

**本科毕业设计(论文)文献综述**

**题目：web 云存储硬盘系统的设计与实现**

|  |  |
| --- | --- |
| **学生姓名：** | **李政** |
| **学 号：** | **1606031127** |
| **班 级：** | **16级软件工程1班** |
| **专 业：** | **软件工程** |
| **院 （系）：** | **计算机工程学院** |
| **指导教师：** | **刘明纲** |

二零二零年一月三十日

**文 献 综 述**

**1．前言**

生活中每时每刻都离不开文件，早在互联网诞生以前，纸质文件一直是人们工作学习的媒介。现如今，随着网络和计算机技术高速发展，产生了各式各样的文件，如何存储和管理这些文件就成了重要问题。随着技术的发展，以往传统的方式可能满足不了用户的需求，比如随着用户文件量增多，本地存储介质容量有限，迫切需要扩容；以及用户经常出现文件误删误改操作，就诞生了文件备份的需求。纵观存储领域的技术发展，不管是硬件还是软件，都在改进和优化技术以达到上述目的。

学校实验室作为一个为学生提供学习研究的场所，日常实验中不乏涉及大量的数据文件和资料，这些数据都是非常宝贵的，对于实验者，应该第一时间保存下来。另外，各种文件数据都应该进行合理备份，避免因各种宕机事故造成损失。此时，首先想到的是市面上的各种公共云存储服务，把文件保存或备份到第三方服务商提供的云存储服务中。然而，众多高校的实验室都没有外界互联网，另外，就算采用第三方服务有几点缺陷，一是容量受限于服务提供商，二是对于安全性和隐私性不能得到保证。那么，能否在实验室或学校内网搭建一个云存储系统呢? 答案是肯定的。这样对于其容量可以很方便地由系统管理员设置，由于是学校内网和由学校内部人员管理，其安全性也大大提升。如果强制要求所有实验数据都保存到该系统中，可以有效避免数据遭到泄露，可以极大程度保护数据。除此之外，可以将系统功能扩展一下，不仅仅只是用于实验室，例如在课堂，老师用于分享教案和学习资料，学生提交作业和实验报告等。

基于上述论述，通过查阅相关文献，了解到很久之前，网盘就是一个研究热点。云计算、大数据、人工智能是近年的热门话题，而云存储是云计算的基础，而其他方面也依赖云计算，所以纵观整个技术领域，云存储特别是分布式存储是基石。目前，国内公共云存储服务商一家独大，在许多方面备受争议，这也是由于云存储某些缺点导致的。比如，成本高昂，由于数据量庞大，导致服务器数量多，从而维护成本高。另外，数据安全和隐私也是重点关注话题，这些问题在未来甚至当下亟待解决。只有在保护好用户的隐私的前提下，才能让用户感到放心；只有让数据被安全地存储和运用，才能利用好大数据技术为用户服务和提供便利。

**2．主 体**

云存储通常由于数据量大，文件数量多，所以采用分布式文件系统。最主要的原因之一就是单机存不下也处理不了，例如云计算涉及海量数据，如果不采用分布式架构，整个系统性能受限于单台机器。分布式存储首先要解决如何决定把文件放到哪台机器上，有两种方式解决这个问题，第一，让客户端决定存储到哪里，最简单的方法就是取hash,比如客户端取得服务器列表后，对其长度取模即可选出目标服务器。但是这种方法不利于服务器数量的增减，而且更为重要的是，其分派逻辑体现在客户端，安全性有失降低。另一种情况，设置一个控制服务器，由它负责统一调度，客户端请求存储或访问时，都由控制服务器管理文件的存储信息。不管采用哪种方式，必须确保对其系统的成员信息、映射规则、监控信息能够掌握。为了实现高可用，应该增设备用控制服务，避免主服务挂掉导致整个系统瘫痪。这样的话，就又引入了一个问题，即master/slave问题，即分布式的协调一致性。现有Paxos,Raft算法，google的分布式锁服务chubby就是基于paxos实现的，也可以直接使用现有的分布式协调系统实现，如Zookeeper等。对于大文件，可以采用目前很流行的分块技术，几乎所有现有的分布式系统都采用了分块技术，比如对于一个4GB的文件,可以按照64MB的块大小分块。接下来，为了实现数据的可靠性和可用性，避免出现服务器宕机或者是某个分块由于网络传输等原因导致了不能访问，在通常情况下，就必须引入冗余技术，所谓冗余就是对于一个分块，同时在系统中存储多份。至于存储多少份，需要结合系统的成本和数据可靠性要求，通常一般系统选择三份即可。对于确实成本有限的系统，可以引入纠错码技术，主要应用在网络传输中避免包的丢失， 存储系统利用它来提高 存储 可靠性。相比多副本复制而言， 纠删码能够以更小的数据冗余度获得更高数据可靠性， 但编码方式较复杂，需要大量计算，目前大多数分布式系统都引入了该技术。在采用冗余技术后，有两种复制方式，一种是直接复制，另一种是链式赋值。直接复制由系统给出存储逻辑后，由客户端方复制，缺点很明显，增加了两倍的带宽占用。链式复制是要求客户端只上传一份到系统，然后由系统内部完成复制的任务。由于系统中同时存储数据的多个副本，那么如何确保数据的一致性，即假如文件修改后，其所有副本也应该更新。一致性策略分为强一致性或最终一致性。前者要求必须完成所有副本更新操作完成后才能响应用户请求，缺点是请求响应时间长，用户体验不好，优点是可以确保用户不会读到脏数据。由于强一致性的种种缺点，最终一致性模型降低了强一致性的要求，例如可以设计为只要一半以上的副本完成更新即可，剩下部分利用系统的“空闲”时间完成。虽然采用冗余技术实现了系统的高可用和数据的可靠性，但是毕竟导致了系统成本上升，为了节省空间，可以采用压缩技术。另外对于文件重复的问题，多个用户可能上传相同的文件，系统应该能够识别并且只保证存储一份。其实现原理是现在本地计算文件的hash指纹，然后再上传到后台进行比对，其实这一点比上述的压缩技术更能节省空间。其实上述论点只是分布式系统中的冰山一角，但是最基础的技术要求。

**3．总 结**

本次研究虽是为了实现局域网内的云存储，但是可以参考比较成熟的公共云存储的实现方案。但是目前国内公共云存储领域一家独大，其限速模式饱受诟病，这是受其于分布式系统有高IO的特性，而不只是服务商的利益欲望。相比传统系统，除了外部访问，其内部也有大量的数据交互。如果不能很好地控制和管理分配IO资源，就会导致某个用户独占整个系统的IOPS。所以说，未来公共云肯定是会向个人云存储方面变化，比如现在已有的家用NAS。一般的NAS支持多协议，比如NFS,CIFS,FTP甚至HTTP。使用NAS，可以根据自己的需求在硬件和软件方面都可以对系统做定制化处理，除了常用的存储和备份之外，还能实现资源的共享和协同工作，免费的离线下载，甚至远程监控网页服务器，打印服务器等。对于企业，可能会更多的选择混合云，私有云和数据中心的某种组合。它通常将由企业拥有和管理的资源与由第三方管理的公共云存储服务相结合。企业将这两种方法结合起来，以平衡保护关键任务资产的需求与公共云存储提供的弹性、可扩展性和成本优势。其实个人使用云存储是为了备份，云存储的目的是将数据和文件的副本保存为备份。这样，如果原始数据或文件被破坏、泄露，则可以在云平台上使用备份进行恢复。如果数据丢失或被破坏，则必须访问云存储才能访问重复文件。而企业除了备份，更多的是为了计算，云计算使计算能力和云存储器可用作服务。与云存储一样，用户可以根据需要访问尽可能多的容量。存储系统必须具备更高的性能，以满足海量巨增数据的实时分析，高效处理需求；必须增强可扩展性，以满足日益增长的数据存储的需求；必须具备更高的安全级别，以保证数据的绝对安全; 必须具备更高的安全级别，以保证数据的绝对安全; 必须具备更高的灵活性，以降低客户易运维易管理的需求；必须具备更高的性价比，以降低客户的使用成本。一直以来，存储性能都远远低于计算性能，这是受其硬件所限，尽管各种厂商和实验室研究设计新型方案试图解决这样的差异，但是效果都不足够理想。在过去的十几年间，CPU的性能提升了8-10倍，DRAM速度提升7-9倍，网络速度提升100倍，Bus速度增长20倍，磁盘速度却仅仅增长了1.2倍——磁盘阵列已经成为IT系统中的短板。随着大数据时代的来临，企业迫切需要对数据进行实时分析，高速处理以满足企业的根本需求。另外，技术是一把双刃剑，即提供了便利，也带来了一些必须值得深思的问题。比如数据的安全性和隐私性，大数据虽然能够给企业带来利益，但是如果使用不当会给数据拥有者带来损失。大数据安全是涉及技术、法律、监管、社会治理等领域的综合性问题，其影响范围涵盖国家安全、产业安全和个人合法权益。同时，大数据在数量规模、处理方式、应用理念等方面的革新，不仅会导致大数据平台自身安全需求发生变化，还将带动数据安全防护理念随之改变，同时引发对高水平隐私保护技术的需求和期待。

**4．参考文献**

[1]田胜利,杜根远,胡子义. 基于用户级权限管理的网络硬盘的设计与实现[J]. 计算机与网络. 2007(06)

[2]胡坤华,粟栗. 网络密文硬盘中的高效认证方案[J]. 微计算机信息. 2006(36)

[3] 刘杰. 浅谈数字化校园中网络硬盘的建设[J]. 福建电脑. 2010(09)

[4] 王洋. 浅谈网络硬盘的应用及其优势以及缺点[J]. 电脑学习. 2009(04)

指导教师（签字）：

年 月 日