

# 2019 杭电云计算课程期末重点个人总结

author: hjs 16041321

题型分布

判断 10 分

单选 30 分

简答 25 分

讨论 20 分

分析 15 分

## 1、云计算服务类型（了解）

- (1) 将基础设施作为服务 IaaS: 将硬件设备等基础资源封装成服务供用户使用
- (2) 将平台作为服务 PaaS: 对资源的抽象层次更进一步, 提供用户应用程序运行环境
- (3) 将软件作为服务 SaaS: 针对性更强, 它将某些特定应用软件功能封装成服务

## 2、云计算特征

超大规模、虚拟化、高可靠性、通用性、高可伸缩性、按需服务、极其廉价

1)弹性服务。服务的规模可快速伸缩, 以自动适应业务负载的动态变化。用户使用的同业务的需求相一致, 避免了因为服务器性能过载或冗余而导致的服务质量下降或资源浪费。

2)资源池化。资源以共享资源池的方式统一管理。利用虚拟化技术, 将资源分享给不同用户, 资源的放置、管理与分配策略对用户透明。

3)按需服务。以服务的形式为用户提供应用程序、数据存储、基础设施等资源, 并可以根据用户需求, 自动分配资源, 而不需要系统管理员干预。

4)服务可计费。监控用户的资源使用量, 并根据资源的使用情况对服务计费。

5)泛在接入。用户可以利用各种终端设备(如 PC 电脑、笔记本电脑、智能手机等)随时随地通过互联网访问云计算服务。

## 3、谷歌三驾马车 GFS、MapReduce、BigTable 三大件

答: (1) GFS 是一个基于分布式集群的大型分布式文件系统, 为 MapReduce 计算框架提供底层数据存储和数据可靠性支撑。

系统架构及组成: (1)Client 客户端: GFS 提供给应用程序的访问接口; (2)Master 主服务器: GFS 的管理节点, 负责整个文件系统的管理; (3)Chunk Server 数据块服务器: 负责具体的存储工作。数据以文件的形式存储在 Chunk Server 上;

实现机制: (1)客户端首先访问 Master 节点, 获取交互的 Chunk Server 信息, 然后访问这些 Chunk Server, 完成数据存取工作。这种设计方法实现了控制流和数据流的分离; (2)Client 与 Master 之间只有控制流, 而无数据流, 极大地降低了 Master 的负载; (3)Client 与 Chunk Server 之间直接传输数据流, 同时由于文件被分成多个 Chunk 进行分布式存储, Client 可以同时访问多个 Chunk Server, 从而使得整个系统的 I/O 高度并行, 系统整体性能得到提高。

特点: (1)采用中心服务器模式: 方便增加 Chunk Server; Master 掌握所有 Chunk Server 情况, 方便负载均衡; 不存在元数据一致性问题; (2)不缓存数据: 因为文件操作为流式读写; (3)用户态下实现

容错机制: (1)GFS 采用副本的方式实现 Chunk Server 的容错; (2)每一个 Chunk 有多个存储副本(默认为三个) (3) 对于每一个 Chunk, 必须将所有的副本全部写入成功, 才视为成功写入; (4)相关的副本出现丢失或不可恢复等情况, Master 自动将该副本复制到其他 Chunk Server; (5) GFS 中的每一个文件被划分成多个 Chunk, Chunk 的默认大小是 64MB; (4)每一个 Chunk 以 Block 为单位进行划分, 大小为 64KB, 每一个 Block 对应一个 32bit 的校验和。

(2)MapReduce 把对数据集的大规模操作, 分发给一个主节点管理下的各分节点共同完成, 通过这种方式实现任务的可靠执行与容错机制。

编程模型: **Map 函数**——对一部分原始数据进行指定的操作。每个 **Map** 操作都针对不同的原始数据, 因此 **Map** 与 **Map** 之间是互相独立的, 这使得它们可以充分并行化。**Reduce** 操作——对每个 **Map** 所产生的一部分中间结果进行合并操作, 每个 **Reduce** 所处理的 **Map** 中间结果是互不交叉的, 所有 **Reduce** 产生的最终结果经过简单连接就形成了完整的结果集。

**Map:**  $(in\_key, in\_value) \rightarrow \{(key_j, value_j) \mid j = 1 \cdots k\}$

**Reduce:**  $(key, [value_1, \cdots, value_m]) \rightarrow (key, final\_value)$

实现机制:

- (1) **MapReduce** 函数首先把输入文件分成 **M** 块
- (2) 分派的执行程序中有有一个主控程序 **Master**
- (3) 一个被分配了 **Map** 任务的 **Worker** 读取并处理相关的输入块
- (4) 这些缓冲到内存的中间结果将被定时写到本地硬盘, 这些数据通过分区函数分成 **R** 个区
- (5) 当 **Master** 通知执行 **Reduce** 的 **Worker** 关于中间  $\langle key, value \rangle$  对的位置时, 它调用远程过程, 从 **Map Worker** 的本地硬盘上读取缓冲的中间数据
- (6) **Reduce Worker** 根据每一个唯一中间 **key** 来遍历所有的排序后的中间数据, 并且把 **key** 和相关的中间结果值集合传递给用户定义的 **Reduce** 函数
- (7) 当所有的 **Map** 任务和 **Reduce** 任务都完成的时候, **Master** 激活用户程序

容错机制:

**MapReduce** 通过重新执行失效的地方来实现容错。**Master** 失效, **Worker** 失效

(3) **Bigtable** 是一个分布式多维映射表, 表中的数据通过一个行关键字 (**Row Key**)、一个列关键字 (**Column Key**) 以及一个时间戳 (**Time Stamp**) 进行索引。**Bigtable** 的存储逻辑可以表示为:  $(row:string, column:string, time:int64) \rightarrow string$

基本架构: **Bigtable** 主要由三个部分组成: 客户端程序库 (**Client Library**)、一个主服务器 (**Master Server**) 和多个子表服务器 (**Tablet Server**)。客户访问 **Bigtable** 服务时, 首先要利用其函数执行 **Open()** 操作来打开一个锁 (实际上就是获取了文件目录), 锁打开以后客户端就可以和子表服务器进行通信了。客户端主要与子表服务器通信, 主服务主要进行一些元数据的操作以及子表服务器之间的负载调度问题, 实际的数据是存储在子表服务器上的。

#### 4、云计算虚拟化

云计算的支撑技术: 分布式技术、虚拟化技术、云平台技术

分布式技术:

##### (1) 分布式系统

分片思想, 低成本、高性能、分布式、多用户、协同工作、高可靠性、高扩展性

**Hadoop** (**HDFS** (提供分布式存储) + **MapReduce** (提供分布式计算))

##### (2) 分布式计算

**MapReduce**

主从架构: 客户端、作业跟踪器、任务跟踪器、任务 (映射 (**Map**) 任务、归约 (**Reduce**) 任务)

##### (3) 分布式存储

分布式文件系统、分布式数据库系统

**HDFS** (分布式文件系统)

架构: 名字节点, 数据节点, 第二名字节点

名字节点: 管理数据块映射; 处理客户端的读写请求; 配置副本策略; 管理 **HDFS** 的名称空间; 数据节点: 负责存储客户端发来的数据块; 执行数据块的读写操作; 第二名字节点: 保存着名字节点的部分信息, 是名字节点的冷备份。

**Hbase** (分布式数据库)

虚拟化技术: 核心思想是利用软件或固件管理程序构成虚拟化层, 把物理资源映射为虚

拟资源。在虚拟资源上可以安装和部署多个虚拟机，实现多用户共享物理资源。

实现技术：

桌面虚拟化：

存储虚拟化：把分布的异构存储设备统一为一个或几个大的存储池。实现方式：基于**主机**的存储虚拟化：通过逻辑卷管理来实现的；基于**存储设备**的存储虚拟化：在存储设备的磁盘、适配器或者控制器上实现虚拟化功能；基于**网络**的存储虚拟化：是在网络设备上实现存储虚拟化功能，包括基于互连设备和基于路由器两种方式。。

网络虚拟化：在底层物理网络和网络用户之间增加一个抽象层。层次：核心层：数据中心核心网络设备的虚拟化、接入层：实现数据中心接入层的分级设计；虚拟机层：

服务器虚拟化：将一个或多个物理服务器虚拟成多个逻辑上的服务器。层次：寄居虚拟化(VMM)、裸机虚拟化(Hypervisor)。底层实现：CPU 虚拟化：虚拟 CPU 的正确运行是要保证虚拟机指令正确运行，现有的实现技术包括模拟执行和监控执行，调度问题是指 VMM 决定当前哪个虚拟 CPU 在物理 CPU 上运行，要保证隔离性、公平性和性能；内存虚拟化：内存虚拟化技术把物理内存统一管理，包装成多个虚拟的物理内存提供给若干虚拟机使用，每个虚拟机拥有各自独立的内存空间；I/O 设备虚拟化：I/O 设备虚拟化技术把真实的设备统一管理起来，包装成多个虚拟设备给若干个虚拟机使用，响应每个虚拟机的设备访问请求和 I/O 请求。

)、

云平台技术

(1) 服务计算技术

SOA 架构

(2) 多租户技术

(3) 容器技术

Docker（镜像、容器、仓库）

## 5、云计算、大数据、人工智能三者之间的发展关系

深度学习为火箭、燃料为大数据、云计算为引擎，三位一体交互发展；云计算为大数据和人工智能提供计算技术支撑；云计算是大数据的基石，保证分布式并行数据挖掘和高效实时挖掘；大数据中的分布式计算和存储本身即是云计算的重要组成部分；大数据的数据分析结果同时反作用于云计算，例如云安全中的态势感知等；大数据处理中的新需求促使云计算的软硬件不断发展，例如 GPU 主机，SSD 硬盘等；深度学习是在云计算和大数据日趋成熟的背景下取得了实质性进展、云计算为深度学习提供了平台，大数据为深度学习提供了基石，云计算与大数据技术的快速迭代，加快了人工智能应用的落地

## 6、云计算的体系结构

云计算的体系结构分为 4 层：SOA 构建层、管理中间件层、管理中间件层和 SOA 构建层。

SOA 构件层：封装云计算能力成标准的 Web Services 服务并纳入到 SOA 体系。

管理中间件层：具有资源管理、任务管理、用户管理和安全管理功能，能够对众多应用任务进行调度，使资源能够高效、安全地为应用提供服务。

资源池层：将大量相同的资源构成同构或接近同构的资源池。计算资源池、存储资源池、网络资源池、数据资源池、软件资源池。

物理资源层：计算机、存储器、网络设施、数据库和软件等

管理中间件层和资源池层是云计算技术的最关键部分，SOA 构建层的功能更多依靠外部设施提供。

## 7、云计算与移动计算的区别

移动计算使用与云计算同样的概念。借助互联网而不是借助单个设备，云计算因数据而

变得活跃。它为用户提供了需要按需获取的数据。在移动计算中，应用程序在远程服务器上运行，为用户提供了访问所存储数据的权限。

#### 8、云数据中心（绿色节能技术有哪些）

（1）配电系统节能技术：高压直流配电技术、市直直供配电技术

（2）空调系统节能技术：高温回风空调系统、低能耗加湿系统、自然冷空调系统

（3）集装箱数据中心节能技术：将数据中心的服务器设备、网络设备、空调设备、供电设备等高密度地装入固定尺寸的集装箱中，使其成为数据中心的标准构建模块，进而通过若干集装箱模块网络和电力的互连互通构建完整的数据中心。

（4）节能策略算法：DVFS 节能技术；基于虚拟化的节能技术；基于主机关闭/开启的节能技术；随机式策略、超时式策略和预测式策略。

$PUE = \text{数据中心总能耗} / \text{IT 设备能耗}$ ，基准是 2，比值越接近 1，表示数据中心的能源利用率越高

#### 9、分析题：给你一个信息系统框图，懂得描述哪些地方用到云计算相关技术，并且如何体现

分布式云存储：淘宝分布式文件系统 TFS

分布式数据库：阿里巴巴 OceanBase

阿里巴巴云服务：云引擎 ACE、大规模分布式计算系统“飞天”

弹性计算服务 ECS(安全可靠、硬件故障自动恢复、防网络攻击、简化部署、降低运维成本)；

开放存储服务 OSS(弹性扩展、大规模并发读写、图片处理优化)；

开放结构化数据服务 OTS(海量数据存储、简单易用表管理、数据管理)；

开放数据处理服务 ODPS(用于数据挖掘、海量运算、数据安全、开箱即用)；

关系型数据库 RDS(安全稳定,数据可靠；自动备份，管理透明；性能卓越，灵活扩容)