第 26 卷 第 3 期
2011 年 6 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol.26 No. 3
Jun. 2011

基于云计算的高校数字化资源整合系统方案

熊聪聪，肖 桐
（天津科技大学计算机科学与信息工程学院，天津 300222）

摘 要：分析了高校数字化资源的现状及所面临的问题，介绍了云计算的结构模型，借助开源云计算系统 Hadoop，给出了基于云计算的数字化资源整合系统方案. 搭建了小型云计算集群，给出了配置方法，并对系统的输入输出性能进行了实验，验证了该方案在数字化资源整合研究中的可行性.

关键词：云计算；数字化资源；整合
中图分类号：TP393 **文献标志码**：A **文章编号**：1672-6510(2011)03-0059-04

Framework of College Digital Resources Integration Based on Cloud Computing

XIONG Cong-cong，XIAO Tong
（College of Computer Science and Information Engineering，Tianjin University of Science & Technology，Tianjin 300222，China）

Abstract：The status of digital resources and the problems it faced was analysed，the basic concepts and structure models of cloud computing were introduced. With the help of open source cloud computing system Hadoop，the framework of digital resources integration based on cloud computing was designed. A small cloud computing cluster was sets up and the configuration was given. An experiment of cluster's I/O performance was conducted to verify the feasibility of this framework in the research of digital resources integration.

Keywords：cloud computing；digital resource；integration

高校校园网的迅速发展，促进了以校园网为基础的教育教学、科学研究、文化娱乐等方面的发展. 而数字化资源的建设又是校园网发展的重要组成部分，各种数字化资源通过校园网可以跨越时间和空间的限制，满足师生对数字化资源的需要. 但是，随着数字化资源的不断积累，一些弊端也显现出来. 数字化资源本身具有形式多样化、数据量大、分布较为分散等特点[1]，使得原先集中管理、集中存储的资源组织管理模式已经不能适应当前校园网的发展趋势，不能满足师生对资源应用的实际需求.

云计算作为一种新的服务形式能够很好解决这些问题. 将云计算应用于数字化资源的整合中，既可以实现对数字化资源整合的自动化，便于师生更为快速的搜索并获取到所需资源，实现数字化资源的共享，又为学校降低了投资成本，简化了对数字化资源的整合和管理工作[2].

本文在分析高校数字化资源的现状及所面临的问题基础上，给出基于云计算的数字化资源整合系统方案，并进行了测试.

1 高校数字化资源现状

各高校在加快信息化建设、数字化校园建设中，建有众多数字化资源平台，如精品课程网站、FTP 服务、音视频点播系统、个人网络硬盘等，存储的数据规模通常可以达到 TB 乃至 PB 级别. 但由于缺乏统一的组织与协作，资源的建设和系统的开发实施都是各自为政，缺乏共建共享的合作. 数字化资源应用现

收稿日期：2010–11–09；**修回日期**：2011–02–24
作者简介：熊聪聪（1961—），女，四川人，教授，xiongcc@tust.edu.cn.

状不尽如人意，主要体现在如下几个方面：(1) 资源分布广泛，存储管理分散. 各类数字化资源各不相关，购买的众多厂商的各类资源. 自主研发的各类系统及其相关的数字资源，这些异构的资源没有有效的整合，无法实现资源在不同层面的转换和集成，无法实现不同资源之间的共享[3]. (2) 资源更新成本高. 技术飞速进步往往要求学校能够为师生提供更多更新的数字化资源，资源的数据量呈线性增长，意味着需要更多的存储设备投入和更多的机房环境设备投入，以及运行维护成本和人力成本的增加. (3) 资源存储安全性要求较高. 对于资源的保存体现为不仅要保存资源载体本身，还要保存资源得以重现的设备和软件，使师生可以随时存取和使用分散分布在校园网中的资源，最终实现各种教育教学资源的高效管理.

与上述问题并存的还包括数字资源的加工、存储和检索[4]等问题. 同时，更丰富的数字化资源必然带来更多的用户和访问，如何合理构建具有开放性、高可用性、高性能、高安全性、高可伸缩性的分布式数字化资源环境，将分布于不同区域的信息资源有机聚合在一起，满足师生日益增长和深化的信息需求，是高校亟待解决的难题之一.

在此现实环境下，建设以统一规划、统一开发、分工建设、共建共享为原则的分布式数字化资源中心，成为解决当前面临问题的最有效途径.

2　云计算

云计算体现为“云平台”，是分布式并行计算存储系统. 在云计算的帮助下，高校数字化资源中心将具备超大处理能力、更高的可靠性、更好的安全性、更低的硬件成本，以及智能化、全自动化的任务管理与调度能力[5].

2.1　云计算核心要素

高校聚集有大量的数字资源，云计算可以将资源无缝隙地提供给师生使用，师生以享用服务的形式享用这些资源，这便是云计算的三个核心要素——资源集中、能力发布和服务模式.

资源集中是指通过把各种资源聚集起来，形成相应的“云”. 例如，把校园网内所有能搜索到的资源以特定的形式汇集起来，从而让师生更方便、迅速地获取所需资源. 这种将资源搜索结果预先汇聚起来的模式便是云计算在资源整合方面的体现.

能力发布是指只有把计算与处理能力以及相应的资源发布，让师生共享，才能实现云计算在数字化资源整合方面的应用. 如果资源和计算能力不能发布供师生使用，是没有实际意义的.

服务模式是指云计算以服务形式而非技术形式面对用户. 用户只需关心可以享受到什么样的服务，而无需关心服务背后涉及什么资源，资源存放在什么位置. 云计算的服务模式大致分为三类：将基础设施作为服务 IaaS、将平台作为服务 PaaS 和将软件作为服务 SaaS[6].

2.2　Hadoop 与 HDFS

开源云计算系统 Hadoop[7]是 Apache 开源组织的一个分布式计算框架，可以在大量廉价的硬件设备组成的集群上运行应用程序，可为应用程序提供稳定可靠的接口，用于构建具有高可靠性和良好扩展性的分布式系统.

HDFS (Hadoop distributed file system) 是 Hadoop 的分布式文件系统，基于云计算的数字资源整合模型主要借助 HDFS 的优点来实现. HDFS 具有高容错性，适合大数据集的应用，有高吞吐率的数据读写能力. HDFS 能够提供对数据可扩展的访问，通过简单地往集群里添加节点便可解决大量客户端同时访问的问题. HDFS 采用 master/slave 架构，其典型部署是在一个 master 上运行 NameNode，集群中其他 PC 各运行一个 DataNode；也可以在运行 NameNode 的 master 上同时运行 DataNode，或者在一台 PC 上运行多个 DataNode[8]. HDFS 采取了副本策略，其目的是为了提高系统的可靠性和可用性.

HDFS 采用副本存放的策略. 大型 HDFS 集群一般运行在多个机架上，不同机架上的机器之间通过交换机进行通信. 一般情况下，HDFS 的副本个数为 3，其中一个副本放在本节点上，另一个副本放在同一机架中的另一个节点上，最后一个放在不同的机架中的另一个节点上. 这个策略可以防止整个机架失效时数据丢失，不会影响到数据的可靠性和可用性，又能保证性能.

3　数字资源整合模型与实现

在使用云计算整合数字资源之前，首先要构建一个云计算集群. 在这个云计算环境中，由动态可扩展的和虚拟化的计算资源来提供数字资源存储和访问服务. 通过云计算，可以将庞大的数字资源自动分拆

成大量较小的数据块，交由多个节点所组成的计算机集群系统进行分散存储.

3.1　整合模型

数字资源通过云计算进行整合后的模型结构如图 1 所示.

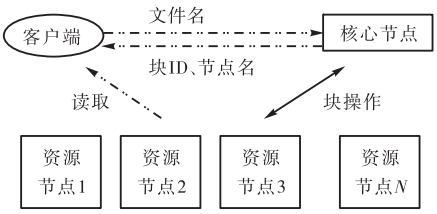

图 1　整合模型结构示意图

Fig.1　Architecture of integration model

从图 1 可知数字资源整合模型的云处理过程：客户端向核心节点请求资源；核心节点根据资源目录查找到该资源所在的多个节点和相应的块 ID，并返回给客户端；客户端从相关资源节点取得资源.

基于云计算的数字资源整合模型可以划分为 5 层，自上而下分别为客户端、资源门户层、应用管理层、存储层和硬件设备层.

客户端用来访问云计算各种应用服务的门户页面. 客户端通常是指各种 Web 浏览器，比如 IE、Firefox、Chrome 等.

资源门户层由 Portal Server 和 Portlet 容器组成. Portal Server 负责接收来自客户端的 HTTP 请求，在 Porlet 容器中调用 Portlet，并将 Portlet 产生的内容聚合到门户页面返回给客户端. 门户页面一般是由多个 Portlet 组件组成，每个 Portlet 提供一种云服务，显示相应的服务内容.

应用管理层负责各种云服务的具体实现，包括用户的管理、资源的整合、资源的管理、资源目录的管理等.

存储层负责将硬件设备层的存储资源虚拟成一个分布式文件系统，提供数据的分布式存储. 如使用 Hadoop 的分布式文件系统（DFS）.

硬件设备层主要包括各种计算资源及存储资源，如服务器、存储柜、交换机等.

3.2　模型的实现

使用 7 台服务器构建小型云计算集群环境，其中 1 台配置为 NameNode，7 台（包括 NameNode）配置为 DataNode. 服务器配置均为双核 CPU，4 GB 内存，1 TB 硬盘，1 000 MB 全双工网卡，之间通过一台 1 000 MB 交换机相联.

7 台服务器所用操作系统为 Red Hat Enterprise Linux AS4 Update8，在服务器上安装 JDK1.6.0_10、Open SSH 3.9p1 和 Hadoop-0.20.1.

构建 Hadoop 云计算集群环境，需要在各服务器上进行如下配置：编辑 hosts 文件，设置集群内各服务器的主机名和 IP 地址；在各服务器上创建相同的用户和 hadoop 目录结构；配置 ssh 保证各服务器之间使用无密码公钥认证的方式来访问；编辑 Hadoop 配置文件 hadoop-env.sh，设置 Java_Home 路径；编辑 Hadoop 配置文件 core-site.xml，设置 fs.default.name 为 NameNode 的 IP 地址和端口号；编辑 Hadoop 配置文件 hdfs-site.xml，设置 dfs.replication 为 HDFS 中每个数据块被复制的次数；编辑 Hadoop 配置文件 mapred-site.xml，设置mapred.job.tracker为 JobTracker 的 IP 地址和端口号；编辑 Hadoop 配置文件 master 和 slaves，设置 NameNode 和 DataNode 的主机名.

结合 Hadoop 云计算环境和已有的各数字资源应用系统，进行整合开发，实现自动同步资源目录、自动整理资源. 用户使用浏览器访问云计算数字资源整合的门户页面，经过身份认证后，可在门户中选择自己感兴趣的资源，也可以通过提供的搜索引擎查找云计算集群中整合的数字资源. Portlet 会将用户的资源访问请求发送给 NameNode，NameNode 返回该资源的数据块所在的所有 DataNode. Portlet 从 DataNode 中选取合适的节点并且并行地发出读请求，不同的请求发送到不同的 DataNode. Portlet 从不同的 DataNode 接收到数据后进行内容汇聚，然后发送给用户.

3.3　性能分析

当前高校数字资源的使用方式，从应用层面分析可分为，文件在线浏览编辑、视频点播、HTTP 下载/上传、FTP 下载/上传、BT 下载、PT 下载、搜索引擎、RSS 订阅资源、网络硬盘等多种多样，都是文件的读写操作.

在云计算系统的实际运行中有大量的并发用户读取资源，为此本文对云计算系统的输入输出性能进行实验. 利用已搭建的小型云计算集群环境，分别实验在不同并发数、不同文件大小条件下写入和读取资源的性能，并以运行时间作为对比依据. 实验基于一个 MapReduce 任务，其中每个Map打开一个文件，并行执行读或写操作并测量I/O 时间，通过一个 Reduce 过程汇聚所有的I/O测量数据，并计算平均值.

实验结果见表 1 和表 2. 可以看出：(1) Hadoop

会根据数据量的大小及并发数来安排集群中的部分机器参与执行任务，当并发数小于 10、数据量小于 100 MB 时，任务实际运行时间较少，但任务初始化时间、中间文件的生成与传输时间相对较多，此时集群未能发挥分布式优势，所以总运行时间差异不大. (2)随着输入输出文件数据量增大、并发数增多，云计算系统将任务细分并分配给不同的 DataNode 处理，此时并行框架的优势得以发挥，任务运行时间远大于分布式系统通信的各种耗时. 通过实验数据可以看出，并发数变化对任务运行时间的影响要大于数据量变化. (3)由于实验环境的硬件条件约束，当并发数、数据量增加到一定阀值后，云计算系统的输入输出性能有所下降，表现为总运行时间成倍增长. 可以通过增加 DataNode 节点、升级各节点硬件来提高这一阀值，以充分发布云计算架构的优势.

表 1　读取数据耗时

Tab.1　Elapsed time of reading data　　ms

文件大小/MB	并发数量			
	1	10	100	1 000
1	23 119	26 125	65 497	415 117
10	24 088	27 037	69 465	526 107
100	24 176	27 163	88 673	1 040 259
1 000	38 200	113 475	164 728	–

表 2　写入数据耗时

Tab.2　Elapsed time of writing data　　ms

文件大小/MB	并发数量			
	1	10	100	1 000
1	23 141	29 286	68 979	519 334
10	24 126	31 432	76 211	557 946
100	26 187	43 366	92 647	2 423 379
1 000	62 339	189 125	274 548	–

4　结　语

本文针对高校目前的数字化资源现状，基于云计算实现了数字资源整合系统，师生们可以在统一界面上检索并获取各种类型数据系统中的数字化资源. 该系统解决了资源共享中存在的诸多问题，可以提高数字化资源的使用率和准确度，有利于提高高校的数字化资源整合水平. 在实际应用中，要充分考虑如何发挥集群中每个节点的性能；考虑在不同数量级的数据处理需求下，如何得到一个适合的高性能集群，还需进一步研究.

参考文献：

[1] 孔繁之，王春梅，彭才洪，等. 数字校园中教学资源库的建设与应用研究[J]. 中国教育信息化，2008(23)：49–50.

[2] Armbrust M，Fox A，Griffith R，et al. A view of cloud computing[J]. Communications of the ACM，2010，53(4)：50–58.

[3] Richard A Brown. Hadoop at home：Large-scale computing at a small college[C]//Proceedings of the 40th ACM Technical Symposium on Computer Science Education. New York：Association for Computing Machinery，2009：106–110.

[4] 朱俊，严明. 企业数字资源整合系统的设计与实现[J]. 情报杂志，2010，29(5)：183–187.

[5] 王庆波，金滢，何乐，等. 虚拟化与云计算[M]. 北京：电子工业出版社，2009.

[6] 叶伟，赵进，叶军，等. 互联网时代的软件革命：SaaS 架构设计[M]. 北京：电子工业出版社，2009.

[7] Jason Venner. Pro Hadoop[M]. Berkeley，CA：Apress，Inc，2009.

[8] 刘鹏. 云计算[M]. 北京：电子工业出版，2009.