利用率，减少访问延迟。

参考文献(References)

[1] 冬瓜头(张东)大话存储II[M].北京:清华大学出版社,2011:22-24.

[2] MATHER T,KUMARASWAMY S,LATIF S.Cloud security and privacy:an enterprise perspective on risks and compliance[M].Cloud Security and Privacy:An Enterprise Perspective on Risks.Sebastopol,CA:OReilly Media,2009:35-72.

[3] 傅颖勋,罗圣美,舒继武.安全云存储系统与关键技术综述[J].计算机研究与发展,2013,50(1):136-145.

[4] 李晖,孙文海,李凤华,等.公共云存储服务数据安全及隐私保护技术综述[J].计算机研究与发展,2014,51(7):1397-1409.

[5] 祁志阳,马满福.信誉度约束下超边际分析的云存储[D].西北师范大学,2015.

[6] HuberN,BrosigF,Kounev S.Model-based self-adaptive resource allocation in virtualized environment[J].in:SEAMS,ACM,2011:90-99.

[7] DuPontC,GiulianiG,HermenierF,et al.Anenergyaware framework for virtual machinePlacement in cloud federated data centers[C].Future Energy System:WhereEnergy,Computing and Communication Meet (e-Energy),2012 Third International Conference on.IEEE,2012:1-10.

作者简介：

陈维华(1978-)，女，硕士，副教授.研究领域：物联网技术及应用.

何彩虹(1980-)，女，硕士，工程师.研究领域：物联网技术及应用.