基于内部云存储的图书馆数据资源存储研究

乔 杨

(郑州轻工业学院 图书馆,河南 郑州 450002)

摘 要: 可靠存储环境是图书馆数字资源为用户提供高效服务的保障。在阐述图书馆数据资源存储技术基础上,分析了云存储内涵及在图书馆中的应用,探讨基于内部云存储的图书馆数字资源数据存储方案,详细分析了其层次架构、访问控制、扩展和部署方法。

关键词: 云存储; 数字图书馆; 云计算; 数据存储
中图分类号: G250.7 文献标识码: A 文章编号: 1004 - 1680(2012) 01 - 0011 - 04

可靠存储环境是图书馆开展数字资源服务的基础保障[1]。随着计算机与网络技术快速发展,数字图书馆和虚拟参考咨询等工作逐步开展,图书馆数据资源信息量呈爆炸式增加,呈现数据量大、数据类型复杂、读者需求复杂等特点。由此,易扩展、高性能、可靠性高的海量数据存储方式成为目前研究的热点,而云存储技术由于其特点从提出开始便受到广泛的关注,在一定程度上改变了我们对于传统存储模式的理解。

在介绍图书馆数据资源存储技术的基础上,分析了云存储内涵及在图书馆的应用,探讨基于内部云存储的图书馆数字资源数据存储方案,详细分析了其组织架构、访问控制、扩展和部署方法,为图书馆数字资源高效、安全存储构建提供借鉴。

1 图书馆数据资源分类

图书馆存储的数据资源类型主要包括电子数据资源和自建数据库数据资源[2]。电子数据资源包括引进的网络数据库、数据镜像站点数据、电子图书、电子期刊和多媒体数据资源等类型。随着数字化图书馆和虚拟参考咨询等工作的开展,电子数据资源每年都在急剧增长。自建数据库数据资源由高校图书馆自建的特色数据库、随书光盘等组成。以郑州轻工业学院图书馆为例,近年来新增了《轻院艺术系优秀学生作品数据库》、《民俗文化研究专题数字图书数据库》、《烟草专题数字图书数据库》等 11 个特色专题数据库,建立了专家库和知识数据库,初步建立了虚拟咨询参考模式。此类数据是图书馆所特有的馆藏或专题资源,对存储安全性有较高的需求。

2 图书馆数据资源存储技术

目前图书馆数据资源存储技术应用较为广泛的主要有 DAS、NAS、SAN 和 ISCSI[3,4]:

直接附加存储(DAS) 是传统图书馆数据资源存储方式,优点为技术架构成熟、费用低廉。不足之处为存储设备通过电缆直接连到服务器,对服务器硬件的依赖性大,使得扩展性较差,只适合小型数字图书馆。

网络附加存储(NAS) 将分布、独立的数据整合为集中管理的数据中心,对不同主机和应用服务器进行访问,具有响应快和带宽高的特点。缺点为单个设备容量受限,受网络环境影响大,适用在并发访问用户不多的中小型图书馆。

存储域网络(SAN) 提供灵活的网络存储访问和链接,服务器可以访问存储区域的任一存储设备,达到跨平台共享目的,易扩展,投资成本较高,适用在大型图书馆数据存储。

Internet 小型计算机系统接口(ISCSI) 技术克服了 DAS 的局限性,可以跨平台共享存储资源,并可以热插拔式扩充存储容量,具有硬件成本低、操作简单等特点。但 IP 网络的效率和延迟是存储数据传输的巨大障碍[5]。

DAS、NAS、SAN 和 ISCSI 等网络存储技术在一定程度上解决了数据存储问题,但也有各自的优缺点和适用范围。

3 内部云存储技术

3.1 内部云存储结构

云存储(cloud storage) 是通过集群应用、网格技术或分布式文件系统等功能,将网络中大量各种不

同类型的存储设备通过应用软件集合起来协同工作,共同对外提供数据存储和业务访问功能的一个系统[6]。内部云存储部署在用户的防火墙内,拥有并控制本部门的私有存储空间,用户可以在咨询机构帮助下建设,也可以由自己来管理和维护,存储拓扑如图 1 所示。使用云存储,并不是使用某一个存储设备,而是使用整个云存储系统带来的一种数据访问服务。

图 1 内部云存储结构

3.2 内部云存储优点

从实施规模上来看,内部云存储在实施规模上可以非常小,可以只有几个节点即可,使得更易于管理,实现云存储的成本低、易管理、易扩展性等特点,使之成为企业和单位部署云存储的可行性方案,具体上内部云存储具有以下特点[7]:对不同服务器的支持,对存储容量扩展支持、易管理性、高数据安全性、系统架构搭建简单、良好的数据读写性能。

3.3 云存储在图书馆中的应用

在国际上,图书馆界和相关机构已开始采用云存储技术来减少成本、提高效率,如: Fedora Commons、DuraSpace、MetaArchive、LOCKSS、Library of Congress 等机构都相继给出云存储研究方案。

Fedorazon[8] 是英国联合信息系统委员会(JISC) 资助的一个项目,通过在亚马逊的云计算平台上部署 Fedora Commons 存储软件,来降低英国高等教育与继续教育的存储成本,同时提高数据的存储性能。DuraSpace[9] 为校验和利用云技术来进行数字资源的长期保存,开展了 DuraCloud 计划,将基本的存储需求交由最好的云存储服务提供者负责,在此基础上增加一些额外的功能来完善存储解决方案,保证数据的长期可用性和易用性。

俄亥俄州大学图书馆联盟通过使用亚马逊网络服务来存储大量的数字资源,如肯特州百年收藏; 哥伦比亚地区公共图书馆使用亚马逊 EC2 服务部署了他们的网站; 东部肯塔基大学图书馆通过使用谷歌、Docs 谷歌日历进行办公和指导会议。

4 基于内部云存储的图书馆数据资源存储方案

图书馆数据资源存储方案构建不仅要考虑调查信息系统的高效、安全等现实需求,而且要考虑存储系统的扩展性、可靠性、易用性等需求,因此,在存储方案设计时,需要考虑以下几个方面:

(1) 存储容量。图书馆数据资源包含文本、图像、声音、视频等格式,所以大容量的存储是必不可少的。

(2) 系统性能。图书馆开展的同步或异步虚拟参考咨询等服务都需要访问数据库(如: 知识库、专家库、资源数据库等) ,进行数据查询、数据统计、数据挖掘等数据库操作,这些服务是通过在线的形式来体现的,用户有个容忍度,高存储系统性能对系统推广具有重要作用。

(3) 可靠性。图书馆数据资源经过若干年的积累而成,来之不易,有些数据属于本单位私有,所以必须把存储安全放在首位。

(4) 扩展性。随着数字化图书馆建设的不断完善,越来越多的特色数据库、数据库镜像站、随书光盘、在线视频等出现,使得资源数据呈几何级数激增。存储方案应通过不同的方法来扩展系统容量,满足将来剧增的用户和访问量需求。

4.1 基于内部云存储的结构模型

图书馆数据资源内部云存储组件部署在单位防火墙后面,保障了数据传输的安全性。在开始构建时,可以从现有的、性能较低的服务器开始,在硬盘插槽中安装低成本硬盘,然后通过逐步添加更多的机器来进行扩展。为取得简单性,文件管理选项限制在创建、读取、更新、删除和移动、拷贝上。方案采用纯软件的解决方案,包括各种操作系统、中间件、数据库及应用程序接口等,允许用户使用不同服务器组成海量、可扩展的存储池,以提供多用户使用,满足易管理、易扩展、高存储容量和安全性能高等需求,其内部云存储结构层次结构图如图 2 所示:

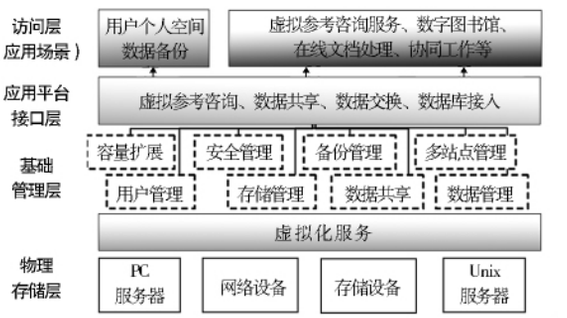

图 2 图书馆数据资源内部云存储结构层次图

(1) 访问层。授权用户通过标准的应用程序接口(API) 来访问内部云存储系统的应用场景。就图书馆工作来说主要包括个人数据资源备份、数字图书馆、虚拟参考咨询服务、协同工作等应用场景,其数据包括文本、音频、视频、图片等类型。

(2) 应用平台接口层。根据访问层不同的应用场景,云存储部署单位开发和调用不同应用接口满足图书馆实际需求,包括虚拟参考咨询服务、数字图书馆数据库接入、数据资源交换、数据资源整合、文档处理、协同工作等。

(3) 基础管理层。云存储核心部分,通过不同的管理、集群技术、分布式文件系统和网格计算等技术,实现云存储中多个存储设备之间的协同工作,使多个存储设备可以对外提供同一种服务。

(4) 物理存储层。在物理存储系统和服务器之间增加一个虚拟层,提供虚拟化服务来管理和控制所有存储设备,提供存储服务。该方式服务器不直接与存储硬件直接通信,存储硬件的增减、调换、分拆、合并对服务器层完全透明,摆脱了物理存储容量的局限性,存储设备可以是 FC 光纤通道存储设备、NAS 和 SCSI 等 IP 存储设备,也可以是 SCSI 或 SAS 等 DAS 存储设备。

4.2 存储访问方法

内部云存储的访问方法途径是传统存储和云存储差异之一,内部云存储的搭建可以由控制节点和数据节点组成,控制节点控制数据在数据节点的存储分配,数据节点个数由用户按需求来分配,通过一定的链接方式进行访问,将不同种类的存储设备通过软件进行连接,协同工作,对外提供数据存储和业务访问服务。

在具体数据存储访问中,可以通过基于具象状态传输(Representational State Transfer,REST) Web APIs 方法来实现,通过基于文件的协议(如: NFS、CIFS 或 FTP 等) 、基于块的协议(如 ISCSI 等) 或基于 Web 的分布式与版本控制(WebDAV) 协议来集成应用程序,从而可以简单有效地来提供数据访问,充分利用云存储,图 3 是图书馆数据资源内部云存储访问示意图。

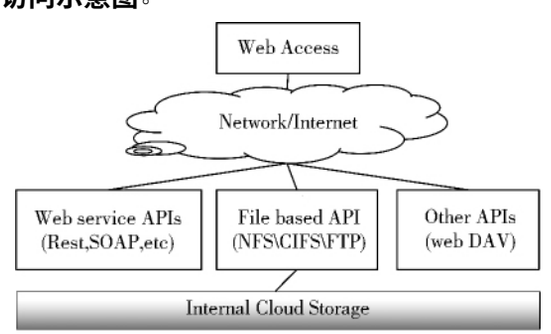

图 3 调查数据内部云存储访问途径

该解决方案可以对现有的基础架构进行整合,通过虚拟化和自动化技术,构建本单位的云存储,实现硬件和软件资源的统一管理、分配、部署、监控和备份。

4.3 存储可靠性

许多云存储解决方案提供了存储控制策略,使用户对其成本有更大控制权。如,Amazon 通过 RSS (Reduced Redundancy Storage) ,为用户提供最小化的存储成本方式。Nirvanix 提供基于策略的复制来对如何以及在何处存储数据提供更细粒度的控制。

内部云存储系统一般情况下可以通过数据的复制、数据节点备份、数据的校验等机制来保证数据的可靠性。数据存储最小单位可以为固定大小的数据分片(block) ,文件可以采用信息分布式算法(Information Dispersal Algorithm,IDA) 被分成多个数据分片,如图 4 所示,IDA 算法支持使用 Reed - Solomon 代码对数据进行切片处理,以便在数据丢失的情况下实现数据重建。允许配置数据分片的数量,可以对一个可接纳故障的数据对象分割成 4 个切片,如图所示,对 8 个可接纳故障的数据对象分割成 20 个切片。有了为数据分片的能力以及纠错码能力,就可以将切片分发到不同的存储位置进行存储。在可能发生物理故障的物理存储节点和网络中断的情况下来提高存储系统的可用性。

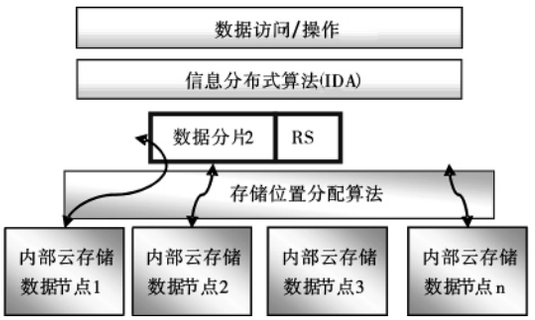

图 4 IDA 最大化数据可靠性方法

按数据节点位置进行数据存储控制,可以降低系统负荷,对块得读取只需向控制节点发送数据分片所存储的位置信息即可,减少了信息交互量。读入的数据分片放入缓存中,这样可以对该数据分片进行重复操作,处理结果统一更新。

4.4 数据存储扩展性

云存储数据的单位采用 File Storage 方式,File Storage 是基于文件级别的存储,把一个文件放在一个硬盘上,即使文件太大需要拆分时,也放在同一个硬盘上。优点是对一个多文件、多用户使用的系统,总带宽可以随着存储节点的增加而扩展,它的架构可以无限制的扩容,且成本低廉。

内部云存储在扩展存储容量时,只需要安装云

存储管理软件在相关存储节点上,然后通过光纤与网络交换机相连。控制节点把侦测到的新增存储节点和存储容量合并到原来的存储池,把一些数据自动迁移到新的存储节点,以便增加存储读写能力。当用户端存储负载变大时,内部云存储管理者可以通过增加数据冗余的方法,把数据复制到不同存储节点,使更多存储节点也可以提供该数据的访问,增加读取性能。存储空间对用户来说是透明的,整个扩容过程在线操作,不影响用户的存储操作,给用户带来方便。

5 讨论

作为新的技术,云存储概念从提出便成为数据存储领域研究的热点,改变了传统存储模式,而内部云存储模式由于其高安全性、可控性等特性为图书馆内部建立云存储提供了可行性方案。构建基于内部云存储研究,为高校图书馆数据资源长期保存提供解决方案。

参考文献:

[1] 高建秀,吴振新,孙 硕. 云存储在数字资源长期保存中的应用探讨[J]. 现代图书情报技术,2010(6):1 - 2.

[2] 陶 蕾. "云"下的图书馆网络存储探讨[J]. 图书馆学研究,2010(7):66 - 67.

[3] David Sacks. Demystifying DAS,SAN,NAS,NAS Gateways,Fibre Channel,and iSCSI. IBM Storage Networking. March 2001. http://mail. ing - steen. se/share/text/school/unix/sysadmin/more/demystifying. pdf.

[4] 贺雪晴,吴景海. 基于云计算的数字图书馆资源存储研究[J]. 情报探索,2010(12):92 - 93.

[5] 邵必林,吴宝江,边根庆. 基于 ISCSI 技术的 SAN 应用研究[J]. 西安建筑科技大学学报(自然科学版),2009,41(1):112 - 114.

[6] Zheng,Weimin,et al. Design a cloud storage platform for pervasive computing environments. Cluster Computing: the Journal of Networks,Software Tools and Applications 13. 2(2010):141 - 142.

[7] Cong Wang,et al. Toward publicly auditable secure cloud data storage services. IEEE Network,2010,24(4):19 - 24.

[8] Fedorazon [EB/OL]. [2010 - 04 - 20]. http://www.ukoln. ac. uk/repositories/digirep/index/Fedorazon.

[9] DuraCloud [EB/OL]. [2010 - 04 - 20]. http://www.duraspace. org/duracloud. php.

Design for Library Data Resources Storage System Based on Internal Cloud Storage

QIAO Yang

(*Library*, *Zhengzhou Institute of Light Industry*, *Zhengzhou* 450002, *China*)

Abstract: Reliable storage environment is the basis for library digital resources providing efficient services for users. Based on the discussion of library digital resources storage different technologies, cloud storage connotation and applications in library are analyzed in this paper. According to the characteristics of library digital resources, a storage architecture based on internal cloud storage for library digital resources is designed. And the hierarchical structure, the methods of access control, expansion and deployment are detailed.

Key words: internal cloud storage; digital library; cloud computing; data storage

作者简介:乔 杨(1981 -),女,硕士,郑州轻工业学院图书馆馆员,主要从事图书馆信息咨询方面的工作。

收稿日期:2011 - 09 - 05
(责任编辑 段麦英)