## 第一章　云时代与大数据

**第一节 “云”与云计算**

**一、“云”及其产生的背景**

近些年，随着计算机及信息科技产业的发展，“云”开始出现并融入各行各业，如云思维、云营销、云原生、阿里云、百度云、云优先等，简直就是无所不用其“云”。

虽然有各种称谓的“云”，但“云”与“云”又有不同。例如，大家熟悉的网盘，现在都改称作云盘。百度云是网盘，而阿里云不是网盘，它是一个提供IT互联网技术资源的平台；天翼云是网盘，但是有一个也叫作“天翼云”的，提供的服务跟阿里云类似。现在政府牵头的“云”也特别多，如各级“政务云”“公安云”“消防云”“农业云”等。这些不同名称的“云”，让我们云里雾里，不知道究竟如何给“云”定义，让很多人认为，“云”就是“高大上”，就是“潮流”，就是“新技术”。

如果要把“云”说得通俗易懂，“云”其实就是服务，更准确地说，就是“按需供给”的服务。举个最简单的例子：我们使用的自来水就是一个“按需供给”的服务。安装了自来水的人，只要有用水的需求，打开水龙头，水就会“自来”。我们只负责喝水，并不用关心水是从哪里来的、怎么来的，背后使用了什么技术过滤水、增压水，以及水网是怎么样的等。

云计算也一样。大部分使用者只要按需使用即可，并不需要关心“云”是什么，或者“云”的背后有什么。从使用者的角度可以认为：“云”=“自来”。云服务一般分为软件即服务（SaaS）、平台即服务（PaaS）和基础设施即服务（IaaS），如图1-1所示。

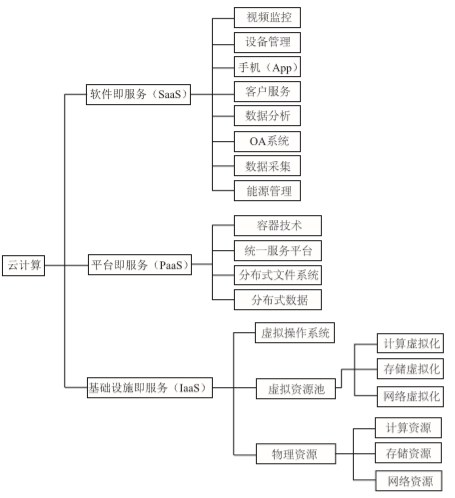


图1-1　云服务的三种常见形式

那么，“云”是在怎样的背景下产生和发展的呢？我们不妨从以下思路开始理解。

在传统模式下，企业要想建立一套完整的IT系统，不仅要购买硬件等基础设施，还要购买软件许可证，更离不开专职人员的维护；企业规模扩大时，为了满足需要，还要继续对各种软硬件设施进行升级。对于企业来说，计算机并不是他们真正需要的，只是一种完成工作的工具；对于个人来说，如果想正常使用计算机，就要安装许多软件，而有些软件是收费的，如果仅仅是偶尔使用，就非常不划算。那么，能否出现一种服务来为我们提供所需要的软件？如此，需要时，只要支付少量租金，就能拥有这些软件服务，从而减少资金投入。

如果用“自来水”来比喻“云”服务系统，就可以这样理解：虽然我们每天都要使用自来水，但并不是每家都要挖一口井，而是由自来水厂统一提供。同样，生活中到处都要使用到电，但并不是每家都要自备发电机，而是由电厂集中提供……这种模式不仅节约了资源，还方便了我们的生活。其实，对于计算机引发的生活问题，同样可以像水和电一样，只要将计算机资源充分利用起来，云计算自然也就出现了。

过去，为了完成某项任务，我们可能会将其交给一台计算机，任务小，就用小计算机；任务大，就用大计算机。而如今，我们完全可以将一项任务交给一堆计算机，由它们共同协作完成。这一堆计算机的工作状态犹如一台超级计算机，如果把一台计算机看作一滴水，那么一堆计算机就是一片云，云计算需要解决的问题是：如何让一堆计算机像一台大计算机一样工作；如何才能根据需要添加新计算机，提高大计算机的计算能力。

其实，云计算就是一种以互联网为媒介的超级计算模式，在远程的数据中心，众多计算机和服务器连接在一起，成为一片计算机云。用户完全可以通过计算机、手机等方式接入数据中心，租用这些数据，供自己所用。也就是说，计算能力也可以像商品一样进行流通，方便取用，价格不高。只不过，它是通过互联网进行传输的，具有一定的革命性。

互联网时代，这种革命是互联网应用取得突破的关键：同传统通信相比，在信息传输网上，互联网前向增加了信息采集（即传感网）、后向增加了无限的信息处理环节，有利于庞大数据的处理。未来，只要使用一台计算机或一部手机，就可以通过网络服务来实现需要的一切，甚至包括超级计算。

从这个角度来说，用户也就成了云计算的真正拥有者。

**二、云计算及其特征**

云计算是传统计算机和网络技术发展融合的产物，如分布式计算、并行计算、效用计算、网络存储、虚拟化、负载均衡、热备冗余等。早期的云计算，就是一种非常简单的分布式计算，解决了任务分发的问题，只要将计算结果合并在一起即可。

曾经，云计算还有一个别名——网格计算。早些年，有些大企业之所以会使用网格计算，可能只是为了解决效率与计算问题，后来它们发现外部也能使用云计算，便出现了公共云计算，也就是把计算机的计算能力直接放在网上进行售卖。

如今，对于云计算的定义都太复杂，其实只要使用一句话就足够：通过网络以自助服务的方式获得所需的IT资源的模式。云计算有三个关键点：获取路径——通过网络；获取方式——自助服务；获取对象——IT资源（包括但不限于计算能力、存储能力、带宽、域名系统等）。

云计算不仅仅是一种技术，而是多种传统IT技术的整合，更是一种管理运营模式。传统的IT工程师都可以往云计算方向转型。以普通人视角来看，云计算跟当初的互联网一样，都经历了“窄带—宽带—无线互联”的演变路径，云计算就是从传统的IT业务演变而来的。

云计算的出现，主要还是源于互联网的飞速发展，特别是社交网络及移动互联网的急速发展让每天产生的数据呈指数型增长，随着需求的不断增长，网络和网络服务也出现了持续增长，从而让网络数据和流量激增。

随着信息与数据的不断增加，在科学、工程和商业等领域都需要处理海量数据。采用传统基础架构，要想实现系统的可扩展性，不仅需要加大硬件投入，还要面对电量不足、空间不够、服务器容量有限等问题，为了解决这些问题，就要对传统基础架构进行扩展。

面对存储数据的急剧膨胀，为了节省更多的成本、实现闲置资源的再利用、联机处理海量信息、提升系统的可扩展性，云计算的概念自然也就被提了出来。概括起来，云计算主要有下面几个明显特征。

（1）呈大规模分布式。云计算规模非常庞大，Google云计算、Amazon、IBM、微软、阿里云等知名的云供应商，服务规模都多达数百万。“云”以分布式服务器为基础，能够为使用者提供绝无仅有的计算能力。只要身边存在网络，都能直接用手机、计算机等设备跟“云”进行链接，该计算能力的网络接入非常宽泛，通过网络，能够给用户提供很多服务。用户可以在任何时间和地点使用已有资源或购买所需的新服务。

（2）虚拟化可实现资源共享。云计算采用的是一种虚拟化技术，用户只要选择一家云服务提供商，注册一个账号，登录到云控制台，就能购买自己需要的服务，如云服务器、云存储、CDN（内容分发网络）等，经过简单的配置之后，可以通过自己的应用向外提供服务，方法简单，操作方便，还能随心所欲地用个人计算机或移动设备来控制自己的资源。

（3）具有弹性，可动态伸缩。利用云计算，用户就能随时随地访问需要的服务；同时，用户的系统规模变化时，还能根据用户的需求对系统做出相应的调整，如硬件配置、网络带宽、存储容量等。云计算不仅具有可动态扩展和配置特性，知名的云计算供应商还会用数据多副本容错、计算节点同构可互换等方法来保障服务的可靠性。

云服务的应用可以持续对外提供全天候服务，为了满足应用和用户规模增长的需要，“云”的规模还可以进行动态伸缩。

（4）按需付费，更加经济。在“云”的资源池中，用户不仅可以像公共资源一样来购买自己需要的应用和资源，还可以购买自己需要的其他服务，甚至按使用量来进行精确计费。如此，不仅能节省IT成本，还能显著改善资源的整体利用率。

（5）安全系数高。如今，网络安全是所有企业或个人创业者必须面对的问题，企业或个人都无法应对来自网络的恶意攻击，如果将“云”服务充分利用起来，就能借助专业的安全团队来有效降低风险，安全性更高。

**三、支撑云计算的五大关键技术**

在正式讨论云计算技术之前，先要了解一下摩尔定律。

摩尔定律由英特尔（Intel）创始人之一戈登·摩尔（Gordon Moore）提

出，该定律认为，只要价格不变，集成电路上可容纳的元器件数量每隔18～24个月就会增加一倍，性能则是原来的一倍。也就是说，只要花1美元购得计算机性能，每隔18～24个月，就会翻番。

过去的二十多年，PC（个人计算机）产业的发展都离不开技术的推动，云计算时代的到来，让一切回归本源。云计算时代，多数计算都在“云”中完成，用户终端得到了解放，对终端计算能力的要求相应降低，推动PC产业前进的摩尔定律逐渐失去了魔力。

云计算时代，人们渐渐发现，只要使用一个性能普通的终端，就能满足日常需求。在“云”时代，只要使用一个简单开放的Linux操作系统、缓慢的CPU、256MB内存、少于10GB的硬盘，就可以满足人们的日常需要了。多数计算都在“云”中完成，性能强大的PC已经失去了存在的意义，“云”时代终端设备的代名词将变为能够上网、开放和廉价。所以，手机上网也就逐渐代替了计算机上网。

云计算改写了摩尔定律的运行法则，那么云技术的核心技术又是什么呢？云计算系统运用了多种技术，但其中有几个核心技术。确切地说，云计算是大规模分布式计算技术及其配套商业模式演进的产物，它的发展主要有赖于虚拟化、分布式数据存储、数据管理、编程模式、信息安全等各项技术、产品的共同发展。

1．虚拟技术

在云技术中，虚拟技术是重要的核心技术之一，主要为云计算服务提供基础架构层面的支撑，驱动着ICT（信息通信和技术）服务快速走向云计算。使用虚拟化技术，软件应用与底层硬件就会隔离开来，既包括裂分模式（将单个资源划分成多个虚拟资源），也包括聚合模式（将多个资源整合成一个虚拟资源）。

根据对象的不同，可以将虚拟化技术分成存储虚拟化、计算虚拟化、网络虚拟化等，而计算虚拟化又分为系统级虚拟化、应用级虚拟化和桌面虚拟化。

虚拟化共有两种应用模式：一种是将一台性能强大的服务器虚拟成多个独立的小服务器，为不同的用户提供服务。第二种是将多个服务器虚拟成一个强大的服务器，实现特定功能。两种模式的核心都是统一管理、动态分配资源和提高资源利用率，在云计算中，两种模式的应用都比较多。

2．编程技术

从本质上来说，云计算是一个多用户、多任务、支持并发处理的系统。其核心理念是高效、简捷和快速，能够通过网络把强大的服务器计算资源方便地分发到终端用户手中，降低成本，为用户带来良好的体验。

目前，MapReduce是云计算主流并行编程模式之一，由Google开发的Java、Python、C++编程模型，是一种简化的分布式编程模型，还是一种高效的任务调度模型，可以被用于大规模数据集（大于1TB）的并行运算。

严格的编程模型能让云计算环境下的编程变得异常简单。所谓MapReduce模式，就是将打算解决的问题分解成Map（映射）和Reduce（化简）等方式，通过Map程序将数据切割成不相关的区块，分配（调度）给计算机进行处理，然后用Reduce程序将结果进行汇整和输出。

3．存储技术

云计算涉及的数据众多，使用分布式数据存储技术，就能将数据存储在不同的物理设备中，提高扩展性，即使是海量数据，也能快速、高效地处理，更好地满足用户需求的变化。

目前，比较重要的云计算存储技术主要有两个：一个是Google开发的GFS；另一个是Hadoop 团队开发的HDFS技术。当然，使用最广泛的还是后者。

（1）GFS是Google File System的简称，该存储技术开源扩展，分布式呈现，硬件价格比较低，提供了足够的容错功能，主要应用于分布式的数据访问。

（2）HDFS是Hadoop Distributed File System的简称，即分布式文件系统，由多个存储数据的终点和管理节点构成。中心服务器是namenode，在每个节点上，都对应一台普通计算机。

4．管理技术

云计算对海量数据进行处理和分析时，要使用两个数据管理技术：一个是Google开发的BT（Big Table，分布式数据存储系统）；一个是Hadoop团队开发的开源数据管理模块HBase。这里重点说一下BT。

BT是一个大型分布式数据库，以数据为研究对象，绘制成一个大表格，将大规模结构化数据存储下来。在Google上，很多项目都是用BT来存储数据的，如网页查询、Google Earth和Google金融。只不过，应用程序不同，要求也不同，服务方式灵活多样。

5．平台管理技术

云计算资源规模庞大，服务器数量众多，分布在不同的地点，数百种应用同时运行，必须有效地管理这些服务器，让整个系统提供不间断的服务。云计算系统的平台管理技术，能够让众多服务器协同工作，及时进行业务部署，快速发现和恢复系统故障，通过自动化、智能化等手段，保证大规模系统的可靠运营。