大数据与云计算技术漫谈

赵　凯　李玮瑶　著

光明日报出版社

大数据与云计算技术漫谈

赵　凯　李玮瑶　著

光明日报出版社

图书在版编目（CIP）数据

大数据与云计算技术漫谈 / 赵凯，李玮瑶著. -- 北京 : 光明日报出版社，2015.8
ISBN 978-7-5112-9250-6

Ⅰ. ①大… Ⅱ. ①赵… ②李… Ⅲ. ①计算机网络－应用 Ⅳ. ①TP393

中国版本图书馆 CIP 数据核字(2015)第 219452 号

大数据与云计算技术漫谈

著　　者：赵　凯　李玮瑶
责任编辑：靳鹤琼　　封面设计：崔新新
责任校对：傅泉泽　　责任印制：曹　诤
出版发行：光明日报出版社
地　　址：北京市东城区珠市口东大街 5 号，100062
电　　话：010-67078258（咨询），67078870（发行），67019571（邮购）
传　　真：010-67078227，67078255
网　　址：http：//book.gmw.cn
E-mail：jn7k@163.com
法律顾问：北京德恒律师事务所龚柳方律师
印　　刷：山东华盛印刷有限公司
装　　订：山东华盛印刷有限公司
本书如有破损、缺页、装订错误，请与本社联系调换
开　　本：787×1092
字　　数：100 千字　　印　　张：8
版　　次：2016 年 10 月第 1 版 印　　次：2016 年 10 月第 1 次印刷
书　　号：ISBN 978-7-5112-9250-6
定　　价：34.00 元

前 言

纵观历史，过去的数据中心无论应用层次还是规模大小，都仅仅是停留在过去有限的基础架构之上，采用的是传统精简指令集计算机和传统大型机，各个基础架构之间都相互孤立，没有形成一个统一的有机整体。在过去的数据中心里面，各种资源都没有得到有效充分地利用。而且传统数据中心资源配置和部署大多采用人工方式，没有相应的平台支持，使大量人力资源耗费在繁重的重复性工作上，缺少自助服务和自动部署能力，既耗费时间和成本，又严重影响工作效率。而当今越来越流行的云计算、虚拟化和云存储等新 IT 模式的出现，又再一次说明了过去那种孤立、缺乏有机整合的数据中心资源并没有得到有效利用，并不能满足当前多样、高效和海量的业务应用需求，于是，大数据技术应运而生。

大数据技术是云计算技术的延伸。大数据技术涵盖了从数据的海量存储、处理到应用多方面的技术，包括海量分布式文件系统、并行计算框架、NoSQL 数据库、实时流数据处理以及智能分析技术如模式识别、自然语言理解、应用知识库等等。大数据技术可以为我们带来新的机会。大数据在网络应用中可以涵盖多个方面，包括企业管理分析如战略分析、竞争分析，运营分析如用户分析、业务分析、流量经营分析，网络管理维护优化如网络信令监测、网络运行质量分析，营销分析如精准营销、个性化推荐等。

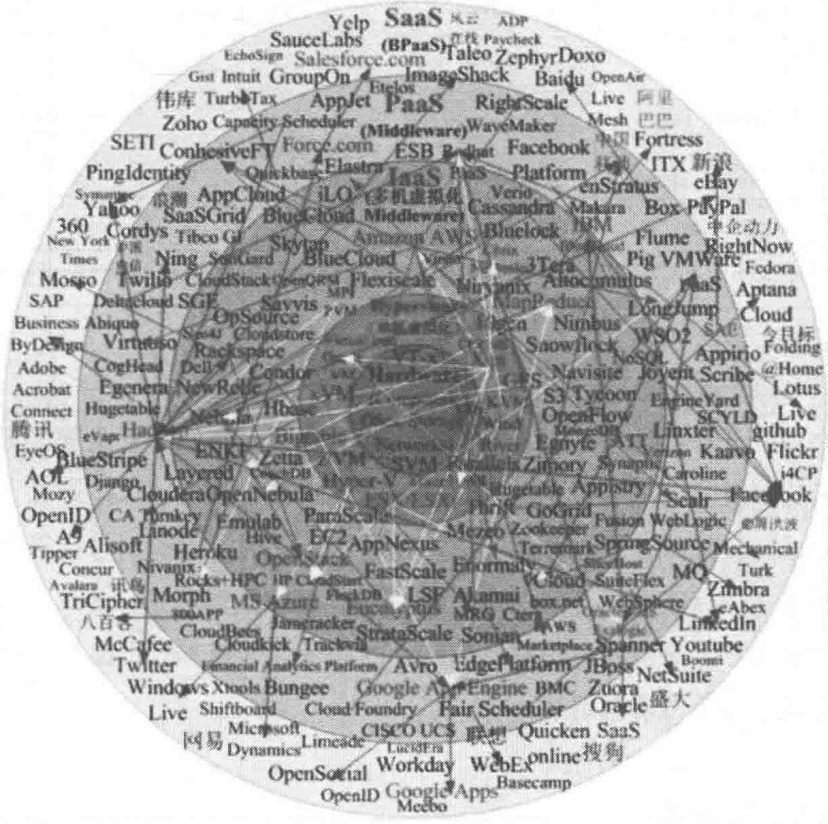
云计算和大数据是一个硬币的两面，大数据正在引发全球范围内深刻的技术和商业变革。如同云计算的出现，大数据也不是一个突然而至的新概念。云计算是大数据成长的驱动力，由于数据越来越多、越来越复杂、越来越实时，这就更加需要云计算去处理，所以二者之间是相辅相成的。本书就从云计算和大数据的概念和类别出发，基于网络和技术两个层面去剖析大数据和云计算技术及其发展，由于时间仓促，本书分析过程中不够深刻在所难免，敬请读者谅解。

目 录

1.云计算

1.1 云计算的概念

云计算的各方面定义很多，基于用户的视角来看，目的就是让使用者在不需了解资源的具体情况下做到按需分配，将计算资源虚拟化为一片云。站在高处看，当前的主流云计算更贴切于云服务，个人认为可理解为早先运营商提供数据中心服务器租用服务的延伸。以前用户租用的是一台台物理服务器，现在租用的是虚拟机，是软件平台甚至是应用程序。公认的三个云计算服务层次是IaaS（Infrastructure as a Service）、PaaS（Platform as a Service）和SaaS（Software as a Service），分别对应硬件资源、平台资源和应用资源。对于用户来说：



1、当提供商给你的是一套 a 个核 CPU、b G 大小内存的主机、c M 带宽网络以及 d G 大小存储空间，需要你自己去装系统和搞定应用程序，那么这就是 IaaS，举例如 Amazon EC2；

2、当提供的是包含基本数据库和中间件程序的一套完整系统，但你还需要根据接

口编写自己的应用程序时，那么就是 PaaS，举例如 Google AppEngine、Microsoft Azure 和 Amazon SimpleDB, SQS；

3、最简单的方式自然是连应用程序都写好了，例如你只需要告诉服务提供商想要的是个 500 人的薪酬管理系统，返回的服务就是个 HTTPS 的地址，设定好帐号密码就可以访问过去直接使用，这就是 SaaS 了，如 SalesForce、Yahoo Hadoop 和 Cisco Webex: Collaboration SaaS 等。

服务 属性	Amazon EC2	Google App Engine	Microsoft Azure	Yahoo Hadoop
架构	IaaS/PaaS	PaaS	PaaS	SaaS
服务形态	Compute/ Storage	Web application	Web and non-web	Software
管理技术	OS on Xen hypervisor	Application container	OS through Fabric controller	Map/Reduce Architecture
使用者界面	EC2 Command-line tools	Web-based Administration console	Windows Azure portal	Command line and web
APIs	Yes	Yes	Yes	Yes
收费	Yes	Yes	Yes	no
编程语言	AMI(Amazon Machine Image)	Python	.NET framework	Java

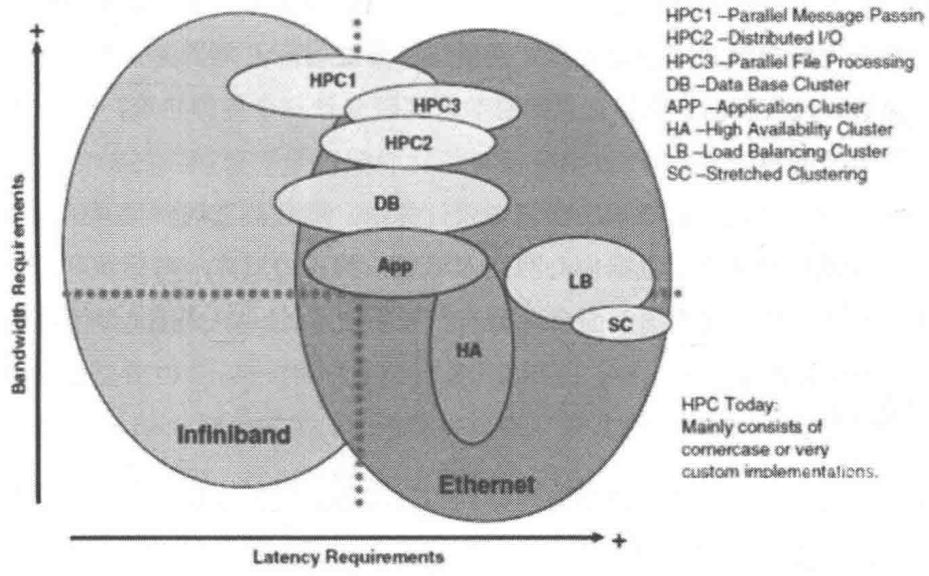
简单来说，云计算的核心首先是计算，什么网络、存储、安全等等都是外延，从技术上讲云计算就是计算虚拟化。最早的云计算来自于网格计算，通过一堆性能较差的服务器完成一台超级计算机才能完成的计算任务，简单的说就是计算多虚一。但是现如今一虚多（VM/XEN 等）也被一些厂商扯着大旗给忽悠进来，并且成为主流。但是单从技术角度来看，这两者是南辕北辙的。因此云计算技术在下面被作者主观的分为集中云与分散云两个概念来阐述。

1.2 云计算的分类

1.2.1 集中云

集中云，根正苗红的多虚一，最早期的也是目前最大的一个典型实际用户就是

Google 了(注意这里说的不是现在 Google 云服务)。搜索引擎是超级消耗资源的典型应用，从你在网页上一个关键词的搜索点击，到搜索结果的产生，后台是经过了几百上千台服务器的统一计算。随着互联网的发展，现在的开心、淘宝、新浪微博等等，虽然使用者看到的只是在简单的页面进行点击输入，但是后台的工作量已经远远不是少量几台大型服务器能够胜任的了，即使天河一号也不见得能搞定。集中云的应用主力就是这些大型的互联网内容提供商们，当然还有一些传统应用如地震、气象和科研项目的计算也会存在此类需求。



了解了需求，下面简单谈下技术，上图是Cluster集群多虚一技术的简单分布，除了按照承载网络类型可分成Infiniband和Ethernet外，根据技术分，还可分为Active-Standby主备与LoadBalance负载均衡两类。

主备模式好理解，所有的Server里面只有一台干活，其他都是候着的，只有侦听到干活的歇菜了，才开始接管处理任务。主备模式大部分就二虚一提供服务，多了如三虚一什么的其实意义都不太大，无非是为了再多增加些可靠性。主备模式以各类HA集群技术为代表。

而负载均衡模式复杂一些,在所有的 LB 技术中都存在两个角色,协调者与执行者,协调者一般是一个或多个（需要主备冗余时），主要工作就是接活儿和分活儿；而执行者就只处理计算了，分到什么就完成什么。从流量模型上来说，LB 集群技术有来回路径一致和三角传输两种，来回路径一致指流量都是客户发起连接，请求协调者进行处理，协调者分配任务给执行者进行计算，计算完成后结果会都返回到协调者，再由协调者应答客户。

这种结构简单，计算者不需要了解外界情况，由协调者统一作为内外接口，安全性最高。此模型主要应用于搜索和地震气象科研计算等业务处理中。三角传输模型指计算者完成计算后直接将结果反馈给客户，此时由于计算者会和客户直接通信，造成安全性降低，但返回流量减少了协调者这个处理节点，性能得到很大提升。此模型主要应用于腾讯新浪的新闻页面和阿里淘宝的电子商务等 WEB 访问业务。

集中云在云服务中属于富人俱乐部的范围，不是给中小企业和个人使用的，实际上都是各大互联网服务提供商自行搭建集中云以提供自己的业务给用户，不会说哪天雅虎去租用个Google的云来向用户提供自己的新闻页面访问。集中云服务可能的租用对象是那些高度科研项目，因而也导致当前集中云建设上升到国家宏观战略层面的地位。你能想象哪天百度的云服务提供给总装研究院去计算个导弹轨迹，核裂变什么的，这是完全不可能的事。

最后是多虚一对网络的需求。在集中云计算中，服务器之间的交互流量多了，而外部访问的流量相对减少，数据中心网络内部通信的压力增大，对带宽和延迟有了更高的要求，自然而然就催生出后面会讲到的一些新技术（L2MP/TRILL/SPB等）。

1.2.2 分散云

分散云，是目前的主流，也是前面提到的云服务的关键底层技术。由于有VMware和Citrix等厂家在大力推广，而且应用内容较集中云更加平民化，随便找台PC或服务器，装几个虚拟机大家都能试一试，也就使其的认知度更加广泛。

一虚多的最主要目的是为了提高效率，力争让所有的CPU都跑到100%，力争让所有的内存和带宽都占满。以前10台Server干的事，整两台Server每台跑5个虚拟机VM（Virtual Machine）就搞定了，省电省空间省制冷省网线，总之省钱是第一位的（用高级词儿就是绿色环保）。技术方面从实现方案来看，目前大致可分为三类：

操作系统虚拟化OS-Level

在操作系统中模拟出一个个跑应用程序的容器，所有虚拟机共享内核空间，性能最好，耗费资源最少，一个 CPU 号称可最多模拟 500 个 VPS(Virtual Private Server)或 VE(Virtual Environment)。缺点是操作系统唯一，如底层操作系统跑的 Windows，VPS/VE 就都得跑 Windows。代表是 Parallels 公司（以前叫 SWsoft）的 Virtuozzo（商用产品）和 OpenVZ（开源项目）。Cisco 的 Nexus 7000 猜测也是采用这种方案运行的 VDC 技术，但不太清楚为什么会有最多 4 个 VDC 的数量限制，也许是基于当前应用场景进行规格控制的一种商业手段。

主机虚拟化Hosted

先说下 Hypervisor 或叫做 Virtual Machine Monitor（VMM），它是管理虚拟机 VM 的软件平台。在主机虚拟化中，Hypervisor 就是跑在基础操作系统上的应用软件，与 OS-Level 中 VE 的主要区别在于：

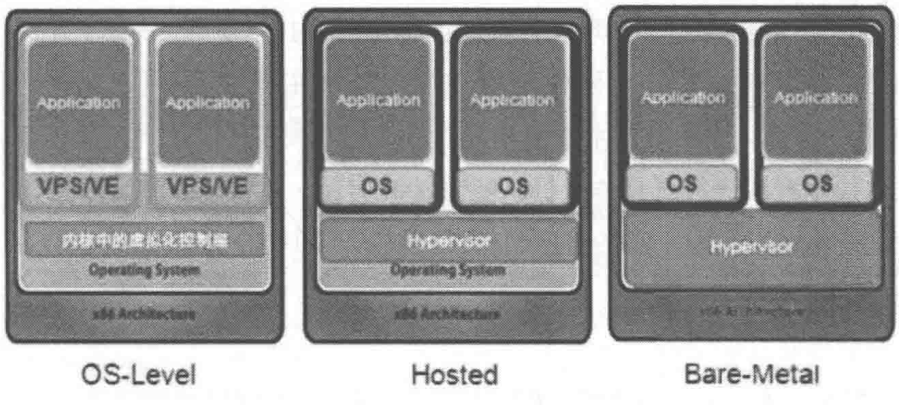
Hypervisor构建出一整套虚拟硬件平台（CPU/Memory/Storage/Adapter），上面需要你再去安装新的操作系统和需要的应用软件，这样底层和上层的OS就可以完全无关化，诸如Windows上跑Linux一点儿问题没有；

VE则可以理解为盗用了底层基础操作系统的资源去欺骗装在VE上的应用程序，每新创建出一个VE，其操作系统都是已经安装好了的，和底层操作系统完全一样，所以VE比较VM（包括主机虚拟化和后面的裸金属虚拟化）运行在更高的层次上，相对消耗资源也少很多。

主机虚拟化中VM的应用程序调用硬件资源时需要经过:VM内核->Hypervisor->主机内核,导致性能是三种虚拟化技术中最差的。主机虚拟化技术代表是VMware Server(GSX)、Workstation和Microsoft Virtual PC、Virtual Server等。

裸金属虚拟化Bare-metal

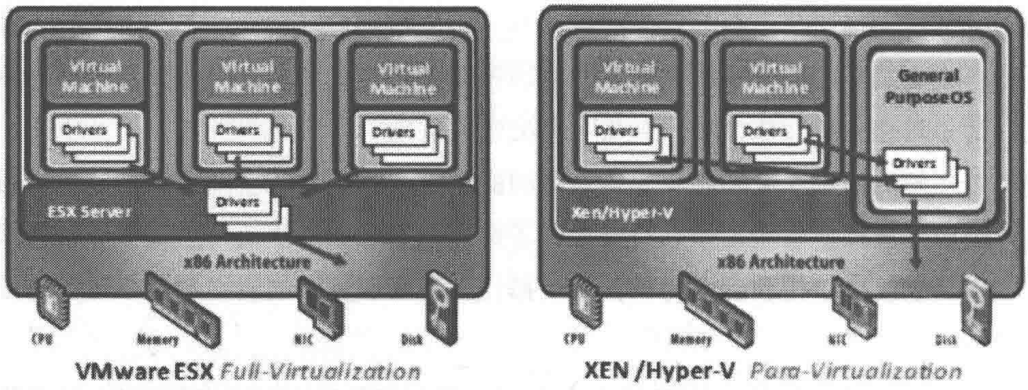
裸金属虚拟化中 Hypervisor 直接管理调用硬件资源，不需要底层操作系统，也可以理解为 Hypervisor 被做成了一个很薄的操作系统。这种方案的性能处于主机虚拟化与操作系统虚拟化之间。代表是 VMware ESX Server、Citrix XenServer 和 Microsoft Hyper-V。



上图描述了三种虚拟化方案的形态区别。当前分散云数据中心服务器虚拟化使用的主要是Bare-Metal方案。分散云给数据中心网络带来了新的挑战，虚拟机之间的数据通信管理需求促使了一系列网络新技术的发展。在OS-Level与Hosted方案中，虚拟机都是架设于操作系统之上的，因此VM/VE之间的通信主要由同样运行于基础操作系统之上的网络交换应用程序来完成。而在最主流的Bare-Metal结构中，由于Hypervisor薄操作系统的引入，性能、管理、安全和可靠性等多维度的考虑，造成VM间网络通信

管理发展出不同的技术道路（EVB与BPE），后文会对这些技术方向加以详述。

VMware ESX 与 Xen/Hyper-V 的 Bare-Metal 方案实现结构有所不同，简单如下图所示。



分散云除了给网络带来上述的VM通信问题，同样由于其对服务器硬件能力的极端榨取，造成网络中的流量压力增大，与集中云一样存在着带宽扩展的需求。原本一台服务器一个操作系统跑一个应用只需要10M流量带宽就够了，现在装了10个VM跑10个应用，带宽可能就需要100M了。

大型机与小型机的一虚多技术早在 30 年前 IBM 就做出来了，现在 RISC 平台上已经相当完善了，相比较而言 X86 架构的虚拟化才处于起步阶段，但 X86 架构由于性价比更高成为了分散云计算的首选。

X86架构最早期是纯软件层面的Hypervisor提供虚拟化服务，缺陷很多，性能也不够，直到2006年Intel推出了实现硬件辅助虚拟化的VT技术CPU产品后才开始迅猛发展（AMD也跟着出了VM技术）。硬件辅助虚拟化技术主要包括CPU/Chipset/Network Adapter等几个方面，和网络技术紧密相关的就是网卡虚拟化了，后文会对如SR-IOV等网卡虚拟化技术应用进行更具体分析。随着2007年Intel VT FlexMigration技术的推出，虚拟机迁移成为可能，2009年Intel支持异构CPU间动态迁移再次向前迈进。

vMotion

这里再多唠叨几句 vMotion 技术。vMotion 是 VMware 公司提出的虚拟机动态迁移技术名称（XEN 也有相应的 XENMotion 技术），由于此名称被喊得最早，范围最广，认知度最高，因此下文提到虚拟机迁移技术时大都会使用 vMotion 来代称。

先要明确 vMotion 是一项资源管理技术，不是高可靠性技术，如果你的某台服务器或 VM 突然宕机了，vMotion 是无助于应用访问进行故障切换和快速恢复的。vMotion 是将一个正常的处于服务提供中的 VM 从一台物理服务器搬家到另一台物理服务器的技术，vMotion 的目的是尽可能方便的为服务管理人员提供资源调度转移手段，当物

理服务器需要更换配件关机重启啦，当数据中心需要扩容重新安排资源啦，这种时候vMotion就会有用武之地了。

设想一下没有vMotion上述迁移工作是怎么完成的，首先需要将原始物理服务器上的VM关机，再将VM文件拷贝到新的物理服务器上，最后将VM启动，整个过程VM对外提供的服务中断会达到几分钟甚至几小时的级别。而且需要来回操作两台物理服务器上的VM，对管理人员来说也很忙叨。

使用 vMotion 后，两台物理服务器使用共享存储来保存 VM 文件，这样就节省了上述步骤 2 中的时间，vMotion 只需在两台物理服务器间传递当前的服务状态信息，包括内存和 TCP 等上层连接表项，状态同步的拷贝时间相对较短，而且同步时原始 VM 还可以提供服务使其不会中断。同步时间跟 VM 当前负载情况及迁移网络带宽有关，负载大了或带宽较低使同步时间较长时，有可能会导致 vMotion 出现概率性失败。当状态同步完成后，原始物理服务器上的 VM 会关闭，而新服务器上的 VM 激活（系统已经在状态同步前启动完毕，始终处于等待状态），此时会有个较短的业务中断时间，可以达到秒级。再者 vMotion 是通过 VMware 的 vCenter 管理平台一键化完成的，管理人员处理起来轻松了许多。

这里要注意vMotion也一定会出现业务中断，只是时间长短区别，不要轻易被一些宣传所忽悠。想想原理，不论怎么同步状态，只要始终有新建发生，在同步过程中原始服务器上新建立的客户连接，新服务器上都是没有的，切换后这部分连接势必被断开重建，零丢包只能是理想值。VMware也同样建议将vMotion动作安排在业务量最少的时候进行。

vMotion什么场景适用呢？首先肯定得是一虚多的VM应用场景，然后是对外业务中断恢复的可靠性要求极高，一般都是7*24小时不间断应用服务才用得上，最后是计算节点规模始终在不断增长，资源调度频繁，管理维护工作量大的数据中心。

另外共享存储这个强制要求会给数据中心带来了整体部署上的限制，尤其是下面提到的跨数据中心站点 vMotion 时，跨站点共享存储的问题解决起来是很麻烦的，由于这部分内容和网络关系不大，属于存储厂商的地盘，对跨站点共享存储技术有兴趣的读者可以参考 EMC/IBM 等厂商的资料看看，本文就不过多介绍了。

vMotion 的出现推动了数据中心站点间大二层互联和多站点动态选路的网络需求，从而导致 OTV 和 LISP 等一系列新网络技术的出现。

通过前面的描述，希望大家能对云计算有个较为清晰的概念。云计算还有一大块内容是平台管理资源调度方面（目前很多厂家吆喝的云计算都是云平台）。这部分主要针对客户如何更便捷的创建与获取虚拟化服务资源，实际过程就是用户向平台管理

软件提出服务请求，管理平台通过应用程序接口API（Application Program Interface）将请求转化为指令配置下发给服务器、网络、存储和操作系统、数据库等，自动生成服务资源。需要网络做的就是设备能够识别管理平台下发的配置，从技术创新的角度讲，没有啥新鲜东西，就不多说了。当前的云平台多以IaaS/PaaS为主，能做到提供SaaS的极少。但在今后看来，SaaS将会成为云服务租用主流，中小企业和个人可以节省出来IT建设和维护的费用，更专注于自身的业务发展。

总结一下云计算给数据中心网络带来的主要变化：

1、更高的带宽和更低的延迟

2、服务器节点（VM）规模的增加

3、VM间通信管理

4、跨数据中心站点间的二层互联以承载 vMotion

题外再多说两句，计算虚拟化中一虚多与多虚一结合使用才是王道。但目前云计算服务提供商能够提供的只是先将物理服务器一虚多成多台 VM，再通过 LB/集群计算等技术将这些 VM 对外多虚一成一个可用的资源提供服务。个人感觉，如果能做到先将一堆物理服务器虚拟成一台几万个核 Super Computer，用户再根据自己的需要几个几十个核的自取资源，这样才更有云计算的样子，Super Computer 就是那朵云。当然计算虚拟化的时候不光是核的调配，还要包括 IO/Memory 等一起进行调度，这里只是简单举例。

1.3 云存储与云安全

1.3.1 云存储

云存储是一种网络在线储存(online storage)的模式，即把资料存放在通常由第三方代管的多台虚拟服务器，而非专属的服务器上。代管(hosting)公司营运大型的数据中心，需要数据储存代管的人，则透过向其购买或租赁储存空间的方式，来满足数据储存的需求。数据中心营运商根据客户的需求，在后端准备储存虚拟化的资源，并将其以储存资源池(storage pool) 的方式提供，客户便可自行使用此储存资源池来存放数据或文件。实际上，这些资源可能被分布在众多的伺服主机上。云存储这项服务透过 Web 服务应用程式接口(API)，或是透过 Web 化的使用者接口来存取。

云存储的优点：

- 用户只需要为实际使用的存储容量付费

• 用户不需要在在他们自己的数据中心或者办公环境中安装物理存储设备，这减少了 IT 和托管成本

• 存储维护工作（例如备份、数据复制和采购额外存储）转移至服务提供商，让企业机构把精力集中在他们的核心业务上

云存储的潜在问题：

• 当在云存储提供商那里保存敏感数据时，数据安全就成为一个潜在隐患

• 性能也许低于本地存储

• 可靠性和可用性取决于 WAN 的可用性以及服务提供商所采取的预防措施等级

• 具有特定记录保留需求的用户，例如必须保留电子记录的公共机构，可能会在采用云计算和云存储的过程中遇到一些复杂问题

1.3.2 云安全

“云安全”(Cloud Security)是网络时代信息安全的最新体现，它融合了并行处理、网格计算、未知病毒行为判断等新兴技术和概念，通过网状的大量客户端对网络中软件行为的异常监测，获取互联网中木马、恶意程序的最新信息，传送到 Server 端进行自动分析和处理，再把病毒和木马的解决方案分发到每一个客户端。

云计算中的安全控制其主要部分与其它 IT 环境中的安全控制并没有什么不同，然而，基于采用的云服务模型、运行模式以及提供云服务的技术，与传统 IT 解决方案相比云计算可能面临不同的风险。

即使有些运行责任落在某些第三方伙伴身上，云计算的一个独特点就是能够在适度地失去控制的同时又能保持可纠责性(accountability)。

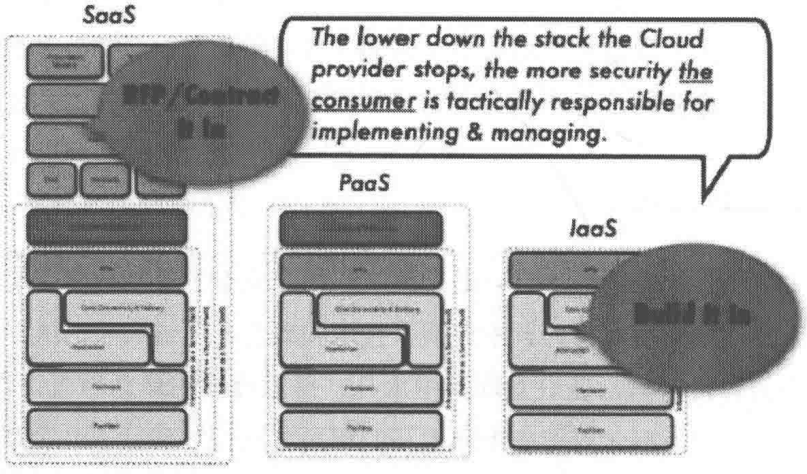
一个机构的安全态势的特征取决于成熟度、有效性以及实现基于风险调节的安全控制的完全程度，这些安全控制可以在一层或多层上实现，包括设备(物理安全)、网络基础设施(网络安全)、IT 系统(系统安全)，一直到信息和应用(应用安全)，更多的控制还包括人员和过程层面的，职责分离和变更的管理等。

在不同云服务模型中，提供商和用户的安全职责有很大的不同。例如，Amazon 的 AWS EC2 架构作为服务包括了一直到 hypervisor 安全的供应商的安全责任，也就是说它们只能解决物理安全、环境安全和虚拟化安全这些安全控制，而用户则负责与 IT 系统(事件)相关的安全控制，包括操作系统、应用和数据。

Salesforce.com 的客户资源管理 CRM SaaS 提供的正好相反，因为整个“栈”都由 Salesforce.com 提供，提供商不仅负责物理和环境安全还必须解决基础设施、应用和数据相关的安全控制，这减轻了用户的许多运行责任。

云计算的吸引力之一在于由经济上的可扩展性、重用和标准化提供的成本效率，为了支撑这种成本效率，云提供商提供的服务必须足够灵活，以服务最大可能的用户数、最大化他们的市场，不幸的是，将安全集成到这些服务方案中常被认为使得方案变得僵化。

这种僵化常与传统 IT 相比，表现在云环境不能部署同等的安全控制，这主要是由于基础设施的抽象化、缺少可视化、缺少集成多种熟悉的安全控制手段的能力，特别是在网络层上。



安全是如何集成的

上图表明了这些问题：在 SaaS 环境中，安全控制及其范围在服务合同中进行协商；服务等级、隐私和符合性等也都在合同中关系到。在 IaaS 中，低层基础设施和抽象层的安全保护属于提供商职责，其它职责则属于客户。PaaS 则居于两者之间，提供商为平台自身提供安全保护，平台上应用的安全性及如何安全地开发这些应用则为客户的职责。

中国企业的“云安全”概念

中国企业的“云安全”，在国际云计算领域独树一帜。中国企业云安全通过网状的大量客户端对网络中软件行为的异常监测，获取互联网中木马、恶意程序的最新信息，推送到服务端进行自动分析和处理，再把病毒和木马的解决方案分发到每一个客户端。整个互联网，变成了一个超级大的杀毒软件，这就是云安全计划的宏伟目标。

发展趋势

未来杀毒软件将无法有效地处理日益增多的恶意程序。来自互联网的主要威胁正在由电脑病毒转向恶意程序及木马，在这样的情况下，采用的特征库判别法显然已经过时。云安全技术应用后，识别和查杀病毒不再仅仅依靠本地硬盘中的病毒库，而是

依靠庞大的网络服务，实时进行采集、分析以及处理。整个互联网就是一个巨大的“杀毒软件”，参与者越多，每个参与者就越安全，整个互联网就会更安全。

云安全的概念提出后，曾引起了广泛的争议，许多人认为它是伪命题。但事实胜于雄辩，云安全的发展像一阵风[1]，瑞星、趋势、卡巴斯基、MCAFEE、SYMANTEC、江民科技、PANDA、金山、360 安全卫士等都推出了云安全解决方案。瑞星基于云安全策略开发的 2009 新品，每天拦截数百万次木马攻击，其中 1 月 8 日更是达到了 765 万余次。趋势科技云安全已经在全球建立了 5 大数据中心，几万部在线服务器。据悉，云安全可以支持平均每天 55 亿条点击查询，每天收集分析 2.5 亿个样本，资料库第一次命中率就可以达到 99%。借助云安全，趋势科技现在每天阻断的病毒感染最高达 1000 万次。

思想来源

云安全技术是 P2P 技术、网格技术、云计算技术等分布式计算技术混合发展、自然演化的结果。

云安全的过程值得一提的是，云安全的核心思想，与刘鹏早在 2003 年就提出的反垃圾邮件网格非常接近。刘鹏当时认为，垃圾邮件泛滥而无法用技术手段很好地自动过滤，是因为所依赖的人工智能方法不是成熟技术。垃圾邮件的最大的特征是：它会将相同的内容发送给数以百万计的接收者。

为此，可以建立一个分布式统计和学习平台，以大规模用户的协同计算来过滤垃圾邮件：

首先，用户安装客户端，为收到的每一封邮件计算出一个唯一的“指纹”，通过比对“指纹”可以统计相似邮件的副本数，当副本数达到一定数量，就可以判定邮件是垃圾邮件；

其次，由于互联网上多台计算机比一台计算机掌握的信息更多，因而可以采用分布式贝叶斯学习算法，在成百上千的客户端机器上实现协同学习过程，收集、分析并共享最新的信息。

反垃圾邮件网格体现了真正的网格思想，每个加入系统的用户既是服务的对象，也是完成分布式统计功能的一个信息节点，随着系统规模的不断扩大，系统过滤垃圾邮件的准确性也会随之提高。用大规模统计方法来过滤垃圾邮件的做法比用人工智能的方法更成熟，不容易出现误判假阳性的情况，实用性很强。反垃圾邮件网格就是利用分布互联网里的千百万台主机的协同工作，来构建一道拦截垃圾邮件的“天网”。

反垃圾邮件网格思想提出后，被 IEEE Cluster 2003 国际会议选为杰出网格项目在香港作了现场演示，在 2004 年网格计算国际研讨会上作了专题报告和现场演示，引