应用型本科 物联网专业“十三五”规划教材

云计算与大数据

主编 陶皖

- 内容新颖：新知识、新技术、新工艺
- 特色鲜明：突出“应用、实践、创新”
- 定位准确：面向工程技术型人才培养
- 质量上乘：应用型本科专家全力打造

西安电子科技大学出版社
http://www.xduph.com

应用型本科 物联网专业“十三五”规划教材

云计算与大数据

主　编　陶　皖
副主编　李　钧　李臣龙

西安电子科技大学出版社

内容简介

本书在阐述大数据和云计算关系的基础上，介绍了大数据和云计算的基本概念、技术及应用。全书分为基础篇、技术与应用篇和实践篇，主要内容包括：绪论、大数据环境下的云计算架构、大数据关键技术与应用、云存储、云服务与云安全、云计算应用、虚拟化技术、Hadoop和Spark平台、分布式文件系统及并行计算框架、分布式数据存储与大数据挖掘。

本书是结合实际应用及实践过程来讲解相关概念、原理和技术的，实用性较强，适合作为本科院校计算机、大数据及信息管理等相关专业的教材，也可作为相关研究人员、爱好者的参考用书。

图书在版编目(CIP)数据

云计算与大数据/陶皖主编. —西安：西安电子科技大学出版社，2017.1
应用型本科 物联网专业“十三五”规划教材
ISBN 978-7-5606-4377-9

Ⅰ. ① 云… Ⅱ. ① 陶… Ⅲ. ① 云计算 ②数据处理
Ⅳ. ① TP393.027 ②TP274

中国版本图书馆 CIP 数据核字(2016)第 320924 号

策划编辑 高 樱
责任编辑 杨 璠
出版发行 西安电子科技大学出版社(西安市太白南路 2 号)
电 话 (029)88242885 88201467 邮 编 710071
网 址 www.xduph.com 电子邮箱 xdupfxb001@163.com
经 销 新华书店
印刷单位 陕西华沐印刷科技有限责任公司
版 次 2017 年 1 月第 1 版 2017 年 1 月第 1 次印刷
开 本 787 毫米×1092 毫米 1/16 印张 14.5
字 数 341千字
印 数 1～3000 册
定 价 29.00 元
ISBN 978-7-5606-4377-9/TP

XDUP 4669001-1

西安电子科技大学出版社
应用型本科　物联网专业“十三五”规划教材
编审专家委员名单

主　任： 吴其胜（盐城工学院 材料工程学院 院长/教授）

副主任： 丁红燕（淮阴工学院 机械工程学院 院长/教授）
杨　莉（常熟理工学院 机械工程学院 副院长/教授）
朱协彬（安徽工程大学 机械与汽车工程学院 副院长/教授）

成　员：（按姓氏拼音排列）
陈　南（三江学院 机械学院 院长/教授）
胡爱萍（常州大学 机械工程学院 副院长/教授）
刘春节（常州工学院 机电工程学院 副院长/副教授）
卢雅琳（江苏理工学院 材料工程学院 院长/教授）
王荣林（南理工泰州科技学院 机械工程学院 副院长/副教授）
王树臣（徐州工程学院 机电工程学院 副院长/教授）
王章忠（南京工程学院 材料工程学院 院长/教授）
叶原丰（金陵科技学院 材料工程学院 副院长/副教授）
吴懋亮（上海电力学院 能源与机械工程学院 副院长/副教授）
吴　雁（上海应用技术学院 机械工程学院 副院长/副教授）
徐启圣（合肥学院 机械系 副主任/副教授）
张可敏（上海工程技术大学 材料工程学院 副院长/教授）
张晓东（皖西学院 机电学院 院长/教授）

前　言

随着互联网、移动互联网、物联网的快速发展，以及社交网络、微博、微信等新一代信息技术的应用和推广，人类产生的数据正以几何级数的速度增长。不言而喻，人类已进入了大数据时代。数据的种类繁多、流动迅速，数据中蕴含的价值也越来越受到人们的重视。需求驱动技术的发展，为了应对海量的数据及对其应用处理的要求，近年来，云计算和各类大数据技术层出不穷。本书将视角放在云计算和大数据技术上，通过介绍云计算和大数据的概念、技术现状及其产业发展趋势，为读者提供了该领域的基础性知识；同时，本书还结合实际应用为读者提供了应用实践指南。

全书分 3 篇，共 10 章。

第一篇(基础篇)简单介绍了云计算和大数据的起源，并对云计算和大数据的定义、特征和作用作了较详细的说明，同时还介绍了云计算和大数据之间的关系，阐述了大数据环境下的云计算架构。

第二篇(技术与应用篇)详细介绍了大数据存储、处理及分析等关键技术，云存储、云服务与云安全以及云计算应用的知识。

第三篇(实践篇)以开源的技术解决方案为主体，从虚拟化技术、Hadoop 和 Spark 平台、HDFS 文件系统、分布式文件存储 Hbase、Hive 和大数据挖掘分析平台 Mahout 等角度介绍了云计算和大数据的实践过程。第三篇还设置了 7 个实验，以使读者能理论联系实践，进一步熟悉云计算和大数据的概念，并在构建自己的应用时获得启发和帮助。

本书的部分实验素材由李臣龙老师提供，李钧老师编写了第三篇中的部分章节，其余章节由陶皖老师编写。本书的编写得到了西安电子科技大学出版社高樱老师、安徽工程大学计算机与信息学院等的大力帮助和支持，在此表示诚挚的谢意。

由于编者水平有限，书中难免存在不妥之处，敬请广大读者批评指正。

编　者

2016 年 9 月

目　录

第一篇　基　础　篇

第二篇　技术与应用篇

第三篇　实　践　篇

第一篇　基础篇

云计算的概念与特征

大数据的概念与特征

大数据下的云计算架构

第1章 绪 论

云计算(Cloud Computing)和大数据(Big Data)的概念在IT从业人员中耳熟能详，它们的应用正席卷IT行业的各个方面。本章将探究云计算和大数据的背景及来历，介绍大数据和云计算的概念和特征、应用和发展，并对大数据和云计算的关系作出阐述。

1.1 云计算的来历与发展

1.1.1 云计算的萌发

如今“大红大紫”的云计算概念实际上起源于20世纪60年代，在那个绝大多数人还没有用过计算机的时代，来自斯坦福大学的科学家John McCarthy就指出“计算机可能变成一种公共资源”。同时代的Douglas Parkhill在其著作《The Challenge of the Computer Utility》中将计算资源类比为电力资源，并提出了私有资源、公有资源、社区资源等在今天被频繁提起的云计算概念。

这些事实让我们不得不承认人类的想象力和智慧是推动世界进步的巨大动因，同时也可以看到云计算不是一个偶然的技术产物，它可以说是计算机技术演进的必然方向。

1.1.2 云计算的诞生

现代的云计算模式诞生于20世纪90年代末的互联网大潮。1997年，Ramnath Chellappa教授在一次演讲中第一次提出了“云计算”这个词；1999年成立的Salesforce.com公司是公认的云计算先驱，它主要向企业客户销售基于云的SaaS(Software as a Service，软件即服务)产品。

Salesforce.com公司的成功之处在于它第一次证明了基于云的服务不仅仅是大型业务系统的廉价替代品，还可以为提高企业运营效率、促进业务发展提供解决方案，同时也可以在可靠性方面维持一个极高的标准。此后，许多苛刻的企业用户开始全面拥抱云计算，从而迎来了云计算的发展高潮。

1.1.3 云计算的发展

进入21世纪的第一个十年，Amazon接棒Salesforce.com推动了云计算的快速发展。Amazon是在线零售商，非常重视客户体验，在发展到一定规模后，它发现自己的数据中心在大部分时间只有不到10%的利用率，剩下90%的资源都闲置着。于是Amazon开始寻找一种更有效的方式来利用自己庞大的数据中心，它的目的是将计算资源从单一的、特定的业务中解放出来，在空闲时提供给其他有需要的用户。Amazon首先在内部实施这一计划，得到了不错的反响，接着将这个服务开放给外部用户，并命名为AWS(Amazon Web Service，亚马逊网络服务)。初期的AWS是一个简单的线上资源库，并没有引起太大的关注，令AWS

声名大噪的是 Amazon 在 2006 年发布的 EC2(Elastic Compute Cloud)。EC2 是业界第一款面向公众提供基础构架的云服务产品，它将云计算的服务对象带入了更广阔的领域。除 EC2 之外，Amazon 还发布了 S3、SQS 等其他云计算服务，组成了一个完整的 AWS 产品线，促使云计算成为 IT 产业的主流声音。

继 Amazon AWS 之后，各种云计算产品层出不穷，Microsoft、Google 等巨头纷纷涌进这个领域。除了数量的增长，云计算类型也日益丰富。除了 Salesforce.com 和 Amazon AWS 分别代表的 SaaS 和 IaaS(Infrastructure as a Service，基础设施即服务)两种云计算服务，第三种服务 PaaS(Platform as a Service，平台即服务)也快速发展起来了，如 2009 年发布的 Google App Engine 服务。此外，围绕在线资源的应用亦快速出现。2009 年，第一个基于 Amazon AWS API(Application Program Interface)的私有云平台 Eucalyptus 出现。同年，信息研究机构 Gartner 预测企业用户将从基于设备的 IT 建设模型往基于单个用户需求的云计算模型转变。

进入 21 世纪的第二个十年，云计算进入了百花齐放的时代。人们已经不再讨论云计算是否可行，而是探讨起云计算未来的发展方向，研究在大数据时代怎样将云计算的潜力充分发挥出来，从而更好地利用数据的价值。图 1-1 所示为云计算与“数字一代”的示意图。

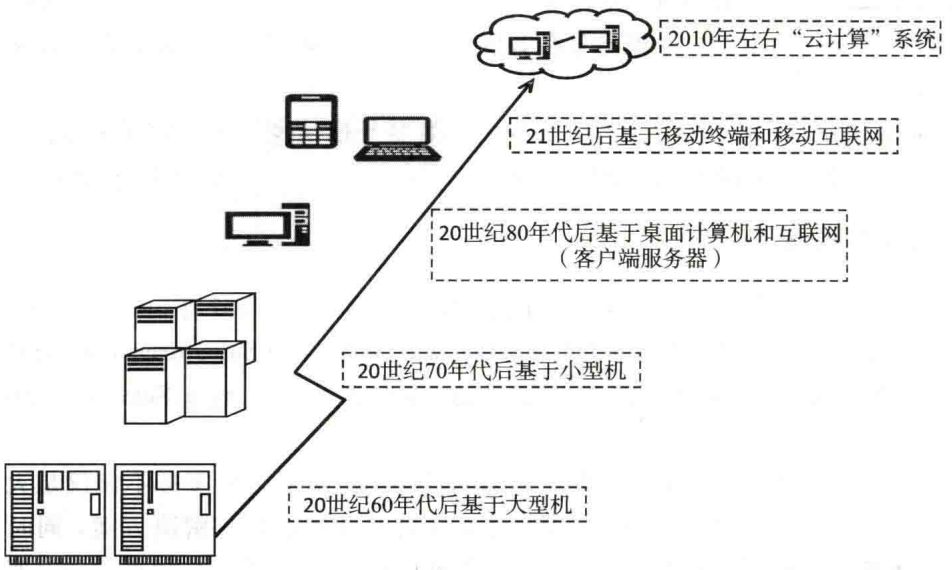

图 1-1 云计算与“数字一代”

百度百科中给出的云计算发展简史：

• 1983 年，太阳电脑(Sun Microsystems)提出“网络是电脑(The Network is the Computer)”，2006 年 3 月，亚马逊(Amazon)推出弹性计算云(Elastic Compute Cloud，EC2)服务。

• 2006 年 8 月 9 日，Google 首席执行官埃里克·施密特(Eric Schmidt)在搜索引擎大会(SES San Jose 2006)首次提出“云计算(Cloud Computing)”的概念。Google“云端计算”源于 Google 工程师克里斯托弗·比希利亚所做的“Google 101”项目。

• 2007 年 10 月，Google 与 IBM 开始在美国大学校园，包括卡内基梅隆大学、麻省理工学院、斯坦福大学、加州大学柏克莱分校及马里兰大学等，推广云计算的计划，希望能通过这项计划降低分布式计算技术在学术研究方面的成本，并为这些大学提供相关的软硬件设备

及技术支持(包括数百台个人电脑及 BladeCenter 与 System x 服务器，这些计算平台将提供 1600 个处理器，支持包括 Linux、Xen、Hadoop 等开放源代码平台)。而学生则可以通过网络开发各项以大规模计算为基础的研究计划。

• 2008 年 1 月 30 日，Google 宣布在台湾启动“云计算学术计划”，将与台湾台大、交大等学校合作，从而将这种先进的云计算技术大规模、快速地推广到校园。

• 2008 年 2 月 1 日，IBM(NYSE：IBM)宣布将在中国无锡太湖新城科教产业园为中国的软件公司建立全球第一个云计算中心(Cloud Computing Center)。

• 2008 年 7 月 29 日，雅虎、惠普和英特尔宣布了一项涵盖美国、德国和新加坡的联合研究计划，推出云计算研究测试床，以推进云计算。该计划要与合作伙伴创建 6 个数据中心作为研究试验平台，每个数据中心配置 1400 个至 4000 个处理器。这些合作伙伴包括新加坡资讯通信发展管理局、德国卡尔斯鲁厄大学 Steinbuch 计算中心、美国伊利诺伊大学香槟分校、英特尔研究院、惠普实验室和雅虎。

• 2008 年 8 月 3 日，美国专利商标局网站信息显示，戴尔正在申请“云计算(Cloud Computing)”商标，此举旨在加强对这一未来可能重塑技术架构的术语的控制权。

• 2010 年 3 月 5 日，Novell 与云安全联盟(CSA)共同宣布一项供应商中立计划，名为“可信任云计算计划(Trusted Cloud Initiative)”。

• 2010 年 7 月，美国国家航空航天局和包括 Rackspace、AMD、Intel、戴尔等支持厂商共同宣布“开放源代码(OpenStack)”计划。微软在 2010 年 10 月表示支持 OpenStack 与 Windows Server 2008 R2 的集成，而 Ubuntu 已把 OpenStack 加至 11.04 版本中。

• 2011 年 2 月，思科系统正式加入 OpenStack，重点研制 OpenStack 的网络服务。

1.2 云计算的概念及特征

1.2.1 什么是云计算

什么是云计算？这是一个被反复提到、反复回答的问题，这说明云计算本身是一个非常抽象的概念，要准确把握其内涵不是一件容易的事情。

云计算的解释有许多种。

• 百度百科中的解释是：云计算是基于互联网的相关服务的增加、使用和交付模式，通常涉及通过互联网来提供动态易扩展且经常是虚拟化的资源。云是网络、互联网的一种比喻说法。过去往往用云来表示电信网，后来也用来表示互联网和底层基础设施的抽象。因此，云计算甚至可以让你体验到 10 万亿次每秒的运算能力，拥有这么强大的计算能力可以模拟核爆炸、预测气候变化和市场发展趋势。用户通过电脑、笔记本、手机等方式接入数据中心，按自己的需求进行运算。

• 维基百科中的解释是：云计算是继 20 世纪 80 年代大型计算机到客户端-服务器的大转变之后的又一种巨变，是一种基于互联网的计算方式。通过这种计算方式，共享的软硬件资源和信息可以按需求提供给计算机和其他设备。用户不再需要了解“云”中基础设施的细节，不必具有相应的专业知识，也无需直接进行控制。云计算描述了一种基于互联网的新的 IT 服务增加、使用和交付模式，通常涉及通过互联网来提供动态易扩展且经常是虚拟化的资源。

• CSA(Cloud Security Alliance，云计算安全联盟)在“Security Guidance For Critical Areas of Focus In Cloud Computing V3.0”的解释是：云计算的本质是一种服务提供模型，通过这种模型可以随时、随地、按需地通过网络访问共享资源池的资源，这个资源池的内容包括计算资源、网络资源、存储资源等，这些资源能被动态地分配和调整，在不同用户之间灵活地划分，凡是符合这些特征的IT服务都可以称为云计算服务。

CSA的解释很好地说明了云计算的本质。NIST(U. S. National Institute of Standards and Technology，美国国家标准与技术学院)提出了一个定义云计算的标准——“NIST Working Definition of Cloud Computing/NIST800－145”。此标准提出云计算的五大要素是：自助服务、通过网络分发服务、可衡量的服务、资源的灵活调度以及资源池化。云计算按服务模式分为三类：SaaS、IaaS和PaaS；按部署模式分为四种：公有云、私有云、社区云和混合云。NIST800－145被业界普遍接受的原因是其提出的云计算五大要素非常简练地说明了一个云计算系统的特征，通过这五个特征能够快速地将云计算系统同传统IT系统区分开来。图1－2所示为NIST提出的云计算概念图。

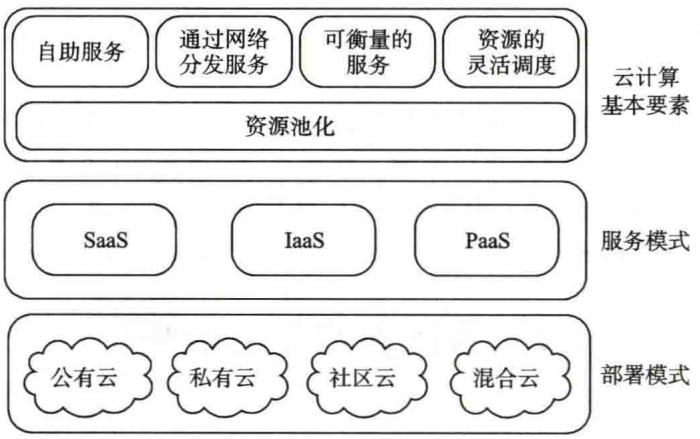

图1－2 NIST提出的云计算概念图

1.2.2 云计算的特征

本小节先介绍五大要素的具体内容，即云计算的五个特征，部署模式和服务模式将在第2章中作说明。

1. 自助服务

在云计算服务中，用户通过自助方式获取服务。自助服务是区分简单的B/S架构与真正云计算的重要标准。自助式的服务方式充分发挥了云计算后台架构强大的运算能力，也使用户获得了更加快捷、高效的体验。

例如WebEx公司推出的在线会议系统，可使用户自助挑选会议类型、设定参会人数、上传会议资料。WebEx的后台服务器会在指定时间将参会人员连接到一个虚拟在线会议室中。云服务提供商并不需要人工干预这个流程，所有的细节都由用户自己在网页上选择决定。而以往的会议服务仅仅将提供多方会议服务的设备从分散的用户机房集中到统一的中心机房，其软硬件设施仍然是僵化的传统框架，无法自动跟上用户需求的快速变更，需要人工干预才

能完成资源的重新划分和调整。

2. 通过网络分发服务

通过网络分发服务打破了地理位置的限制，打破了硬件部署环境的限制，实现了只要有网络就有计算，从而革命性地改变了人们使用电脑的习惯。

以书写为例，当云计算真正普及后，只需要在诸如 iPad 的终端上登录 Google Docs，就能够进行在线写作。Google Docs 提供了大部分的电子文档编辑功能，不需要新添置计算机，不需要购买 Office 软件，只需要通过任何带 Web 浏览器的设备就可进行工作。

3. 可衡量的服务

一个完整的云计算平台会对存储、CPU、带宽等资源保持实时跟踪，并将这些信息以可量化的指标反映出来。基于这些指标，云计算平台运营商或管理企业内部私有云的 IT 部门，能够快速地对后台资源进行调整和优化。

4. 资源的灵活调度

资源的灵活调度是资源池化的下一步发展。由于计算资源已经被池化，云计算供应商可以非常快速地将新设备添加到资源池中，以满足客户不断增长的需求。而对于用户来说，好像只要愿意付账单，就可以即时要求云计算供应商提供无限制的资源。

例如 WebEx、Amazon EC2 等总是可以满足用户不断增长的需求。在 WebEx 平台上，已经召开过同时上千人参与的全球视频会议，而 WebEx 表示人数仍然没有达到上限。

5. 资源池化

在云计算中，计算资源——CPU、存储、网络等有了新的组织结构，称为资源池。资源池化是指将所有设备的运算能力放在一个池内，再进行统一分配。对于 IT 部门来说，计算资源不再以单台服务器为单位，云计算打破了服务器机箱的限制，将所有的 CPU 和内存等资源解放出来，汇集到一起，形成一个个 CPU 池、内存池、网络池，当用户产生需求时，便从这个池中配置能够满足需求的组合。

NIST 提出的五个特征非常形象地提炼出不同云计算模式的共性。在绝大部分获得成功的云服务身上，都能轻易找到这五点特征。Amazon AWS 的 EC2 就是一个典型的例子。Amazon EC2 的服务全部可以在 Amazon AWS 的网站上自助开通，用户通过网络可获取 Amazon EC2 的后台计算资源。Amazon EC2 有一个完善的后台管理系统，能够在不同的数据中心之间调配资源，以满足瞬息万变的用户需求。这些服务特点将 Amazon EC2 塑造成一个优秀的、成功的云服务提供商，也界定出优秀云计算服务的基本模型。

1.3 云计算的应用及与其他计算服务模式的区别

1.3.1 云计算的应用

云计算为用户提供动态、可扩展的计算资源，也就是说，用户享用的计算资源可以根据客户流量需要随时增减。云计算的特点对于现有的企业，特别是对计算资源要求随时间变化的企业具有相当大的吸引力。利用云计算的弹性资源，企业解决了因需求量突然增加而出现计算资源不足的问题，同时避免了因闲置过剩计算资源而造成的浪费。

云计算也特别适合刚刚起步的IT企业。新生的企业如果要提供网络服务，通常需要购买一定数量的服务器等硬件设备和软件，甚至还要招聘管理和支持这些服务器和设备的信息技术管理人员。这对新企业而言是一笔不小的启动资金。利用云计算服务，企业可以花费较少的资金从云计算服务商那里获得所需的网络计算资源，随着业务的发展，再决定是否逐步增加租用云计算服务，甚至设立自己的数据中心。如果企业决定改变经营方向，也不用丢弃现有设备另起炉灶，从而降低风险。

随着云计算的普及，人们开发的软件将会越来越多地借助互联网的强大功能，更多的软件将在互联网上直接为用户提供服务，这将给软件开发者(无论企业还是个人)带来他们的黄金时代。如果软件开发者有自己的思想和创意，那么在没有足够经费购买硬件和软件的情况下，借助云计算就有望开发出独特的软件。也就是说，云计算服务在软件开发方面将起到积极的推动作用，软件的开发也会借此东风向前迈进一大步。

但是，应该指出，不是所有的软件都需要搬到云计算中，云计算也不是对每个开发商都适用。从目前来说，对计算资源需求不大、所需资源没有大起大落的网上软件，云计算并不能带来特别的好处。此外，一些国家和地区有明确的法律和规章，不容许有关的数据信息存储在其他国家的数据中心。毫无疑问，云计算在这些国家和地区的使用将受到一定的限制。

1.3.2 云计算与其他计算服务模式的区别

1. 云计算与一般托管环境的区别

云计算和一般数据中心的服务器托管听起来很相似，但实际上存在着差别。

首先，工作环境建立有所不同。目前的数据中心提供的托管环境有共享的，也有专用的，有硬件服务器，也有虚拟服务器，但计算资源对于每个托管的软件都是有限的，如果需要更多的资源，就得增加服务器。而云计算的环境可以随时提供所需资源。例如，微软的云计算，开发者不需要和服务器直接打交道，而是与服务模块打交道。为了服务更多的客户，开发者只需指定有多少个软件同时运行。至于数据中心的服务器的启动和管理，由体系管理器来负责。

其次，两者的收费方式也有所不同。服务器托管服务环境通常是按月向用户收取固定费用，云计算服务商则根据计算的时间、信息存储量、计算量等向用户收费。当存储量增大、计算量增大或信息流量增大时收费也随之增加。

2. 云计算与网格计算的区别

Ian Foster将网格定义为：支持在动态变化的分布式虚拟组织间共享资源、协同解决问题的系统。所谓虚拟组织，就是一些个人、组织或资源的动态组合。图1-3和图1-4所示分别为云计算系统和网格计算系统的结构示意图。图1-3显示云计算系统采用以太网等快速网络将若干机群连接在一起，用户通过互联网获取云计算系统提供的各种数据处理服务。图1-4显示网格计算系统是一个资源共享模型，资源提供者也可以成为资源消费者。网格侧重研究怎样将分散的资源组合成动态虚拟组织。云计算与网格计算的一个重要区别在于资源调度模式：云计算采用机群来存储和管理数据，运行的任务以数据为中心，即调度计算任务到数据存储节点运行；而网格计算则以计算为中心，计算资源和存储资源分布在互联网的各个角落，不强调任务所需的计算和存储资源同处一地。由于网络带宽的限制，网格计算中的数据传输时间将占用总运行时间的很大一部分。

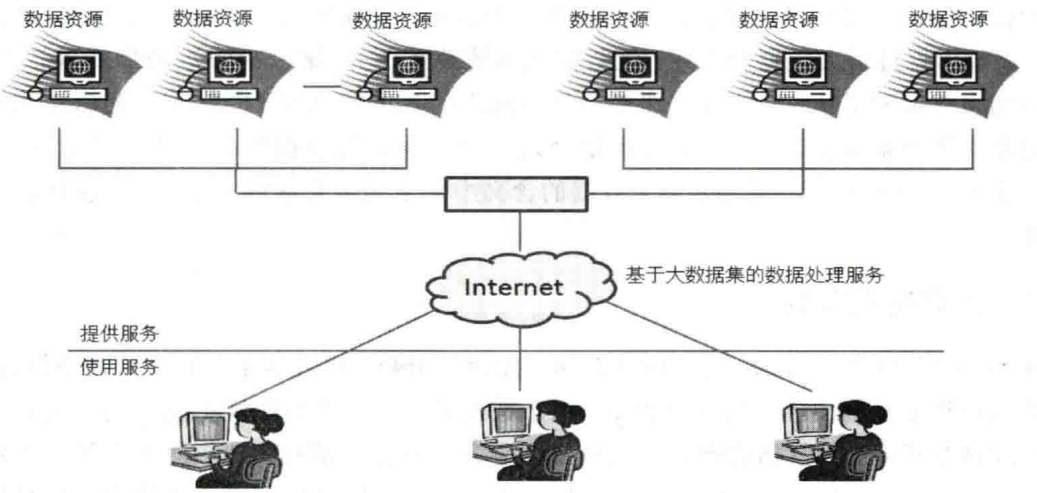

图 1-3 云计算系统的结构

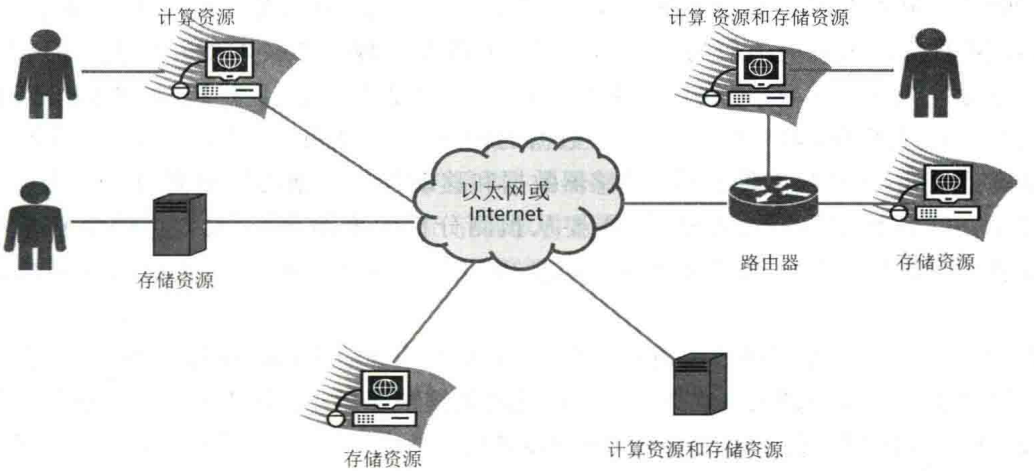

图 1-4 网格计算系统的结构

3. 云计算系统与传统超级计算机的区别

云计算系统和传统超级计算机最大的区别在于，超级计算机是应用超高的技术来实现一台超级计算机的强大处理功能的，而云计算系统则是用计算机机群来分工处理的。相对于超级计算机的研发困难来说，由更多的低配计算机(这里的低配计算机是相对于超级计算机来说的)来搭建处理中心相对来说要简单得多。如果这个云计算系统是由多台超级计算机组成的，那它的处理能力会成几何级数增长。相对于超级计算机要投入大量时间和资源的研发，云计算系统更适合社会应用。

1.4 大数据的提出及发展

1.4.1 大数据的提出

大数据一词源于英文词组“Big Data”，以往类似的词语如“信息爆炸”“海量数据”都很难

去准确描述这个词的具体内涵。早在1980年，著名未来学家阿尔文·托夫勒的《第三次浪潮》一书中，就出现过对大数据的表述，大数据被热情地赞颂为“第三次浪潮的华彩乐章”。但如果要追溯大数据的最初出处，就必然要提及Apache org的开源项目Nutch。当时大数据的意思是更新网络搜索索引，同时还需要批量处理和分析大量的数据集。谷歌的Map Reduce和Google File System(GFS)发布后，大数据的含义中除了涵盖大量数据之外，还包括数据处理的速度。

1.4.2 大数据的发展

在20世纪90年代后期，气象学家在作气象地图分析、物理学家在建立大物理仿真模型、生物学家在建立基因图谱的分析过程中，由于数据量巨大，他们已经不能再用传统的计算技术来完成这些任务，大数据的概念在这些科学研究领域首先被提出来。面对大量科学数据在获取、存储、搜索、共享和分析中遇到的技术难题，一些新的分布式计算技术被研究和开发出来了。

2008年，随着互联网和电子商务的快速发展，当雅虎、谷歌等大型互联网和电子商务公司用传统手段无法再解决他们的业务问题时，大数据的理念和技术就开始被他们接受和应用。此时的共性问题是，处理的数据量通常很大(那时是PB级，1个PB的数据相当于50%的全美学术研究图书馆藏书资讯内容)，数据的种类很多(文档、日志、博客、视频等)，数据的流动速度很快(包括流文件数据、传感器数据和移动设备数据的快速流动)。而且，这些数据经常是不完备甚至是不可理解的(需要从预测分析中推演出)。大数据的新技术和新架构正是在这种背景下被不断开发出来，以便能有效地解决这些现实中的互联网数据处理问题。

2010年，全球进入Web 2.0时代，Twitter(推特)、Facebook(脸书)、博客、微博、微信等社交网络将人类社会带入自媒体时代，互联网数据快速激增。随着苹果、三星等智能手机的普及，移动互联网时代也随之到来，移动设备所产生的数据也海量涌入网络。为了实现更加智能的应用，物联网技术也逐步被推广，随之而来的是更多实时获取的视频、电子标签(RFID)、传感器等数据也被联入互联网，数据量进一步暴增。根据美国市场调查公司IDC的预测，人类产生的数据量正在呈指数级增长，大约每两年翻一番，这个速度在2020年之前会一直保持下去。全球在2010年正式进入ZB时代(1 ZB数据相当于全世界海滩上的沙子数量的总和)，预计到2020年，全球将总共拥有35 ZB的数据量。这意味着人类在最近两年产生的数据量相当于之前产生的数据量总和。人类真正进入了一个数据的世界，大数据技术有了用武之地，大数据技术和应用空前繁荣起来。

2011年，全球著名战略咨询公司麦肯锡的全球研究院(MGI)发布了《大数据：创新、竞争和生产力的下一个新领域》研究报告，这份报告分析了数字数据和文档爆发式增长的状态，阐述了处理这些数据能够释放出的潜在价值，分析了大数据相关的经济活动和业务价值链。这篇报告在商业界引起极大的关注，为大数据从技术领域进入商业领域吹响了号角。

2012年3月29日，奥巴马政府以“大数据是一个大生意(Big Data is a Big Deal)”为题发布新闻，宣布投资2亿美元启动“大数据研究和发展计划”，该计划涉及包括美国国家科学基金、美国国防部在内的6个联邦政府部门，通过大力推动和改善与大数据相关的收集、组织和分析工具及技术，来推进从大量的、复杂的数据集合中获取知识和洞见的能力。这表明，