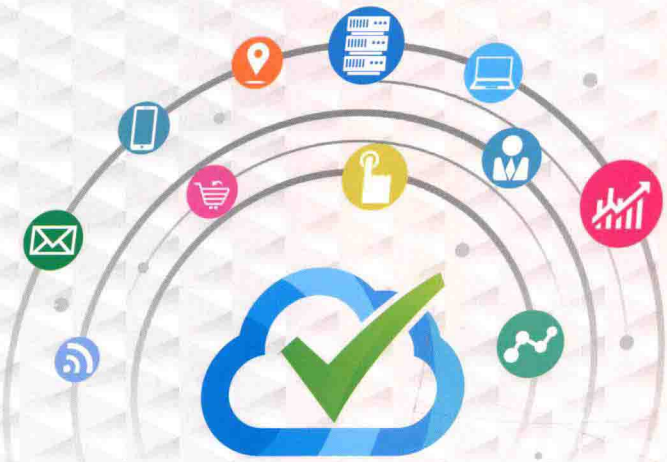
大数据技术与应用专业规划教材

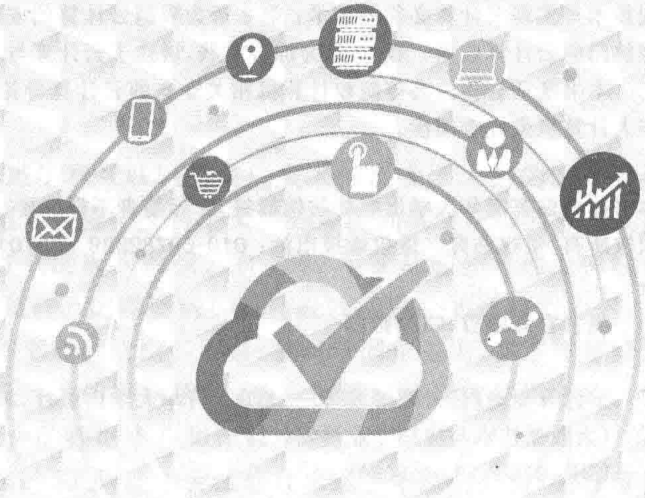


云计算导论

◎ 吕云翔 张璐 王佳玮 编著

清華大學出版社

大数据技术与应用专业规划教材



云计算导论

◎ 吕云翔 张璐 王佳玮 编著

清华大学出版社
北京

内 容 简 介

本书从云计算最基本的概念开始，由浅入深地带领读者领会云计算的精髓，以梳理知识脉络和要点的方式，带领读者登堂入室。

本书的第 1～3 章为云计算的基础部分，包括云计算的产生、发展、基本概念和实现云计算的机制部分；第 4～7 章为云计算的技术部分，包括虚拟化、分布式文件系统、分布式存储系统和数据处理与并行编程技术等实现云计算必须的技术；第 8 章为限制云计算的安全问题；第 9 章向读者提到了目前存在的一些热门的云计算应用；第 10 章为综合实践，讲述了云计算与 Docker 技术结合的实践内容。

本书既适合作为高等院校计算机相关专业的云计算导论课程的教材，也适合非计算机专业的学生及广大计算机爱好者阅读。

图书在版编目(CIP)数据

云计算导论/吕云翔等编著. —北京：清华大学出版社，2017
(大数据技术与应用专业规划教材)
ISBN 978-7-302-45805-0

Ⅰ. ①云…　Ⅱ. ①吕…　Ⅲ. ①云计算　Ⅳ. ①TP393.027

中国版本图书馆 CIP 数据核字(2016)第 290856 号

责任编辑：魏江江　梅栾芳
封面设计：刘　键
责任校对：徐俊伟
责任印制：宋　林

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课 件 下 载：http://www.tup.com.cn，010-62795954
印 装 者：清华大学印刷厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：12　　**字　　数**：303 千字
版　　次：2017 年 2 月第 1 版　　**印　　次**：2017 年 2 月第1 次印刷
印　　数：1～2000
定　　价：29.00 元

产品编号：065992-01

前　言

在过去的几十年里，计算机技术的进步和互联网的发展极大地改变了人们的工作和生活方式。计算模式也经历了从最初的把任务集中交付给大型处理机到基于网络的分布式任务处理再到目前的按需处理的云计算方式的极大改变。自 2006 年亚马逊公司推出弹性计算云(EC2)服务，让中小型企业能够按照自己的需要购买亚马逊数据中心的计算能力后，云计算的时代就此正式来临。“云计算”的概念随之由 Google 公司于同年提出，其本质是给用户提供像传统的电、水、煤气一样的按需计算的网络服务，是一种新型的计算使用方式。它以用户为中心，使互联网成为每一个用户的数据中心和计算中心。

Gartner 公司早在 2011 年 1 月发布的 IT 行业十大战略技术报告中就已经将“云计算技术”列为十大战略技术之首，目前世界上主要国家和跨国企业都积极地加快着战略部署，推动云计算的高速发展。我国也将云计算上升到了国家战略高度，中央、地方政府、产业界都在共同推动我国云计算的应用和发展，除此之外，像 Google、IBM、Microsoft、Amazon、阿里巴巴、腾讯等在内的知名 IT 企业也都在大力地开发相关云计算产品。

然而在教育普及方面，经作者调研，即使计算机相关专业的学生对于云计算的相关知识也知之甚少，而对用户而言，如果不提前进行了解就去使用云计算是很危险的事情。目前，虽然市场上关于云计算技术相关的书籍较多，但是适合读者进行云计算入门的书籍还较少，因此本书定位为“云计算导论”课程的专业教材，旨在传授读者云计算的基础入门知识，本书也适合非计算机专业学生以及广大的计算机爱好者阅读。

本书的章节内容如下：第 1～3 章为云计算的基础部分，包括云计算的产生、发展、基本概念和实现云计算的机制部分；第 4～7 章为云计算的技术部分，包括虚拟化、分布式文件系统、分布式存储系统和数据处理与并行编程技术等实现云计算必须的技术；第 8 章为限制云计算的安全问题；第 9 章向读者提到了目前存在的一些热门的云计算的应用；第 10 章为综合实践，讲述了云计算与 Docker 技术结合的实践内容。

在本书的编写过程中，尽量做到仔细认真，但由于作者的水平有限，还是可能出现一些不妥之处，在此非常欢迎广大读者进行批评指正。同时也希望广大读者将自己读书学习的心得体会反馈给作者(yunxianglu@hotmail.com)。

编　者

2016 年 8 月

目　录

第1章 云计算概论

本章介绍云计算的定义，旨在让读者对云计算有一个宏观的概念，然后介绍云计算的产生背景，接着介绍云计算的发展历史。通过本章的学习，能让读者对云计算有一个初步的认识。

1.1 什么是云计算

云计算(Cloud Computing)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每秒10万亿次的运算能力，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。用户可通过电脑、笔记本、手机等方式接入数据中心，按自己的需求进行运算。

对云计算的定义有多种说法。对于到底什么是云计算，至少可以找到100种解释。现阶段广为接受的是美国国家标准与技术研究院(NIST)的定义：云计算是一种按使用量付费的模式，这种模式提供可用的、便捷的、按需的网络访问，进入可配置的计算资源共享池(资源包括网络、服务器、存储、应用软件、服务)，这些资源能够被快速提供，只需投入很少的管理工作，或与服务供应商进行很少的交互。

1.2 云计算的产生背景

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。

云计算是分布式计算(Distributed Computing)、并行计算(Parallel Computing)、效用计算(Utility Computing)、网络存储(Network Storage Technologies)、虚拟化(Virtualization)、负载均衡(Load Balance)、热备份冗余(High Available)等传统计算机和网络技术发展融合的产物。

1.3 云计算的发展历史

1983年，太阳微系统公司(Sun Microsystems)提出“网络是电脑”的概念，2006年3月，亚马逊公司(Amazon)推出弹性计算云(Elastic Compute Cloud，EC2)服务。

2006年8月9日，Google公司首席执行官埃里克·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 公司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,这项计划希望能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 公司宣布在中国台湾启动“云计算学术计划”,与台湾台大、交大等学校合作,将云计算技术推广到校园的学术研究中。

2008 年 2 月 1 日,IBM 公司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、惠普和英特尔公司宣布一项涵盖美国、德国和新加坡的联合研究计划,推进云计算的研究进程。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 1400～4000 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宾分校、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 公司与云安全联盟(CSA)共同宣布一项供应商中立计划,名为“可信任云计算计划”。

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 OpenStack 开放源代码计划,微软公司在 2010 年 10 月表示支持 OpenStack 与 Windows Server 2008 R2 的集成;而 Ubuntu 已把 OpenStack 加至其 11.04 版本中。

2011 年 2 月,思科公司正式加入 OpenStack,重点研制 OpenStack 的网络服务。

2013 年,我国的 IaaS(基础设施即服务)市场规模约为 10.5 亿元,增速达到了 105%,显示出旺盛的生机。IaaS 相关企业不仅在规模、数量上有了大幅提升,而且吸引了资本市场的关注,UCloud、青云等 IaaS 初创企业分别获得了千万美元级别的融资。

过去几年里,腾讯、百度等互联网巨头纷纷推出了各自的开放平台战略。新浪 SAE 等 PaaS(平台即服务)的先行者也在业务拓展上取得了显著的成效,在众多互联网巨头的介入和推动下,我国 PaaS 市场得到了迅速发展,2013 年市场规模增长近 20%。但由于目前国内 PaaS 服务仍处于吸引开发者和产业生态培育的阶段,大部分 PaaS 服务都采用免费或低收费的策略,因此整体市场规模并不大,估计约为 2.2 亿元人民币左右,但这并不妨碍人们对 PaaS 的发展前景抱有充足的信心。

无论是国内还是全球,SaaS(软件即服务)一直是云计算领域最为成熟的细分市场,用户对于 SaaS 服务的接受程度也比较高。2015 年 SaaS 市场增长率将达到 117.5%,市场规模将增长至 8.1 亿元人民币。

2015 年以来,云计算方面的相关政策不断。2015 年初,国务院发布了《国务院关于促进云计算创新发展培育信息产业新业态的意见》,明确了我国云计算产业的发展目标、主要任

务和保障措施。2015 年 7 月，国务院又发布了《关于积极推进“互联网＋”行动的指导意见》，提出到 2025 年，互联网＋成为经济社会创新发展的重要驱动力量。2015 年 11 月，工业和信息化部印发《云计算综合标准化体系建设指南》。

1.4 如何学好云计算

云计算是一种基于互联网的计算方式，要实现云计算则需要一整套的技术架构，包括网络、服务器、存储、虚拟化等。云计算目前分为公有云和私有云。两者的区别只是提供服务的对象不同，一个是企业内部使用，一个则是面向公众。目前企业中的私有云都是通过虚拟化来实现的，建议可以了解一下虚拟化行业的前景和发展。

虚拟化目前分为服务器虚拟化（以 VMware 为代表）、桌面虚拟化（思杰要比 VMware 的优势要大）、应用虚拟化（以思杰为代表）。学习虚拟化需要的基础如下：

（1）操作系统，懂得 Windows 操作系统（如 Windows Server 2008、Windows Server 2003、Windows 7、Windows 8、Windows 10 等）的安装和基本操作、懂得 AD 域角色的安装和管理、懂得组策略的配置和管理。

（2）数据库的安装和使用（如 SQL Server）。

（3）存储的基础知识（如磁盘性能、RAID、IOPS、文件系统、FC SAN、iSCSI、NAS 等）、光纤交换机的使用、使用 Open-E 管理存储。

（4）网络的基础知识（如 IP 地址规划、VLAN、Trunk、STP、Etherchannel）。

1.5 小　　结

云计算作为一种新型的计算模式，利用高速互联网的传输能力将数据的处理过程从个人计算机或服务器转移到互联网上的计算机集群中，带给用户前所未有的计算能力。云计算的产生与发展，使用户的使用观念发生了彻底的变化，他们不再觉得操作复杂，他们直接面对的将不再是复杂的硬件和软件，而是最终的服务。云计算将计算任务分布在大量计算机构成的资源池上，使各种应用系统能够根据需要获取计算力、存储空间和各种软件服务。云计算现在还存在着一些问题，如数据安全问题、网络性能、互操作问题等，但是它的优点是毋庸置疑的。云计算不仅大大降低了计算的成本，而且也推动了互联网技术的发展。在众多公司和学者的研究下，未来的云计算将会有更好的发展。在不久的将来一定会有越来越多的云计算系统投入使用。通过对本章的学习，读者应该能对云计算有大体的了解，为后面章节的学习做好铺垫。

1.6 习　　题

1. 美国国家标准与技术研究院（NIST）是如何定义云计算的？
2. 云计算的发展历史经历了哪些过程？
3. 虚拟化指的是什么？

第2章 云计算基础

本章主要介绍关于云计算的各种基础知识，包括分布式计算、云计算的基本概念、实现云计算的几种关键技术以及云交付和部署模式，同时介绍云计算有哪些优势、面临的挑战以及几种典型的云应用。通过本章的学习，应能够对云计算有一个基本的认识。

2.1 分布式计算

分布式计算是一种计算方法，和集中式计算是相对的。随着计算技术的发展，一些应用需要巨大的计算能力才能完成，如果采用集中式计算，则需要耗费很长的时间才能完成。而分布式计算将应用分解成许多更小的部分，分配到多台计算机进行处理，这样可以节省整体计算时间，大大提高计算效率。云计算是分布式计算技术的一种，也是分布式计算这种科学概念的商业实现。

分布式计算的优点就是发挥“集体的力量”，将大任务分解成小任务，分配给多个计算节点同时去计算。分布式计算将计算扩展到多台计算机，甚至是多个网络，在网络上有序地执行一个共同的任务，当然离不开 Web 技术，但在分布式计算发展起来之前的网络协议并不能满足分布式计算的要求，于是产生了 Web Service 技术。

分布式计算的另一种应用是 Web Service，Web Service 是一个平台独立的、低耦合的、自包含的、基于可编程的 Web 的应用程序，可使用开放的 XML（标准通用标记语言下的一个子集）标准来描述、发布、发现、协调和配置这些应用程序，用于开发分布式的、互操作的应用程序。

如图 2.1 所示，Web Service 的体系结构是基于 Web 服务提供者、Web 服务请求者、Web 服务注册中心三个角色和发布、发现、绑定三个动作构建的。简单地说，Web 服务提供者就是 Web 服务的拥有者，等待为其他服务和用户提供自己已有的功能；Web 服务请求者就是 Web 服务功能的使用者，利用 SOAP 消息向 Web 服务提供者发送请求以获得服务；Web 服务注册中心的作用是把一个 Web 服务请求者与合适的 Web 服务提供者联系在一起，它充当管理者的角色，一般是 UDDI（Universal Description Discovery and Integration）。这三个角色是根据逻辑关系划分的，在实际应用中，角色之间很可能有交叉：一个 Web 服务既可以是 Web 服务提供者，也可以是 Web 服务请求者，或者二者兼而有之，显示了 Web 服务角色之间的关系，其中，“发布”是为了让用户或其他服务知道某个 Web 服务的存在和相关信息；“发现”是为了找到合适的 Web 服务；“绑定”则是在提供者与请求者之间建立某种联系。

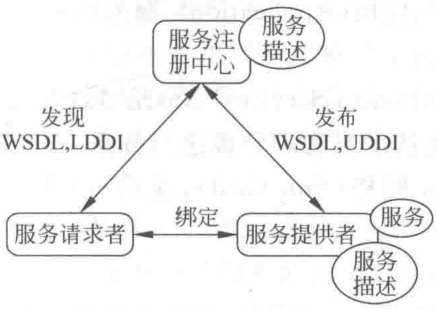


图 2.1　Web Service 的体系结构

简单地说，这种技术的功能和中间件的功能有相似之处：Web Service 技术是屏蔽掉不同开发平台开发的功能模块的相互调用的障碍，从而可以利用 HTTP 和 SOAP 协议使商业数据在 Web 上传输，可以调用这些开发平台不同的功能模块来完成计算任务。这样看来，要在互联网上实施大规模的分布式计算，就需要 Web Service 做支撑。

2.2　云计算的基本概念

云计算已经成为一个大众化的词语，似乎每个人对于云计算的理解各不相同，第 1 章已经对云计算有一个宏观的概念和通俗地理解，如图 2.2 所示，云计算的“云”就是存在于互联网上的服务器集群上的资源，它包括硬件资源（服务器、存储器、CPU 等）和软件资源（应用软件、集成开发环境等），本地计算机只需要通过互联网发送一个需求信息，远端就有成千上万的计算机为用户提供需要的资源并将结果返回给本地计算机。这样，本地计算机几乎不需要做什么，所有的处理都在云计算提供商所提供的计算机群来完成。简而言之，云计算是一种商业计算模型，它将计算任务分布在大量计算机构成的资源池上，使用户能够按需获取计算力、存储空间和信息服务。

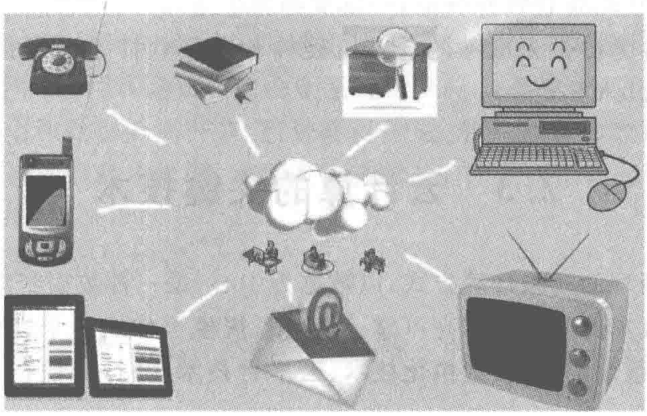


图 2.2　云计算

最简单的云计算技术在网络服务中已经随处可见，例如搜索引擎、网络信箱等，使用者只需要输入简单的指令即能得到大量信息。

云计算的组成可以分为六个部分，它们由下至上分别是：基础设施（Infrastructure）、存

储(Storage)、平台(Platform)、应用(Application)、服务(Services)和客户端(Clients)。

(1) 基础设施

云基础设施(Infrastructure as a Service,IaaS)是经过虚拟化后的硬件资源和相关管理功能的集合,对内通过虚拟化技术对物理资源进行抽象,对外提供动态、灵活的资源服务。具体用于如 Sun 公司的 Sun 网格(Sun Gird)、亚马逊(Amazon)的弹性计算云(Elastic Computer Cloud,EC2)。

(2) 存储

云存储涉及提供数据存储作为一项服务,包括类似数据库的服务,通常以使用的存储量为结算基础。全球网络存储工业协会(SNIA)为云存储建立了相应标准。它既可交付作为云计算服务,又可以交付给单纯的数据存储服务。具体应用如亚马逊简单存储服务(Simple Storage Service, S3)、Google 应用程序引擎的 BigTable 数据存储。

(3) 平台

云平台(Platform as a Service,PaaS)直接提供计算平台和解决方案作为服务,以方便应用程序部署,从而节省购买和管理底层硬件和软件的成本。具体应用如 Google 应用程序引擎(Google App Engine),这种服务让开发人员可以编译基于 Python 的应用程序,并可免费使用 Google 的基础设施来进行托管。

(4) 应用

云应用利用云软件架构,往往不再需要用户在自己的电脑上安装和运行该应用程序,从而减轻软件维护、操作和售后支持的负担。具体应用如 Facebook 的网络应用程序、Google 的企业应用套件(Google Apps)。

(5) 服务

云服务是指包括产品、服务和解决方案都实时地在互联网上进行交付和使用。这些服务可能通过访问其他云计算的部件,例如软件,直接和最终用户通信。具体应用如亚马逊简单排列服务(Simple Queuing Service)、贝宝在线支付系统(PayPal)、Google 地图(Google Maps)等。

(6) 客户端

云客户端包括专为提供云服务的计算机硬件和电脑软件终端,如苹果手机(iPhone)、Google 浏览器(Google Chrome)。

2.3 云计算的关键技术

云计算是一种新型的超级计算方式,以数据为中心,是一种数据密集型的超级计算。云计算的目标是以低成本的方式提供高可靠、高可用、规模可伸缩的个性化服务,要实现这个目标,需要分布式海量数据存储、虚拟化技术、云平台技术、并行编程技术、数据管理技术等若干关键技术支持。

2.3.1 分布式海量数据存储

随着信息化建设的不断深入,信息管理平台已经完成了从信息化建设到数据积累的职能转变,在一些信息化起步较早、系统建设较规范的行业,如通信、金融和大型生产制造等领

域，海量数据的存储、分析需求的迫切性日益明显。

以移动通信运营商为例，随着移动业务和用户规模的不断扩大，每天都产生海量的业务、计费以及网管数据，然而庞大的数据量使得传统的数据库存储已经无法满足存储和分析需求。主要面临的问题如下。

(1) 数据库容量有限

关系型数据库并不是为海量数据而设计，设计之初并没有考虑到数据量能够庞大到 PB 级。为了继续支撑系统，不得不进行服务器升级和扩容，成本高昂，难以接受。

(2) 并行取数困难

除了分区表可以并行取数外，其他情况都要对数据进行检索才能将数据分块，并行读数效果不明显，甚至增加了数据检索的消耗。虽然可以通过索引来提升性能，但实际业务证明，数据库索引作用有限。

(3) 针对 J2EE 应用来说，JDBC 的访问效率太低。

由于 Java 的对象机制，读取的数据都需要序列化，导致读数速度很慢。

(4) 数据库并发访问数太多

由于数据库并发访问数太多，导致 I/O 瓶颈和数据库的计算负担太重两个问题，甚至出现内存溢出崩溃等现象，但数据库扩容成本太高。

理想的解决方案是把大数据存储到分布式文件系统中，云计算系统由大量服务器组成，同时为大量用户服务，因此云计算系统采用分布式存储的方式存储数据，用冗余存储的方式(集群计算、数据冗余和分布式存储)保证数据的可靠性。冗余的方式通过任务分解和集群，用低配机器替代超级计算机的性能来保证低成本，这种方式保证分布式数据的高可用、高可靠和经济性，即为同一份数据存储多个副本。云计算系统中广泛使用的数据存储系统是 Google 的 GFS 和 Hadoop 团队开发的 GFS 的开源实现 HDFS。

2.3.2 虚拟化技术

虚拟化技术是云计算系统的核心组成部分之一，是将各种计算及存储资源充分整合和高效利用的关键技术。云计算的虚拟化技术不同于传统的单一虚拟化，它是涵盖整个 IT 架构的，包括资源、网络、应用和桌面在内的全系统虚拟化。通过虚拟化技术可以实现将所有硬件设备、软件应用和数据隔离开来，打破硬件配置、软件部署和数据分布的界限，实现 IT 架构的动态化，实现资源集中管理，使应用能够动态地使用虚拟资源和物理资源，提高系统适应需求和环境的能力。

虚拟化技术可以提供以下特点。

(1) 资源分享

通过虚拟机封装用户各自的运行环境，有效实现多用户分享数据中心资源。

(2) 资源定制

用户利用虚拟化技术，配置私有的服务器，指定所需的 CPU 数目、内存容量、磁盘空间，实现资源的按需分配。

(3) 细粒度资源管理

将物理服务器拆分成若干虚拟机，可以提高服务器的资源利用率，减少浪费，而且有助于服务器的负载均衡和节能。