

智能网联系统导论



# 物联网数据处理技术

裴欣，副研究员

中央主楼809房间

[peixin@tsinghua.edu.cn](mailto:peixin@tsinghua.edu.cn)



2020/12/15

清华大学自动化系系统工程研究所

# 物联网数据的特点

2



海量

多态



动态

关联



# 物联网数据的特点

3

- ✓ 海量感知设备和节点：  
2020年物联网中对象的数量约为320亿部
- ✓ 传感器节点多数处于全时工作状态

海量

多态

动态

关联



# 物联网数据的特点

4

- ✓ 海量感知设备和节点：  
2020年物联网中对象的数量约为320亿部
- ✓ 传感器节点多数处于全时工作状态

海量

多态

动态

关联

- ✓ 多源异构数据：不同传感器，不同格式，不同数值范围，不同单位，不同精度…
- ✓ 多维数据：集成多个感知设备同时感知同一对象的多个属性





# 物联网数据的特点

5

- ✓ 海量感知设备和节点：  
2020年物联网中对象的数量约为320亿部
- ✓ 传感器节点多数处于全时工作状态

- ✓ 动态变化，实时产生，有周期性，也有不确定性
- ✓ 数据增长速度快，时效性要求高（物联网实时访问控制）

海量

多态

动态

关联

- ✓ 多源异构数据：不同传感器，不同格式，不同数值范围，不同单位，不同精度…
- ✓ 多维数据：集成多个感知设备同时感知同一对象的多个属性



# 物联网数据的特点

6

- ✓ 海量感知设备和节点：  
2020年物联网中对象的数量约为320亿部
- ✓ 传感器节点多数处于全时工作状态

- ✓ 动态变化，实时产生，有周期性，也有不确定性
- ✓ 数据增长速度快，时效性要求高（物联网实时访问控制）

海量

多态

动态

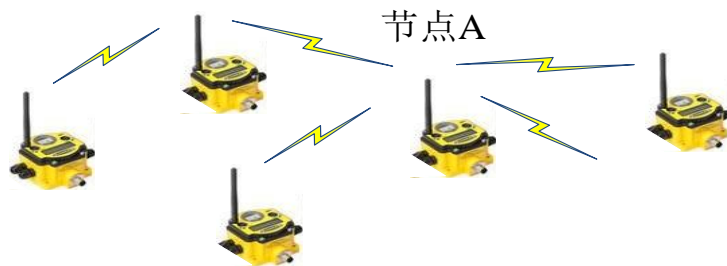
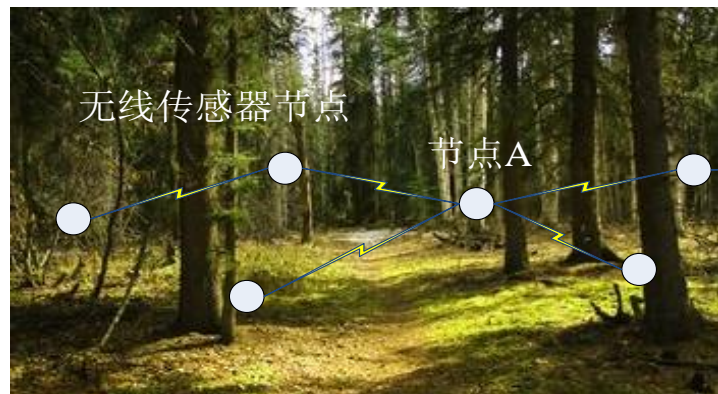
关联

- ✓ 多源异构数据：不同传感器，不同格式，不同数值范围，不同单位，不同精度…
- ✓ 多维数据：集成多个感知设备同时感知同一对象的多个属性
- ✓ 各类传感器或多个节点在时空维度上存在紧密关联性

# 物联网数据-信息-知识

7

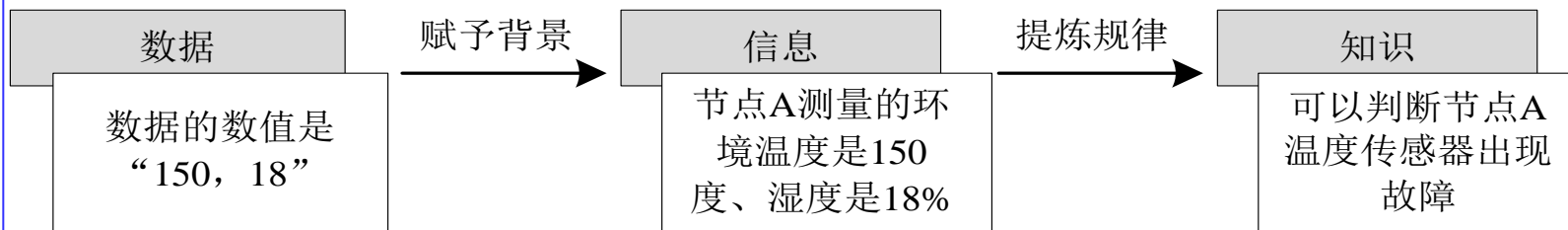
用于森林防火监控的无线传感器网络系统



森林防火监控中心



监控中心  
数据分析系统



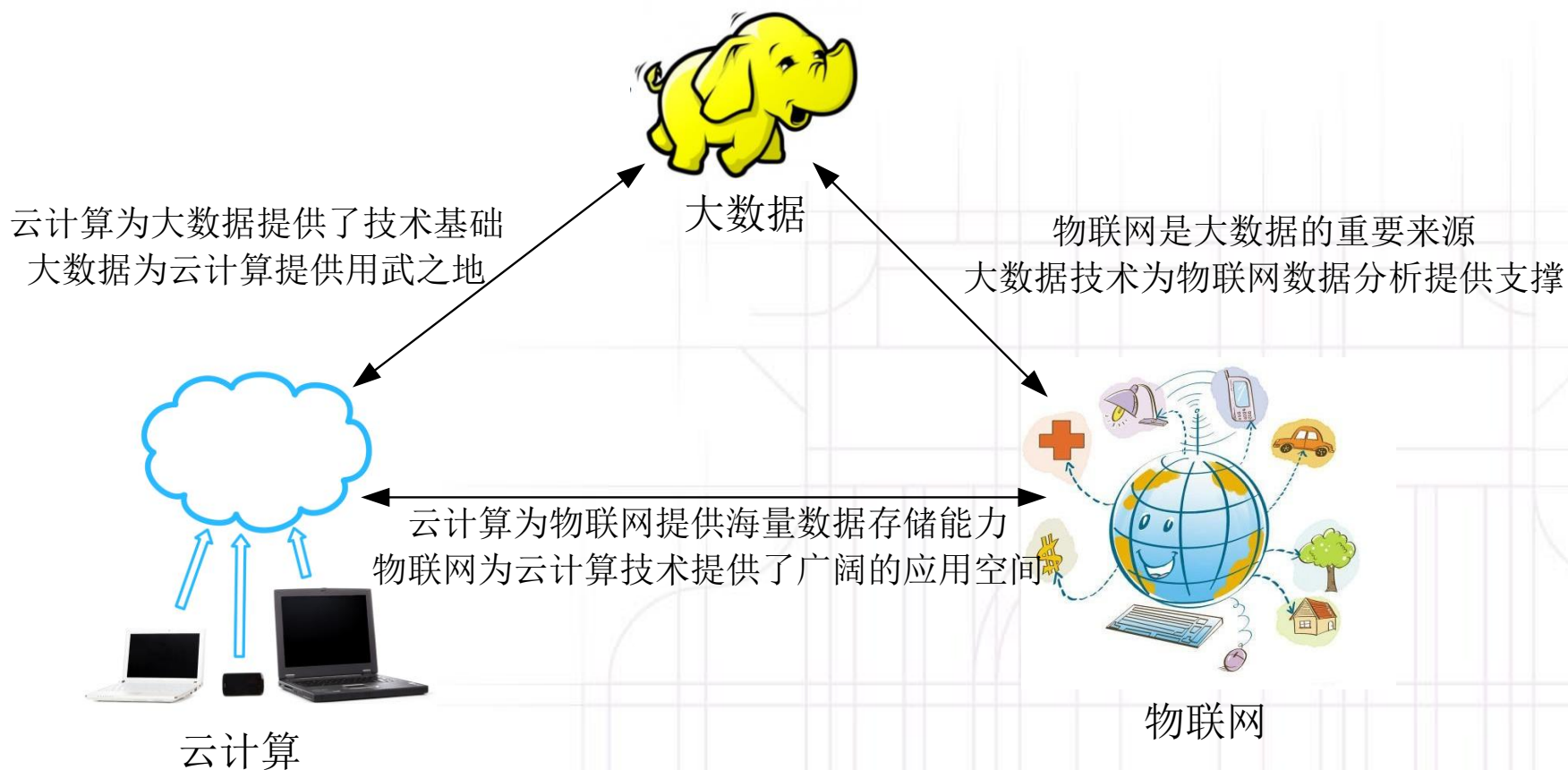
# 物联网数据处理的关键技术

8

- 海量数据存储技术
  - ▣ 数据中心与云存储技术
  - ▣ 数据安全技术
- 多态数据融合技术
  - ▣ 多传感器融合
- 数据挖掘技术
  - ▣ 数据查询与搜索技术
  - ▣ 数据挖掘技术
- 智能决策技术
  - ▣ 从数据中挖掘信息，提炼知识，形成智慧
  - ▣ 发展物联网的最终目标不是互联，而是催生具有计算、通信、控制、协同和自治特征的智能设备和系统，实现实时感知、动态控制和智能服务



# 物联网与大数据、云计算的关系



大数据、云计算和物联网之间的关系

1

# 物联网与大数据技术

# 本节内容

11

- 1 大数据热潮
- 2 大数据存储
- 3 大数据处理
- 4 物联网大数据研究要点

# “大数据”横空出世

12

- 《自然》杂志，最先提出新术语“Big Data”
  - 2008年9月发表“The Next Google”专刊
  - 将全世界的一切物质和信息集合起来组成一个巨大的数据库，打破虚拟和现实的边界，这将是世界可以期待的最大变革。
- 《科学》杂志
  - 2011年2月发表有关大数据的专刊《Dealing With Data》
  - 1700多位反馈者中有91.2%的人认为无法有效驾驭所拥有的数据
- 麦肯锡研究院
  - 2011年5月发布《Big data: The next frontier for innovation, competition and productivity》
  - 定义大数据：大小超出了常规数据库工具获取、储存和分析能力的数据集。



# 数据存储的发展历史

13

**甲骨文→纸质书籍→数字化存储**

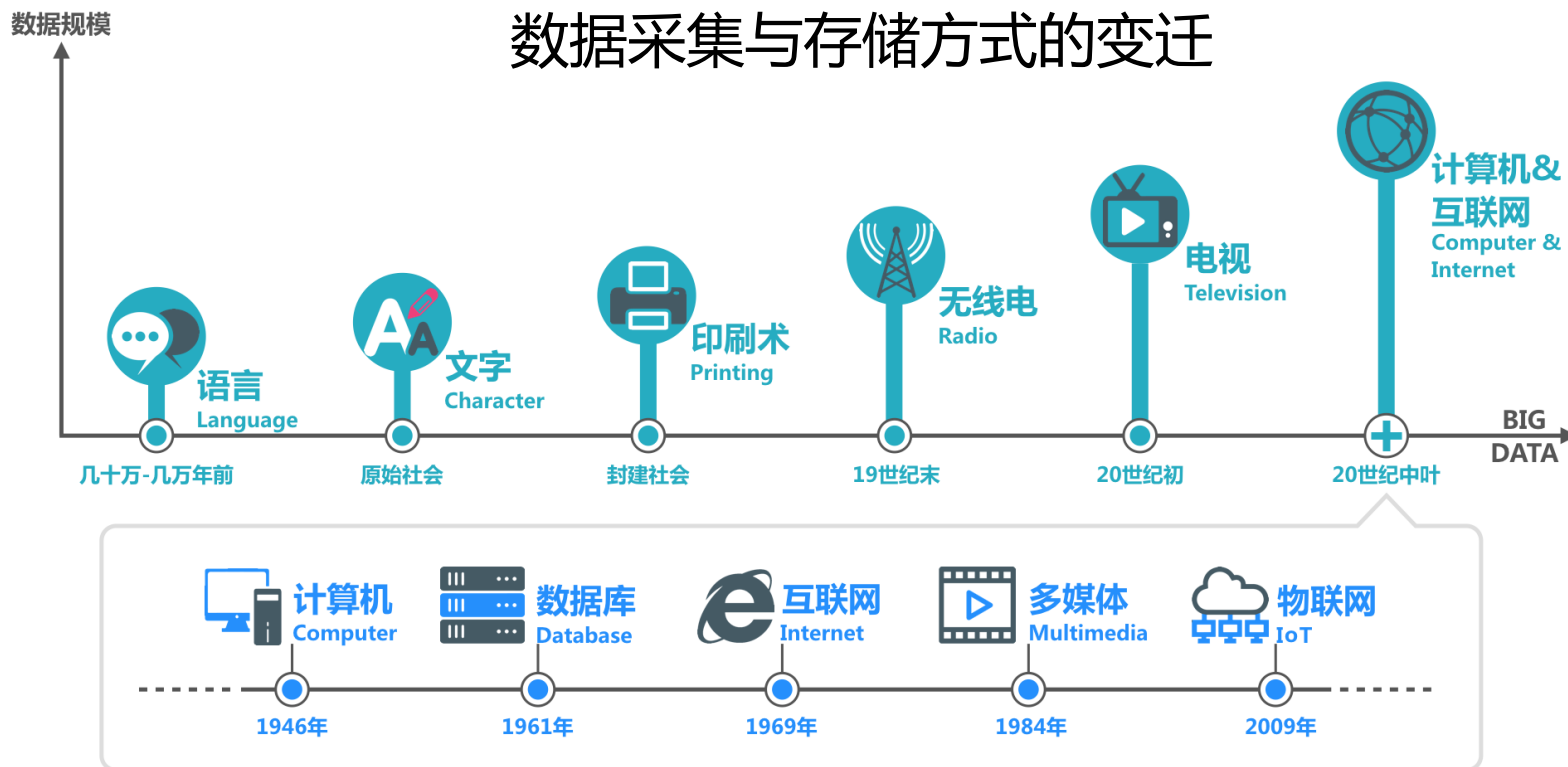


**信息存储介质发生了重要变化**



# 数据增长的历史

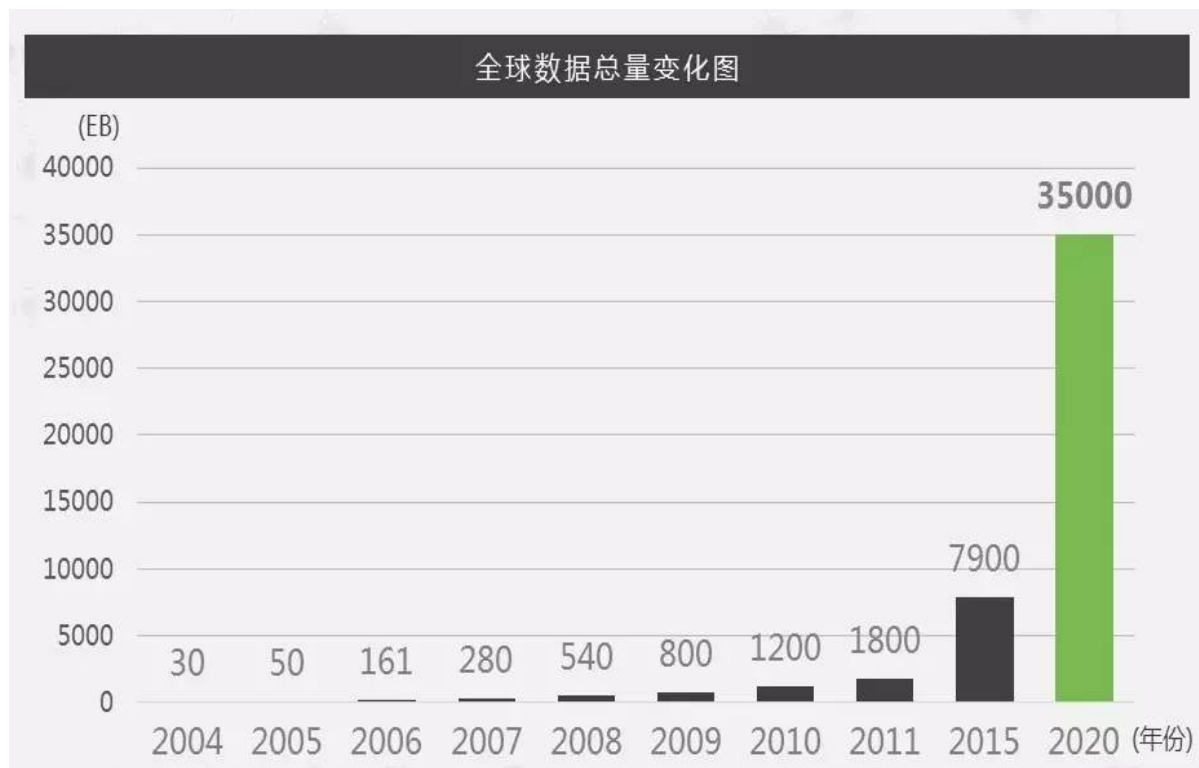
14



- 物联网的出现将联网终端由计算机扩展到传感器、标签、甚至不具备任何计算能力的设备。2020年，物联网中对象的数量约为320亿部

# 数据增长的历史

15



- 2007年产生的数据量为281EB（1EB=10亿GB），2011年1.8ZB，预期2020年35ZB，每两年翻一番
- 当前数据总量的80%是最近两年产生

# 附：数据量单位与换算关系

16

单位	英文标识	单位标识	大小	含意与例子
位	bit	b	0或1	计算机处理数据的二进制数
字节	byte	B	8位	计算机存储数据的基本物理单元，存储一个英文字母用1个字节表示，一个汉字用2个字节表示
千字节	KiloByte	KB	1024字节或 $2^{10}$ 个字节	一张纸上的文字约为5KB个字节
兆字节	MegaByte	MB	$2^{20}$ 个字节	一个普通的MP3格式的歌曲约为4MB
吉字节	GigaByte	GB	$2^{30}$ 个字节	一部电影大约是1GB
太字节	TeraByte	TB	$2^{40}$ 个字节	美国国会图书馆所有书籍的信息量约为15TB，截至2011年底其网络备份数据量为280TB，今后每个月以5TB的速度增长
拍字节	PetaByte	PB	$2^{50}$ 个字节	NASA EOS对地观测系统3年观测的数据量约为1PB
艾字节	ExaByte	EB	$2^{60}$ 个字节	相当于中国13亿人每人一本500页书的数据量的总和
皆字节	ZetaByte	ZB	$2^{70}$ 个字节	截至2010年人类拥有的信息量的总和约为1.2ZB
佑字节	YottaByte	YB	$2^{80}$ 个字节	超出想象 1YB=1024ZB=1 208 925 819 614 629 174 706 176 B
诺字节	NonaByte	NB	$2^{90}$ 个字节	超出想象
刀字节	DoggaByte	DB	$2^{100}$ 个字节	超出想象

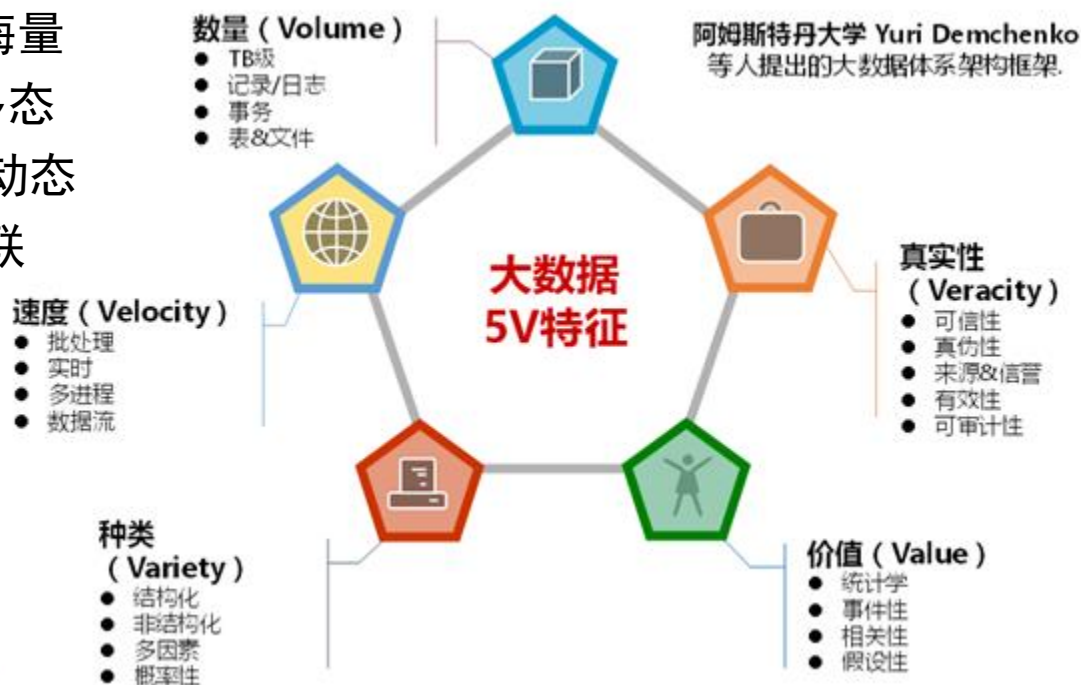
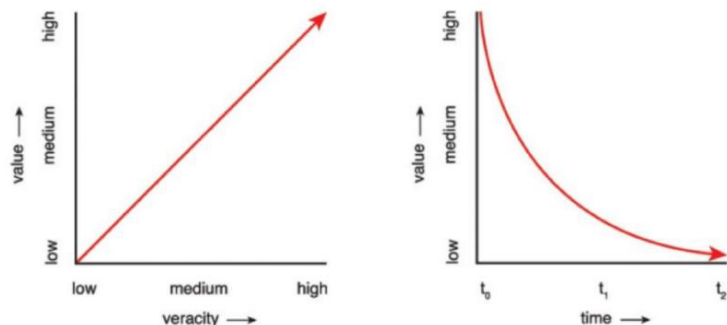
# 大数据特点

17

## □ 大数据

- 无法在一定时间内用常规软件工具对其内容进行抓取管理和处理的数据
- 大数据5V特征-物联网数据特点对比：

- 数量大 (Volume) - 海量
- 种类多 (Variety) - 多态
- 速度快 (Velocity) - 动态
- 价值高 (Value) - 关联
- 真实性 (Veracity) - 关联



# 物联网与大数据

18

- 物联网成为大数据的重要来源之一
  - ▣ 物联网中的设备数量持续快速增长
  - ▣ 数据量和有用数据比率相比传统应用明显升高
  - ▣ 为数据处理带来巨大挑战，同时也推动相关技术的发展
- 大数据为物联网的智能化发展提供有力保障
  - ▣ 建立与应用相关的数学模型
  - ▣ 运算系统的处理和计算
  - ▣ 多维度信息的整合和分析
- 物联网需要大数据技术提供适合其特点的大数据存储与处理技术



# 本节内容

19

- 1 大数据热潮
- 2 大数据存储
- 3 大数据处理
- 4 物联网大数据研究要点

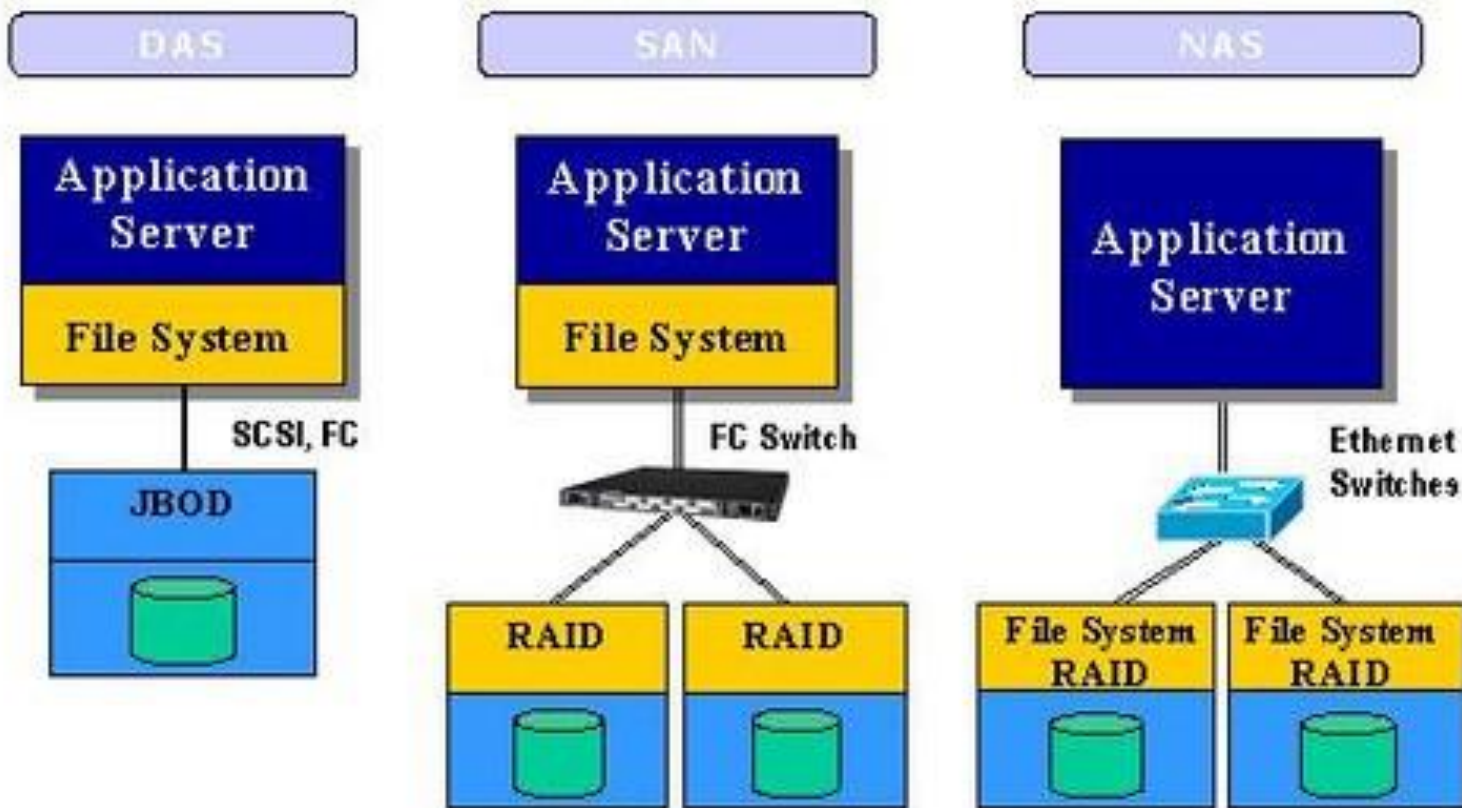
# 常见网络存储方式

20

- **直接附加存储**（Direct-Attached Storage, DAS）
  - 将存储系统通过缆线直接与服务器或工作站相连
  - 一般包括多个硬盘驱动器，与主机总线适配器通过电缆或光纤相连
  - 在存储设备和主机总线适配器之间不存在其他网络设备
- **网络附加存储**（Network Attached Storage, NAS）
  - 文件级的计算机数据存储架构
  - 计算机连接到一个仅为其它设备提供基于文件级数据存储服务的设备
- **存储区域网络**（Storage Area Network, SAN）
  - 通过网络方式连接存储设备和应用服务器的存储架构
  - 由服务器、存储设备和SAN连接设备组成
  - 存储共享
  - 支持服务器从SAN直接启动

# 三种网络存储结构的比较

21



# 三种网络存储结构的比较

22

## □ DAS

- 管理容易，结构简单；是一种对已有服务器的简单扩展，并没有真正实现网络互联；存储资源利用率低，资源共享能力差，造成“信息孤岛”。

## □ NAS

- 网络的存储实体，容易实现文件级别共享；性能严重依赖于网络流量，用户数过多，读写过频繁时性能受限。

## □ SAN

- 存储管理简化，存储容量利用率提高；无直接文件级别的访问能力，但可在SAN基础上建立文件系统。

以上存储方式只能满足中等规模的商用需求，大型公司需要大型数据中心

# 数据中心

23

## □ 维基百科：

- “数据中心是一整套复杂的设施。它不仅仅包括计算机系统和其它与之配套的设备（例如通信和存储系统），还包含冗余的数据通信连接、环境控制设备、监控设备以及各种安全装置。”

## □ Google：

- “多功能的建筑物，能容纳多个服务器以及通信设备。这些设备被放置在一起是因为它们具有相同的对环境的要求以及物理安全上的需求，并且这样放置便于维护。”而不仅仅是一些服务器的集合。



# 数据中心的起源与发展

24



大型机



微型机



大规模数据中心  
(Mega Data Center)

# 数据中心的起源与发展（续）

25

大规模数据中心已经得到推广（面向云计算提供服务）



# 数据中心标准

26

- 数据中心建设者面对的难题
  - ▣ 如何规划一个新的数据中心？
  - ▣ 怎样对数据中心进行升级？
- 数据中心标准对相关经验进行了总结
- ANSI/TIA/EIA-942（简称TIA-942）：数据中心标准
  - 电信产业协会（TIA）提出
  - 美国国家标准学会（ANSI）批准

# 数据中心标准：TIA-942

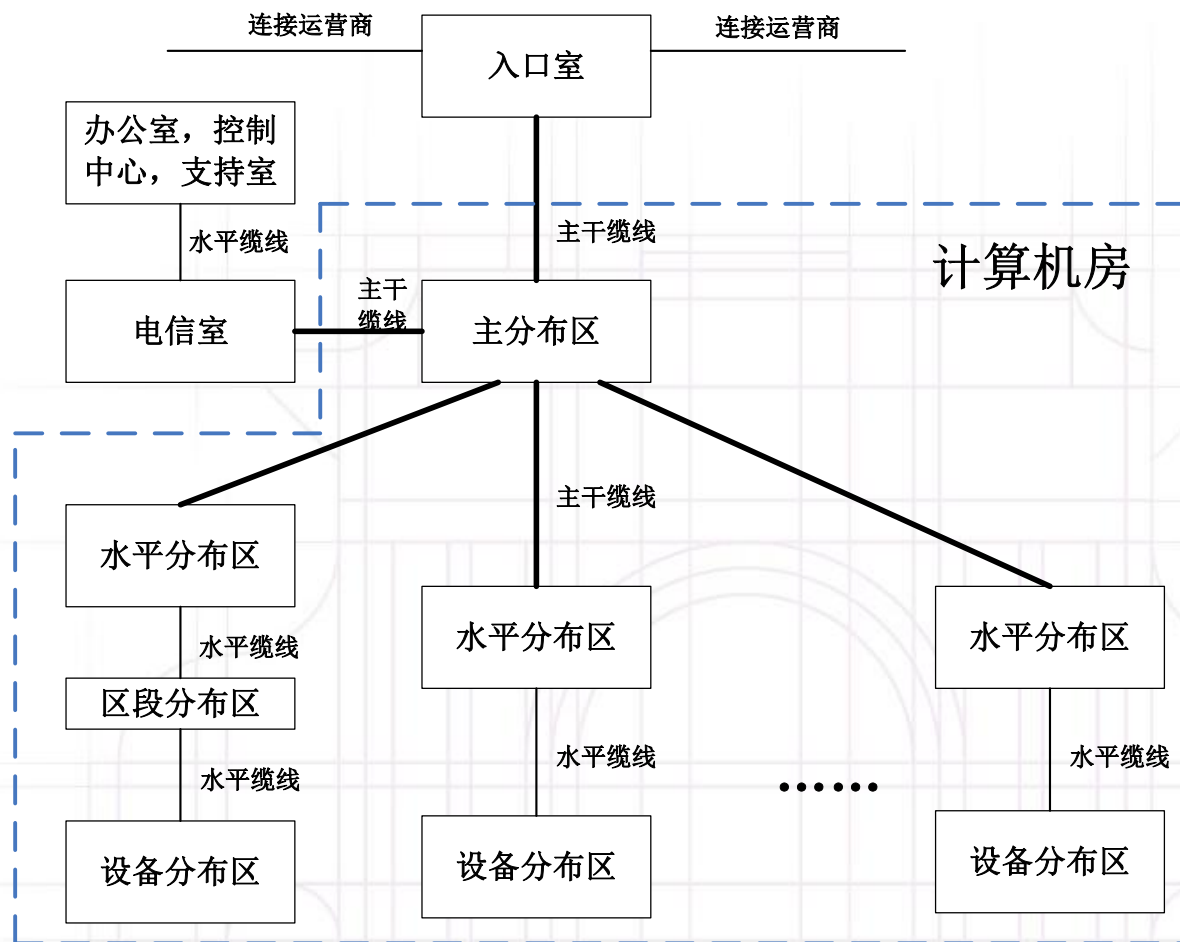
27

□ **选址：**需要考虑多方面因素

- 建设和运营成本
- 应用需求
- 政策优惠
- ...

□ **布局：**

- 按功能区域划分

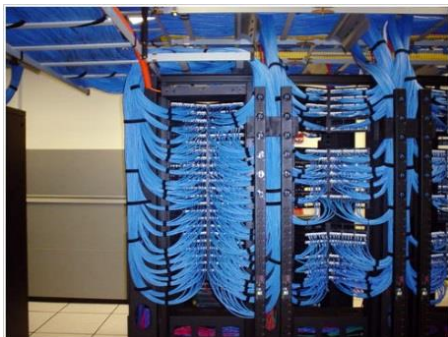


功能区域组成

# 数据中心标准：TIA-942

28

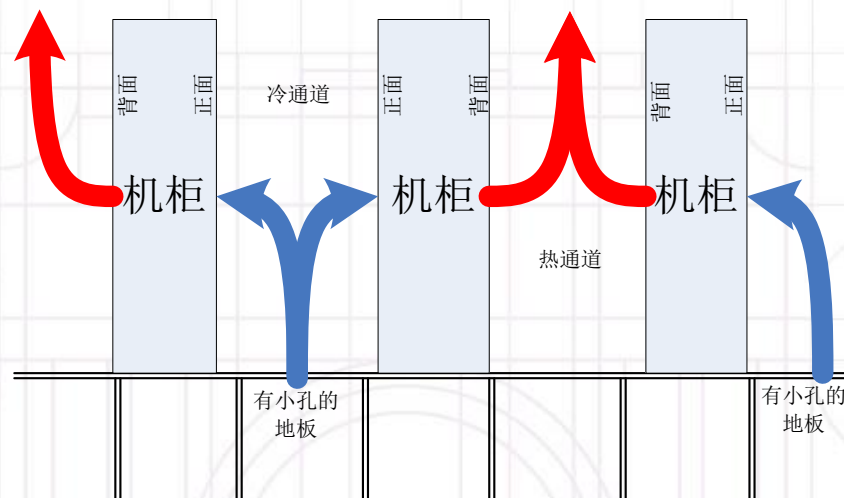
- TIA-942还对缆线系统，可靠性分级，能源系统和降温系统等做了规定。



缆线系统  
✓规格  
✓如何放置缆线



能源系统  
✓外部电力供应  
✓电池组  
✓发电机



降温系统  
✓降温设备  
✓架空地板  
✓冷通道与热通道



# Google数据中心

29

- Google数据中心选址：
  - 能源（廉价、低碳）、降温（大型水源）、空地面积（占地、保密）、与其他数据中心间距离（高速互联）、税收等
  - 在俄勒冈州，规划建设3个约6400平方米的中心机房



Google数据中心在全球的分布

# Google数据中心

30

- 2006年Google在数据中心项目上的花费为19亿美元，而2007年该项支出增加到24亿美元。
- Google在俄勒冈州的数据中心有近100兆瓦的功率，满负荷运行时消耗的电力基本上和纽卡斯尔（Newcastle）一个城市所有家庭的用电量加起来一样多。

Google在俄勒冈州哥伦比亚河边的数据中心



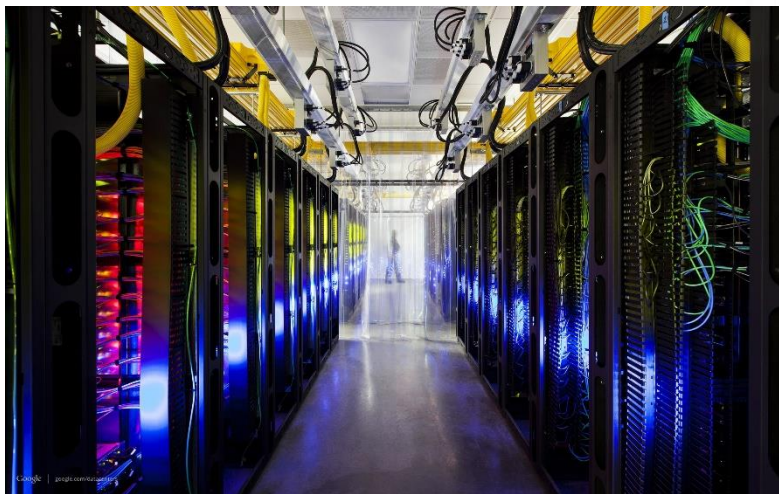
# Google数据中心

31

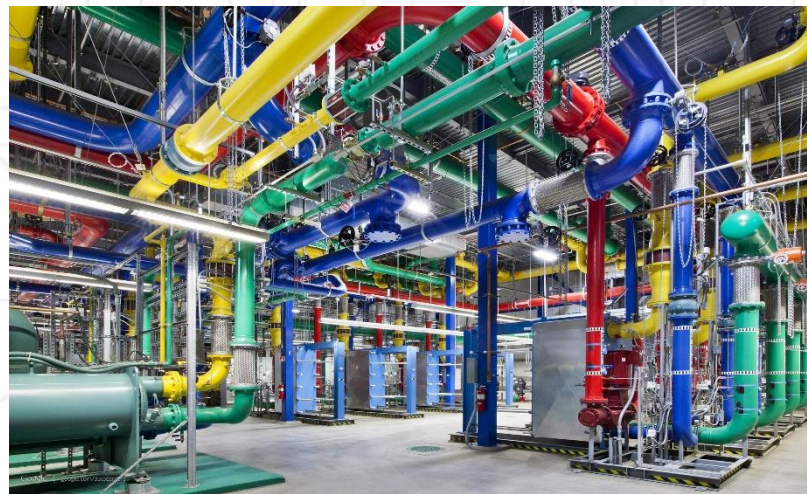
- Google数据中心能耗比低（Power Usage Effectiveness, PUE）
  - ▣ 能耗比=总能耗/IT设备能耗
  - ▣ 业界普遍为2，Google达到1.16，业界领先
- **措施一：数据中心运行高温化**
  - ▣ 数据中心普遍运行在21°C，Google运行在27°C
  - ▣ 条件苛刻：精准预测设备的崩溃点，温度控制的精确程度
- **措施二：特殊定制的能源系统**
  - ▣ 能源系统集成电池，替代UPS，达到99.99%的效率
  - ▣ 提高效率，降低能源系统能耗；同时减少热量输出，进一步节约降温系统的能耗

# Google数据中心

32



数据中心中的服务器及网络  
(黄色线缆为光纤)



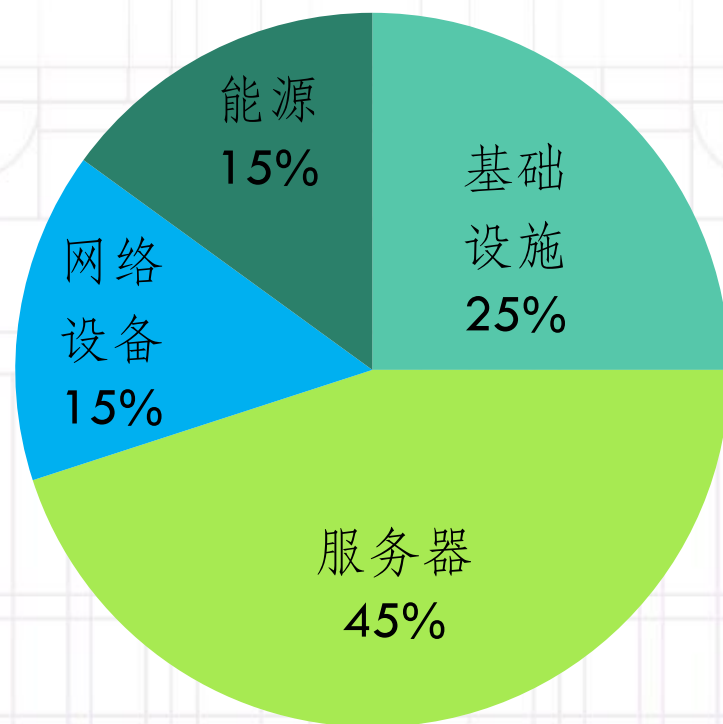
数据中心冷却系统  
(蓝色水管中为冷水，红色水管中为冷却后的热水)



# 数据中心的成本构成

33

- **研究热点：**如何在保证服务质量的前提下降低成本？
- 基础设施部分包括能源系统、降温系统的建设成本、各种防火设备、安保设备成本等。降低这一部分成本往往涉及到机械设备制造技术或政策优惠等因素，与计算机学科的关联程度相对较低。
- 分别从服务器，网络设备，能源三个方面对造成高成本的原因和目前的解决方法进行简要介绍。



# 服务器成本（45%）

34

- 服务器的实际利用效率较低
  - ▣ 分配到各服务器的应用不能完全利用某些组件
  - ▣ 对应用需求的预测比较难，无法做到按需分配
  - ▣ 为了提高系统的可靠性，一般都留有冗余设备
  
- 提高服务器利用率的关键在于及时应对需求的动态变化
  - ▣ 单个数据中心内敏捷性
  - ▣ 数据中心网络具有网络地址与位置无关、一致的带宽和延迟等特性

# 网络设备成本（15%）

35

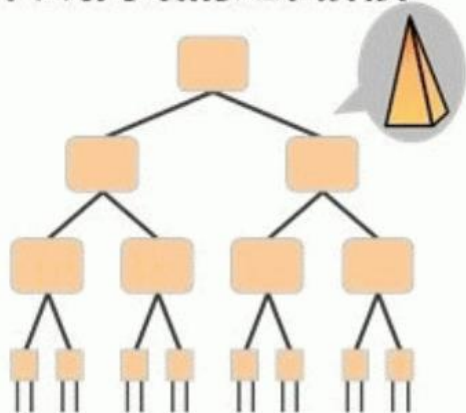
## □ 主要来源

- 交换机、路由器、负载均衡设备
- 传统的数据中心使用树形结构，核心交换机和路由器构成流量瓶颈，且造价昂贵

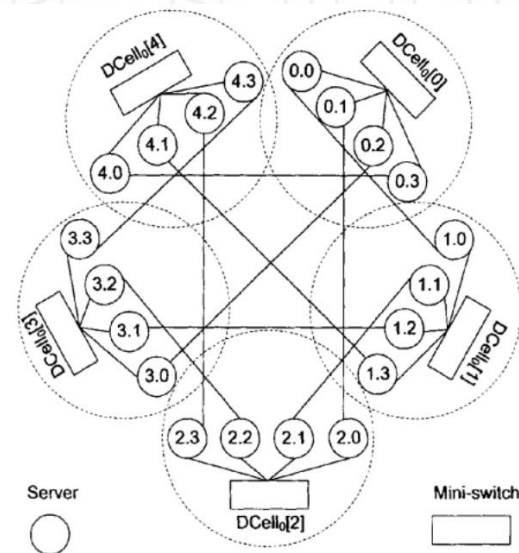
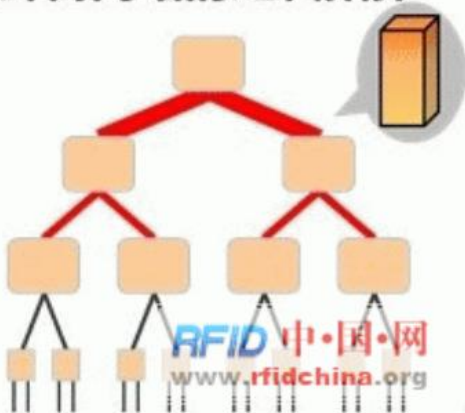
## □ 研究热点：新的数据中心网络结构

- 以交换机为中心的多层树形结构：例如Fat-Tree
- 以服务器为中心的互联结构：例如DCell

传统网络的逻辑拓扑



胖树网络的逻辑拓扑





# 能源成本（15%）

36

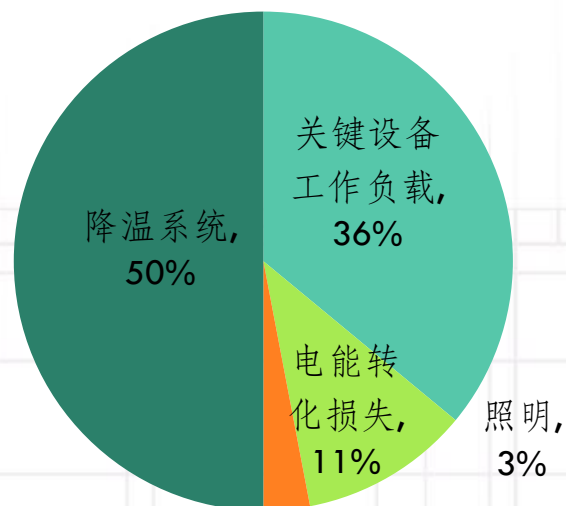
## □ 研究热点

### □ 降低服务器工作能耗

- 降低同等性能设备能耗
- 提高同等能耗设备性能
- 可调整负载的服务器

### □ 减少降温系统能耗

- 精细、精准的温度控制（WSN环境感知）
- 集装箱式模块化数据中心



# 本节内容

37

- 1 大数据热潮
- 2 大数据存储
- 3 大数据处理
- 4 物联网大数据研究要点

# Google数据处理需求

38

## □ Google数据需求

- 每月将近3.8亿用户
- 每月30亿次的搜索查询
- 每天大约处理超过20PB的数据
- 存储数十亿网页地址、数亿用户的个人资料

## □ Google硬件支持

- 全球共建有近40个大规模数据中心
- 独特的硬件设备：定制的以太网交换机、能源系统等

## □ Google软件支持

- 自行研发的分布式系统软件技术
- Google三驾马车：Google File System、MapReduce、BigTable
- 2003-2004以论文形式公开技术细节

# Google File System

39

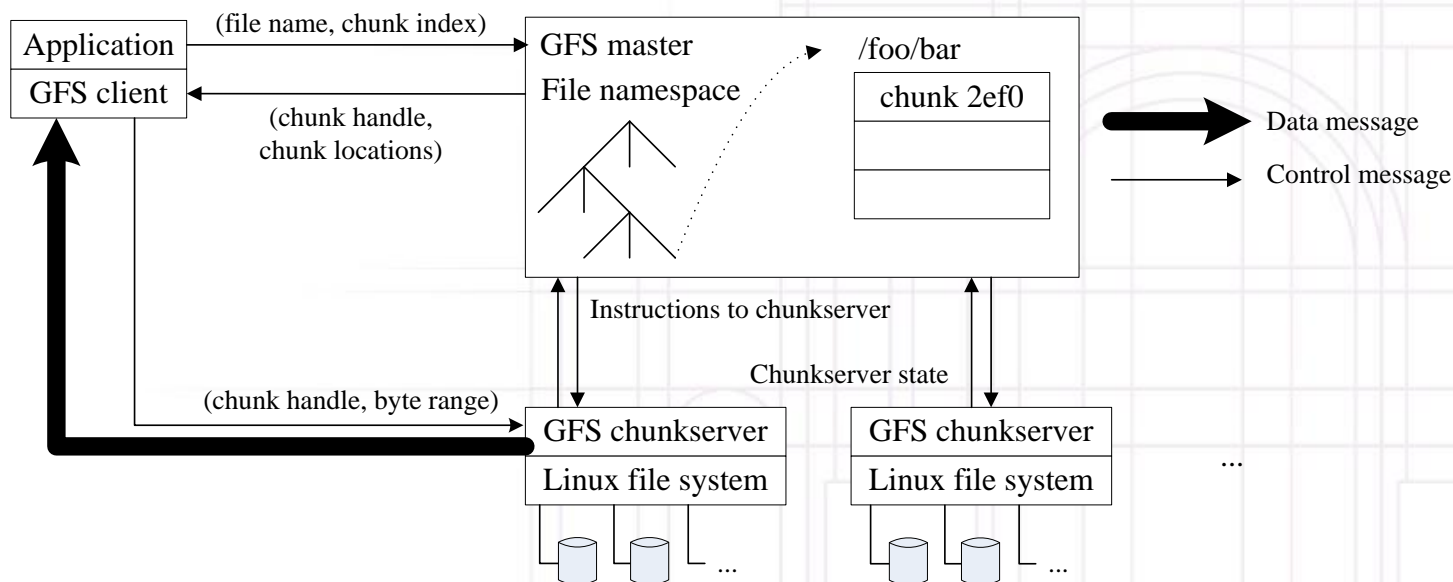
## □ GFS的设计观念

- 组件失效不再被认为是意外，而是被看做正常的现象
- GFS的文件非常巨大
- 对文件的操作以读取为主，写操作采用追加模式
- 应用程序和文件系统API的协同设计提高了整个系统的灵活性

# GFS的设计架构

40

- 一个GFS集群包含一个主服务器和多个块服务器，并被多个客户端访问。
- 文件分成固定大小的“块”。每个块在创建时都由主服务器分配一个固定不变的64位句柄唯一标识。
- 块服务器把块作为Linux文件存储在本地磁盘上（3个副本），并根据指定的块句柄和字节范围对数据块进行读写操作。



# GFS的设计架构

41

- **主服务器**维护所有文件系统的元数据，包括名字空间、访问控制信息、文件到块的映射信息以及块当前的位置。此外，主服务器还控制其它系统级的活动。主服务器周期性地与块服务器通信，以下达指令和收集状态。
- **GFS客户端代码**被嵌入到每个应用中。它实现了文件系统API，实现主服务器与块服务器的通信从而代表应用实现读写操作。客户端与服务器交互从而实现元数据操作，但所有的数据操作都通过直接与块服务器交互而完成。

# MapReduce

42

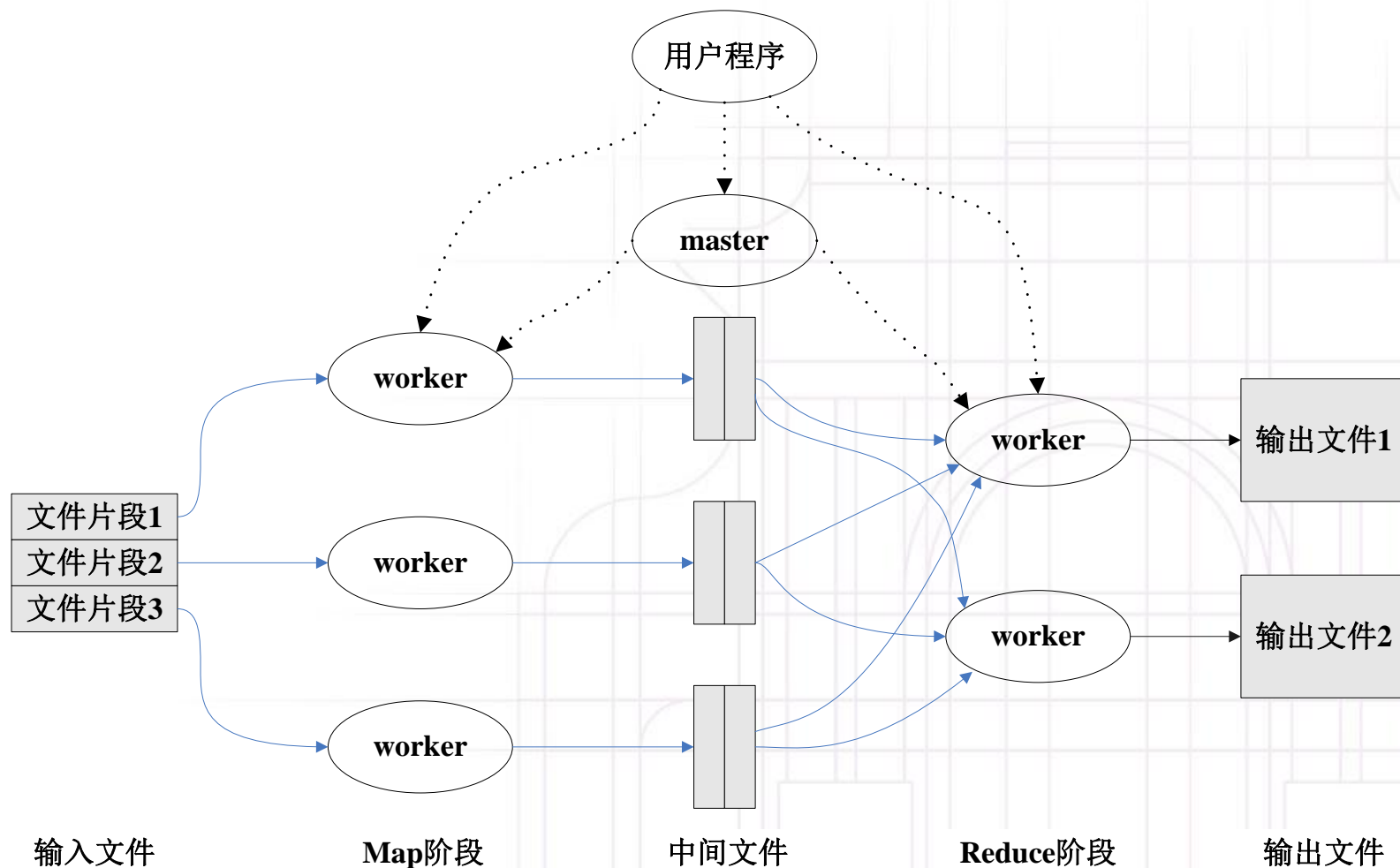
- **MapReduce**是一种针对超大规模数据集的编程框架
  - ▣ 用MapReduce开发出的程序可在大量商用计算机集群上并行执行、处理计算机的失效以及调度计算机间的通信
- MapReduce的基本思想
  - ▣ 一个在计算机集群上执行多个程序实例的框架
  - ▣ 用户写的两个程序：Map和Reduce
    - Map程序：从输入文件中读取数据集合，执行所需的查询或计算，以 (Key, Value) 的形式输出结果。
    - Reduce程序：按照用户定义的规则对Map的输出结果进行合并。



# MapReduce (续)

43

## □ MapReduce程序的执行过程



# MapReduce实例： WordCount程序任务

WordCount程序任务

程序	WordCount
输入	一个包含大量单词的文本文件
输出	文件中每个单词及其出现次数（频数），并按照单词字母顺序排序，每个单词和其频数占一行，单词和频数之间有间隔

一个WordCount的输入和输出实例

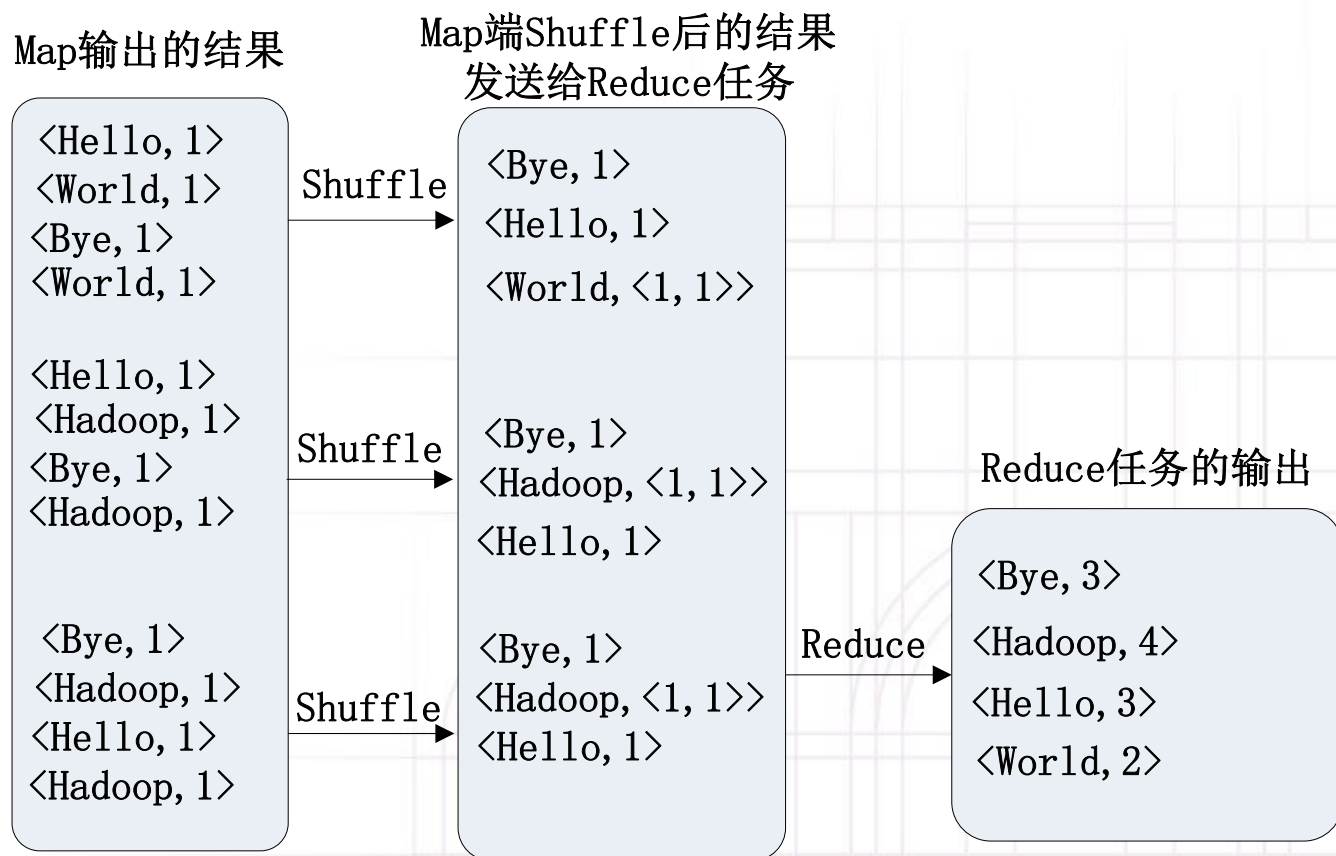
输入	输出
Hello World	Hadoop 1
Hello Hadoop	Hello 3
Hello MapReduce	MapReduce 1
	World 1

# MapReduce实例： WordCount执行过程



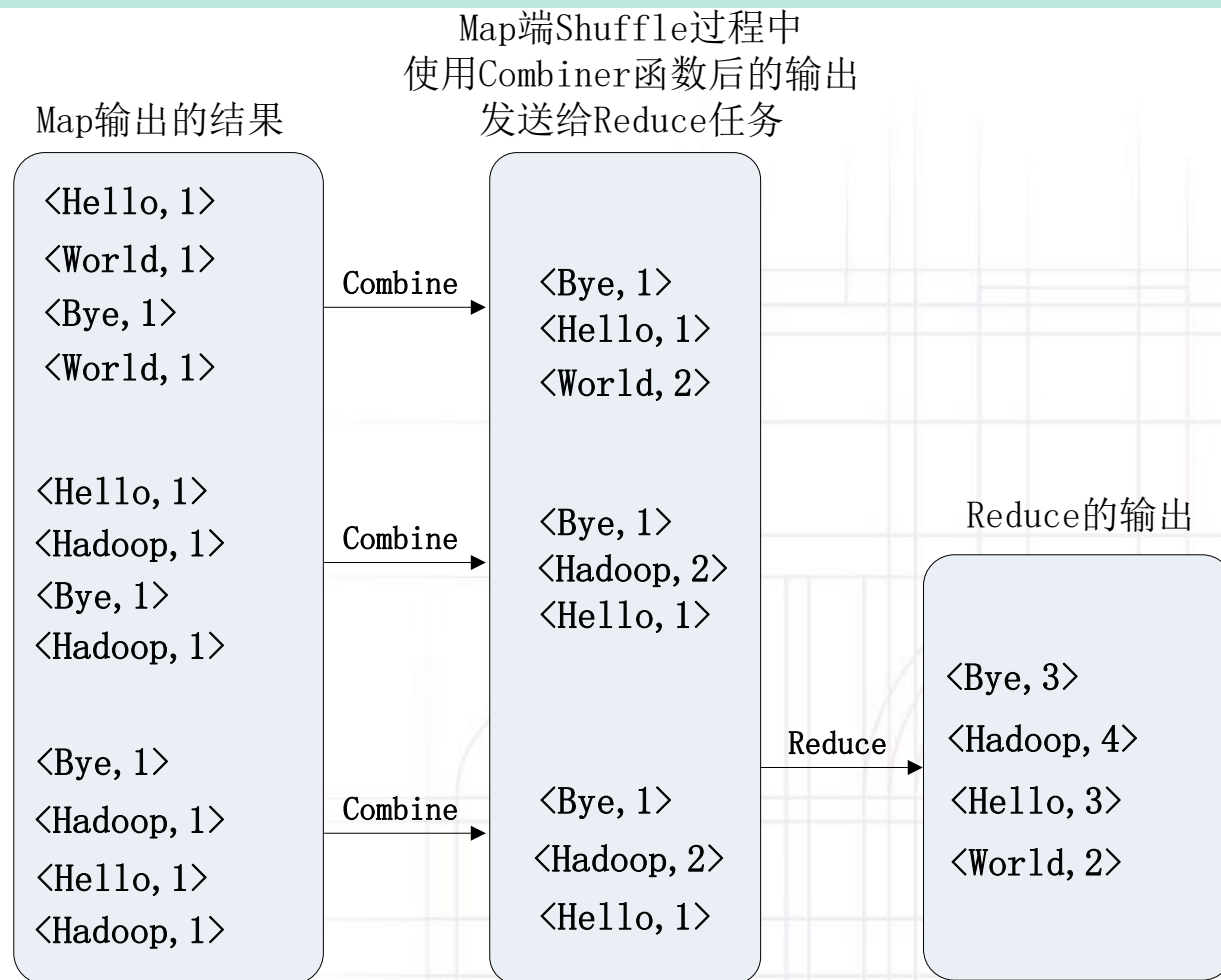
Map过程示意图

# MapReduce实例：WordCount执行过程



用户没有定义Combiner时的Reduce过程示意图

# MapReduce实例：WordCount执行过程

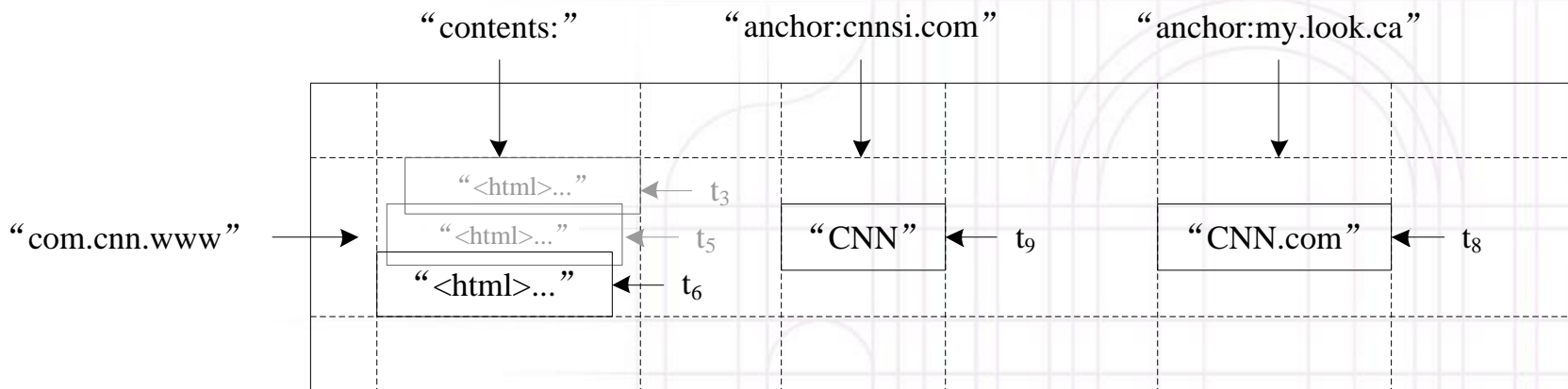


用户有定义Combiner时的Reduce过程示意图

# BigTable

48

- **BigTable**是一种用来在海量数据规模下（例如包含以PB为单位的数据量和数千台廉价计算机的应用）管理结构化数据的分布式存储系统。
- 典型应用：网页索引
- 主要结构：行、列族、时间戳



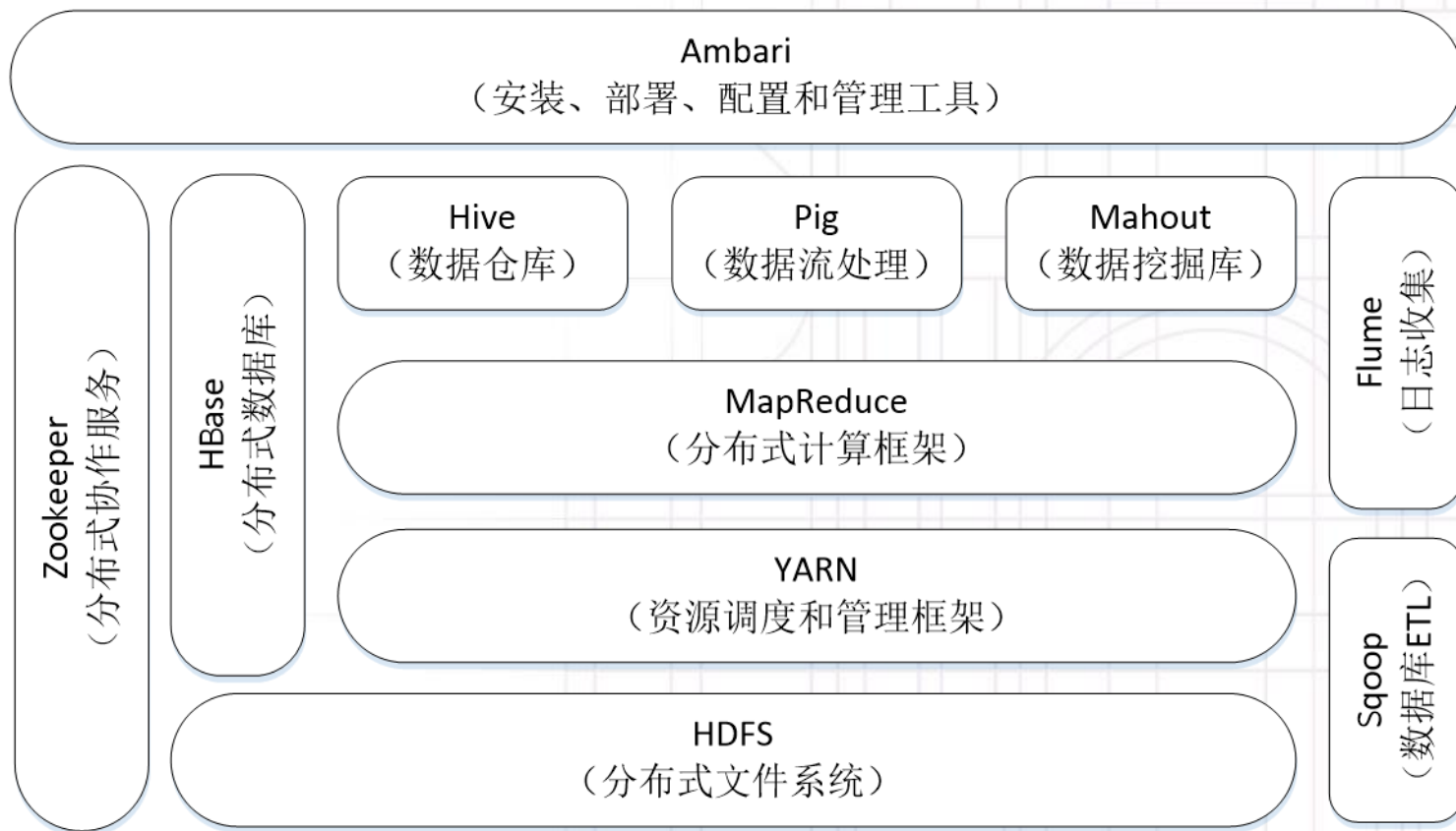


- 什么是Hadoop?
  - ▣ Apache开源组织的一个分布式计算开源框架
  - ▣ 用于在大型集群的廉价服务器设备上，对大量数据进行分布式处理的软件框架
  - ▣ 在早期实际上是Google文件系统与MapReduce分布式计算框架及相关IT基础服务的开源实现
- Hadoop特性
  - ▣ 高可靠性
  - ▣ 高可扩展性
  - ▣ 高容错性
  - ▣ 成本低
  - ▣ 运行在Linux平台上
  - ▣ 支持多种编程语言





- 经过多年的发展，Hadoop生态系统不断完善和成熟，目前已经包含了多个子项目。除了核心的HDFS和MapReduce以外，Hadoop生态系统还包括Zookeeper、HBase、Hive、Pig、Mahout、Sqoop、Flume、Ambari等功能组件。



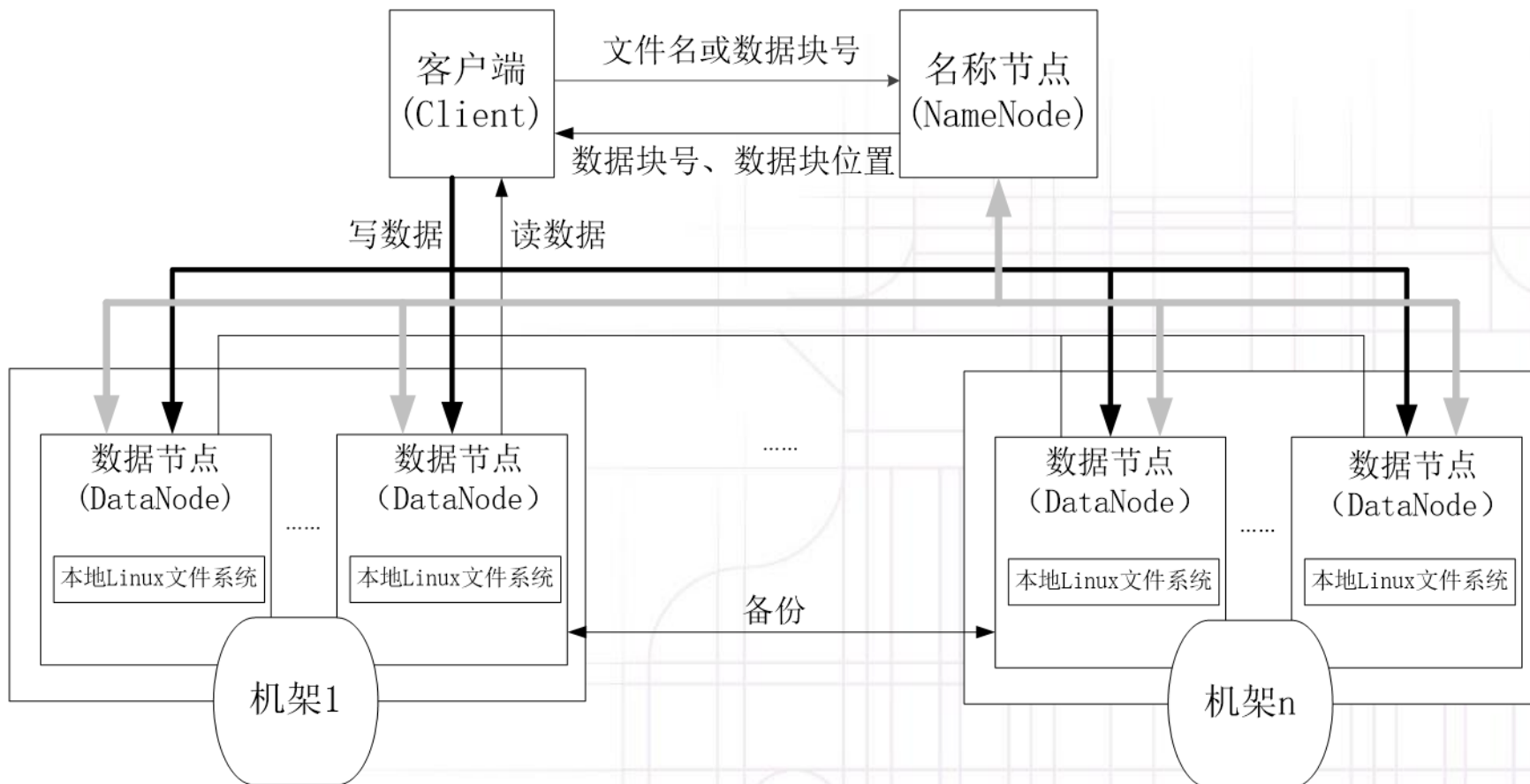
# Hadoop Distributed File System, HDFS

51

- HDFS采用了主从（Master/Slave）结构模型，一个HDFS集群包括一个名称节点（NameNode）和若干个数据节点（DataNode）
  - ▣ 名称节点作为中心服务器，负责管理文件系统的命名空间及客户端对文件的访问。
  - ▣ 集群中的数据节点一般是一个节点运行一个数据节点进程，负责处理文件系统客户端的读/写请求，在名称节点的统一调度下进行数据块的创建、删除和复制等操作。每个数据节点的数据实际上是保存在本地Linux文件系统中的

# Hadoop Distributed File System, HDFS

52



# Hadoop Distributed File System, HDFS

53

- HDFS优点：
  - ▣ 兼容廉价的硬件设备
  - ▣ 流数据读写
  - ▣ 大数据集
  - ▣ 简单的文件模型
  - ▣ 强大的跨平台兼容性
- HDFS特殊的设计，在实现上述优良特性的同时，也使得自身具有一些应用局限性，主要包括以下几个方面：
  - ▣ 不适合低延迟、随机数据访问
  - ▣ 无法高效存储大量小文件
  - ▣ 不支持多用户写入及任意修改文件

# Hadoop问题

54

- Hadoop存在如下一些缺点：
  - ▣ 表达能力有限
  - ▣ 磁盘IO开销大
  - ▣ 延迟高
  - ▣ 任务之间的衔接涉及IO开销
  - ▣ 在前一个任务执行完成之前，其他任务就无法开始，难以胜任复杂、多阶段的计算任务

- Spark最初由美国加州伯克利大学（UC Berkeley）的AMP实验室于2009年开发，是基于内存计算的大数据并行计算框架，可用于构建大型的、低延迟的数据分析应用程序
- 2013年Spark加入Apache孵化器项目后发展迅猛，如今已成为Apache软件基金会最重要的三大分布式计算系统开源项目之一（Hadoop、Spark、Storm）
- Spark在2014年打破了Hadoop保持的基准排序纪录
  - Spark/206个节点/23分钟/100TB数据
  - Hadoop/2000个节点/72分钟/100TB数据
  - Spark用十分之一的计算资源，获得了比Hadoop快3倍的速度
- 相比于Hadoop MapReduce，Spark主要具有如下优点：
  - Spark的计算模式也属于MapReduce，但不局限于Map和Reduce操作，还提供了多种数据集操作类型，编程模型比Hadoop MapReduce更灵活
  - Spark提供了内存计算，可将中间结果放到内存中，对于迭代运算效率更高

# 本节内容

56

- 1 大数据热潮
- 2 大数据存储
- 3 大数据处理
- 4 物联网大数据研究要点



# 物联网大数据研究的特殊性

57

## □ 异构性与多样性

- 物联网的数据来自不同的行业、不同的应用、不同的感知手段，这些数据可以进一步分为：状态数据、位置数据、个性化数据、行为数据与反馈数据，数据具有明显的异构性与多样性

## □ 实时性、突发性与颗粒性

- 物联网感知数据是系统控制命令与策略制定的基础，对物联网数据处理时间要求很高。
- 同时，事件发生往往很突然和超出预判，事先无法考虑周全。
- 物联网感知设备获得的数据很容易出现不全面和噪声干扰。

## □ 非结构化与隐私性

- 物联网应用系统中存在着大量图像、视频、语音、超媒体等非结构化数据，增加了数据处理的难度。
- 物联网应用系统的数据中隐含有大量企业重要的商业秘密与个人隐私信息，数据处理中的信息安全与隐私保护难度大。

# 大数据研究的机遇与挑战

58

- 数据以意想不到的方式在收集和利用
  - ▣ 数据收集无处不在。
  - ▣ 数据收集方式从专用设备（传感器等）向常用设备（路由器）转移。
- 数据以极简的方式在分析处理
  - ▣ “大数据基础上的简单算法比小数据基础上的复杂算法更有效”  
——《数据的非理性效果》 彼得·诺维格
- 数据以真实又诡异的方式在讲故事
  - ▣ 数据是真实的，数据却不会自己得出结论。
  - ▣ 不谨慎的分析和解读会引向不正确的结论。
- 数据是一种重要的资源
  - ▣ 数据创造全新的商业模式，带来经济价值。
  - ▣ 对数据资源的利用仍处于起步阶段。

# 本节小结

59

- 了解物联网的数据特点及其与大数据技术的关系。
- 了解大数据概念的提出过程和主要特点。
- 理解三种基本的网络存储体系结构（DAS，NAS，SAN）的基本概念以及各自的优缺点。
- 理解数据中心的概念，了解保证性能前提下降低数据中心成本的方法（服务器成本，网络设备成本，能源成本）。
- 以Google数据处理技术为例，了解GFS，MapReduce，BigTable等大数据处理技术的基本概念和特点。了解Hadoop分布式计算开源框架的特点。
- 了解物联网大数据研究的特点和难点。

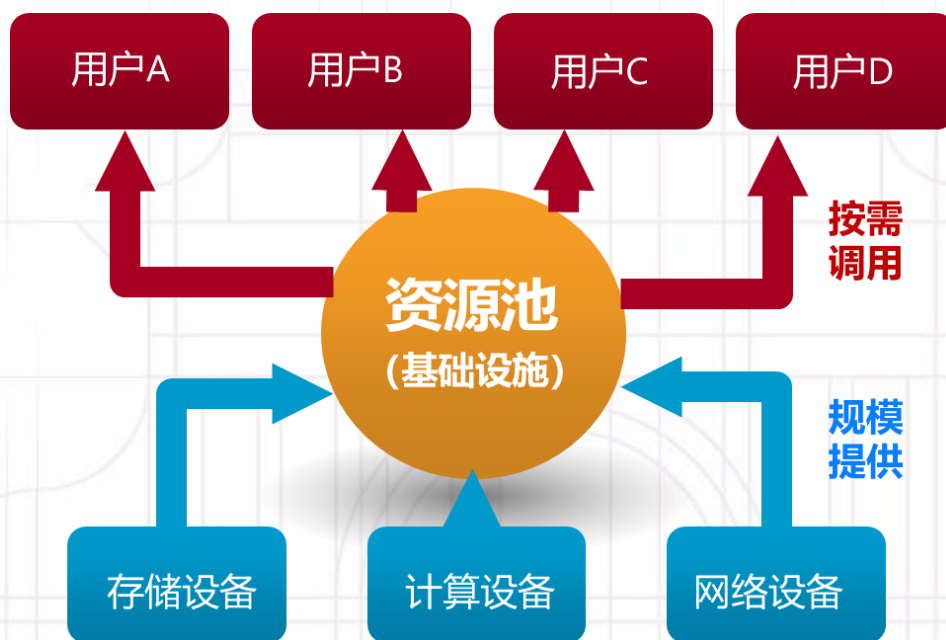
2

# 物联网与云计算技术

# 云计算

61

- 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，或服务供应商进行很少的交互。
- 云计算是分布式计算、并行计算、网格计算、网络存储、虚拟化、负载均衡、热备份冗余等传统计算机和网络技术发展融合的产物。



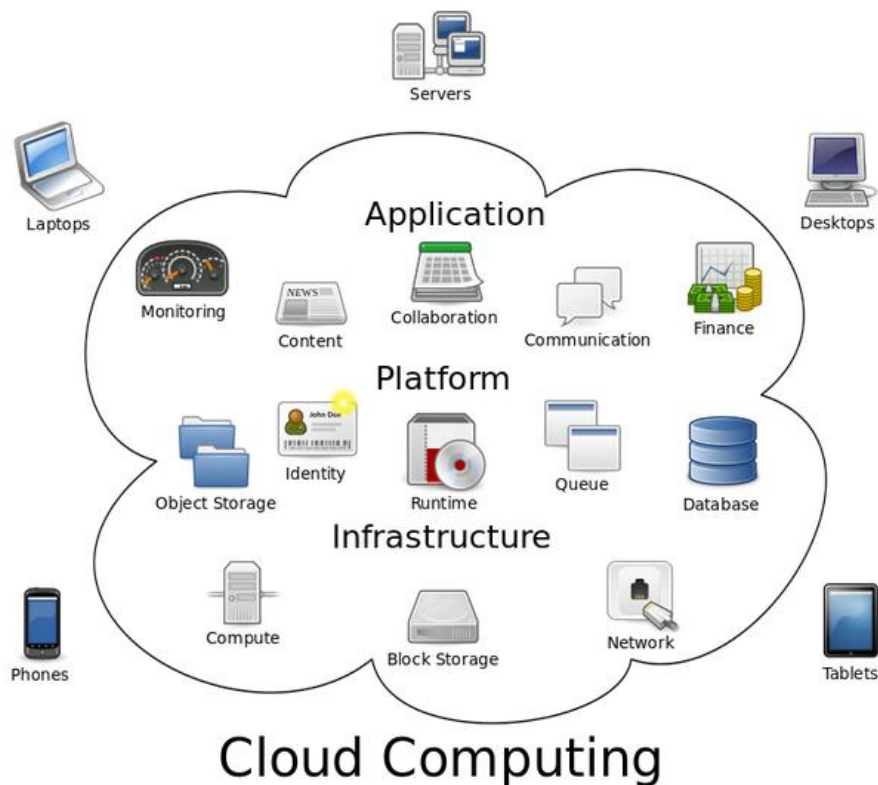
# 本节内容

62

- 1 云计算生态系统
- 2 服务器、操作系统和网络
- 3 虚拟化
- 4 云存储与云下载
- 5 “云物联”的展望

# 云计算的商业模式

63



软件即服务 (SaaS)

平台即服务 (PaaS)

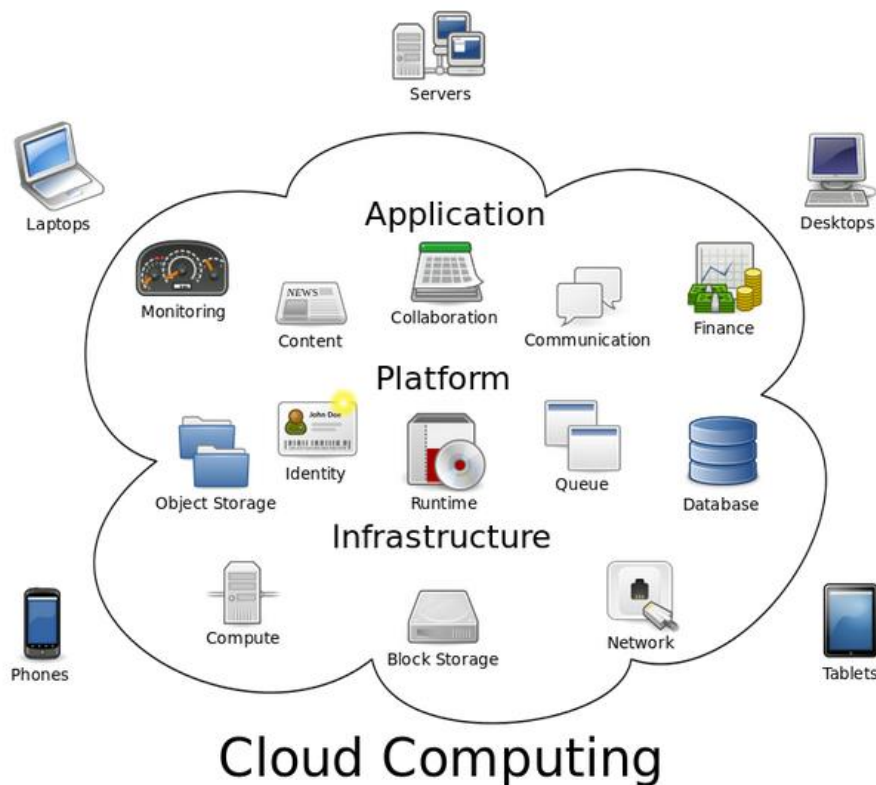
基础设施即服务 (IaaS)

- **基础设施即服务 (Infrastructure as a Service, IaaS)**
  - 提供基础设施资源。包括虚拟化的计算资源、存储资源、网络资源和安全保障等。
  - 例子：亚马逊的EC2、阿里云



# 云计算的商业模式

64



软件即服务 (SaaS)

平台即服务 (PaaS)

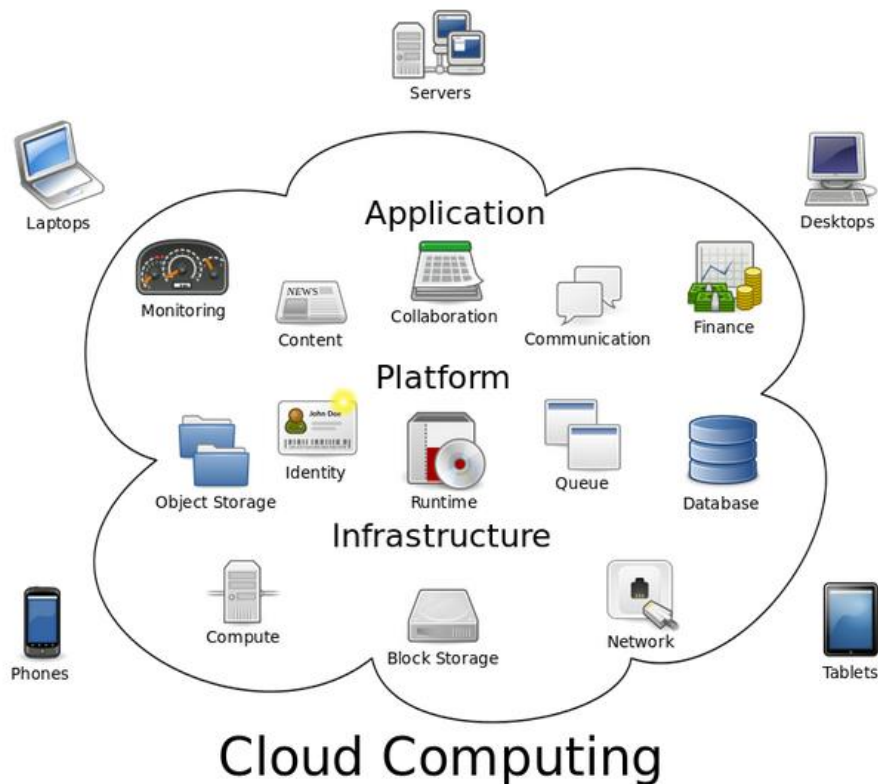
基础设施即服务 (IaaS)

## □ 平台即服务 (Platform as a Service, PaaS)

- 服务终端用户的应用程序。不操控硬件、网络、操作系统等基础资源，也不关心应用是如何开发调试的。
- 例子：微软的Azure App Service、谷歌的App Engine

# 云计算的商业模式

65



软件即服务 (SaaS)

平台即服务 (PaaS)

基础设施即服务 (IaaS)

## □ 软件即服务 (Software as a Service, SaaS)

- 服务应用的开发者。开发者通过这个平台开发、运行和管理应用程序时，无需处理诸如配置开发环境、测试环境等麻烦问题。
- 例子：云盘、文档在线编辑（如谷歌Docs和Evernote）

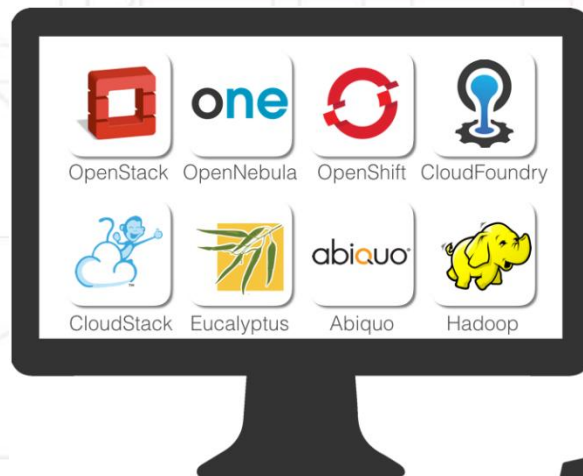
# 云计算的部署模式



# 云计算生态系统

67

## □ 商业云 & 开源云



# 云计算主要供应商



# 云计算的应用商



墨迹天气



# 数据中心与云计算

70

- Google的数据中心属于“自产自销”模式，同时提供产品和服务
- Amazon开创了云存储和云计算的商业模式
  - 弹性计算云：提供海量数据计算服务
  - 简单存储服务(S3)：可伸缩、可靠、高可用、低成本的存储服务
  - Dropbox后台即架设于S3之上
- 未来物联网
  - 不同商业机构共享云存储，而不需要建立自己的数据中心
  - 对海量数据的分析和处理也可以依托云计算进行



# 本节内容

71

- 1 云计算生态系统
- 2 服务器、操作系统和网络
- 3 虚拟化
- 4 云存储与云下载
- 5 “云物联”的展望

# 服务器

72

- 对于任何一个云计算系统，（物理）服务器都是最基础、最重要的硬件。
- 逻辑功能
  - ▣ 服务器和个人电脑不存在本质区别。
- 工作性能
  - ▣ 服务器在硬件配置、对外接口、稳定性、可用性、安全性等方面都远超过个人电脑，所以价格也相应昂贵得多。
- 操作方式
  - ▣ 服务器通常是默认没有图形界面的，操控服务器往往只能在远程登录之后以命令行方式进行。

# 服务器

73

## □ 硬件配置

- ▣ 服务器的CPU多使用英特尔至强（Xeon）系列或AMD皓龙（Opteron）系列，内存高达几十GB甚至上百GB都是很常见的；存储空间则一般不固定，可以很大也可以较小。

## □ 对外接口

- ▣ 服务器提供多个存储扩展接口供使用者随时以热插拔方式添加新的存储介质，且一般会提供两个以上以太网接口（一个接入外部网，另一个接入内部网）。



# 服务器

74

## □ 稳定性

- ▣ 服务器一旦开机运行，就要求高度稳定。

## □ 可用性

- ▣ 服务器通常都是全天候运行，无法访问的概率越低越好，一般应该低于0.1%或0.01%。

## □ 安全性

- ▣ 服务器采用一系列软硬件措施保护数据安全，包括硬件防火墙、软件防火墙、用户权限控制、访问控制列表等。

# 操作系统

75

- 绝大多数的服务器都运行Linux/UNIX系列的操作系统，其中Linux的使用最为广泛。
- Linux内核高度稳定——把容易破坏操作系统稳定性的图形界面放到内核之外。

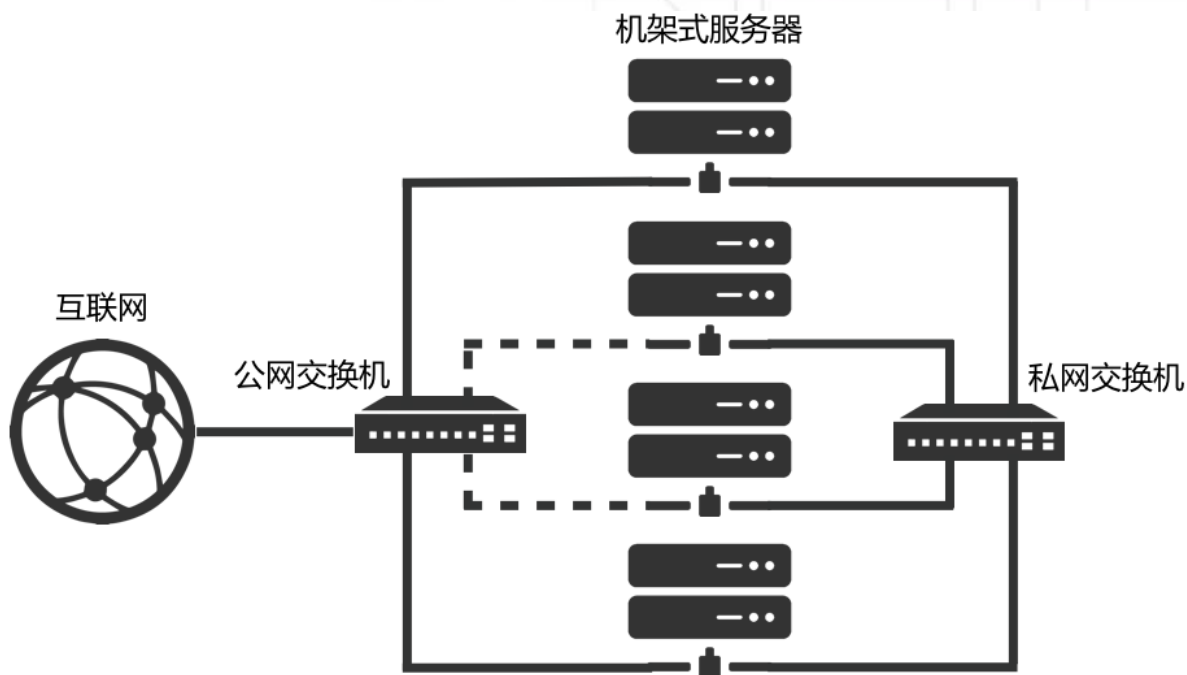


# 网络环境

76

## □ 公网/私网环境

- 最典型的网络环境配置方法就是：将所有服务器连接到一台私网交换机，同时将需要连入互联网的服务器（例如Web服务器）连接到一台公网交换机（通常由数据中心提供）。



# 本节内容

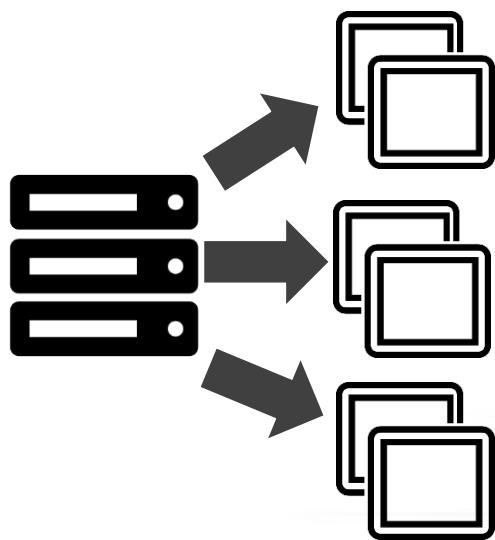
77

- 1 云计算生态系统
- 2 服务器、操作系统和网络
- 3 虚拟化
- 4 云存储与云下载
- 5 “云物联”的展望

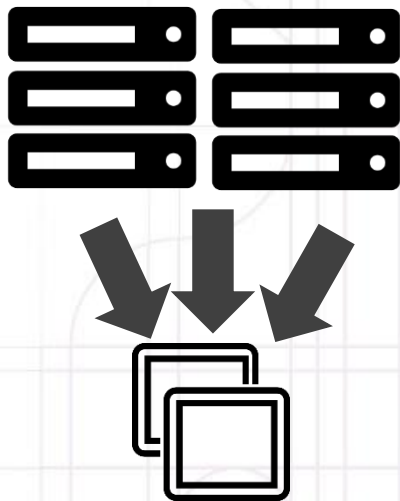
# 虚拟化

78

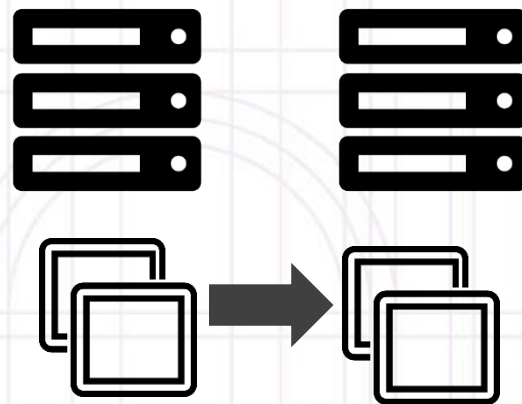
**虚拟化**是云计算的关键技术，它把刚性的物理硬件软件化成柔性的虚拟资源。其功能主要包括：



拆分



组合






动态配置和迁移



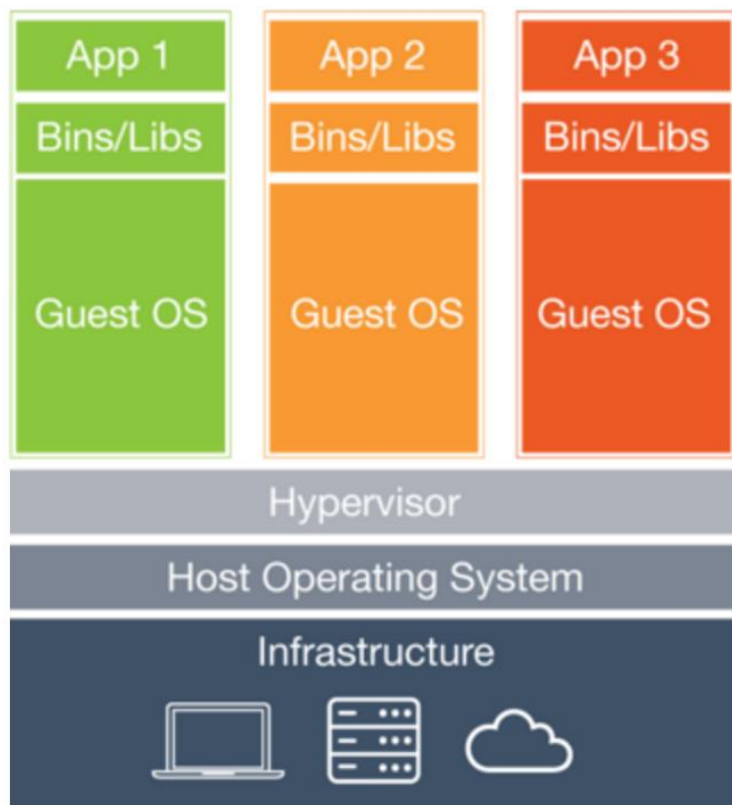
# 虚拟化层次

79

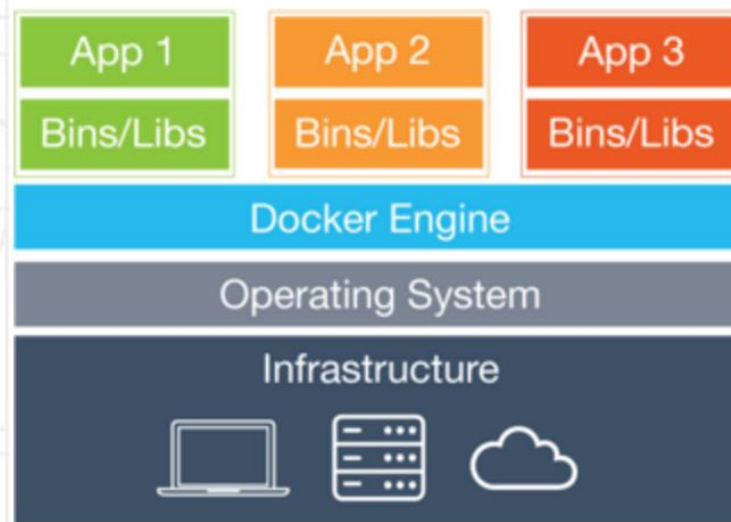
虚拟化层次	代表性系统
应用层虚拟化	  
内核层虚拟化	
半虚拟化	
硬件辅助虚拟化	 
沙盒和LXC	 
Docker	

# 虚拟化技术对比

80



虚拟机



**Docker**容器

# 基于虚拟机的虚拟化

81

它通过一个**软件层的封装**，提供和物理硬件相同的输入输出表现，实现了操作系统和计算机硬件的**解耦**，将OS和计算机间从1对1变成了多对多（实际上是1对多）的关系。

该软件层称为**虚拟机管理器（VMM/Hypervisor）**，它可以直接运行在裸机上（Xen、VMware ESXi），也可以运行在操作系统上（KVM、VMware workstation）。

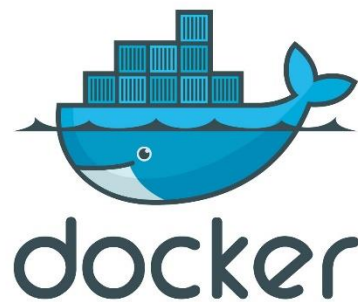
这项技术已经发展了40多年，但仍然存在以下几个问题：

- ❑ 在虚拟机上运行了一个完整的操作系统（GuestOS），在其下执行的还有虚拟化层和宿主机操作系统，一定比直接在物理机上运行相同的服务**性能差**；
- ❑ 有GuestOS的存在，虚拟机镜像往往有几个GB甚至几十个GB，**占用的存储空间大，便携性差**；
- ❑ 想要使用更多硬件资源，需要启动一台新的虚拟机，要等待GuestOS启动，可能**需要几十秒到几分钟不等**。

# 基于虚拟容器的虚拟化

82

- 容器是没有GuestOS的轻量级“虚拟机”，多个容器共享一个OS内核，容器中包含需要部署的应用和它依赖的系统环境，容器大小通常只有几十到几百MB。
- 由于共享操作系统内核，所以容器依赖于底层的操作系统，各个操作系统大都有自己的容器技术和容器工具。
- Docker是目前应用最广泛的一种Linux容器管理工具。
- 由于虚拟容器的出现，诞生了Caas (Container as a service) 服务模式



# Docker容器和虚拟机的对比

83

	Docker容器	虚拟机 (VM)
操作系统	与宿主机共享OS	宿主机OS上运行虚拟机OS
存储大小	镜像小，便于存储与传输	镜像庞大 (vmdk、vdi等)
运行性能	几乎无额外性能损失	操作系统额外的CPU、内存消耗
移植性	轻便、灵活，适应于Linux	笨重，与虚拟化技术耦合度高
硬件亲和性	面向软件开发者	面向硬件运维者
部署速度	快速，秒级	较慢，10s以上

容器技术与传统虚拟机性能对比

# 本节内容

84

- 1 云计算生态系统
- 2 服务器、操作系统和网络
- 3 虚拟化
- 4 云存储与云下载
- 5 “云物联”的展望

# 云存储（1）

85

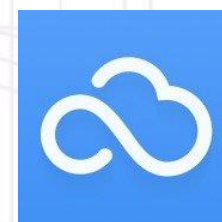
## □ 国外

- Dropbox
- OneDrive
- Google Drive
- iCloud Drive
- .....



## □ 国内

- 百度云盘
- 360云盘
- 腾讯微云
- .....



## 云存储（2）

86

### □ 方便而可靠地存取和分享数据

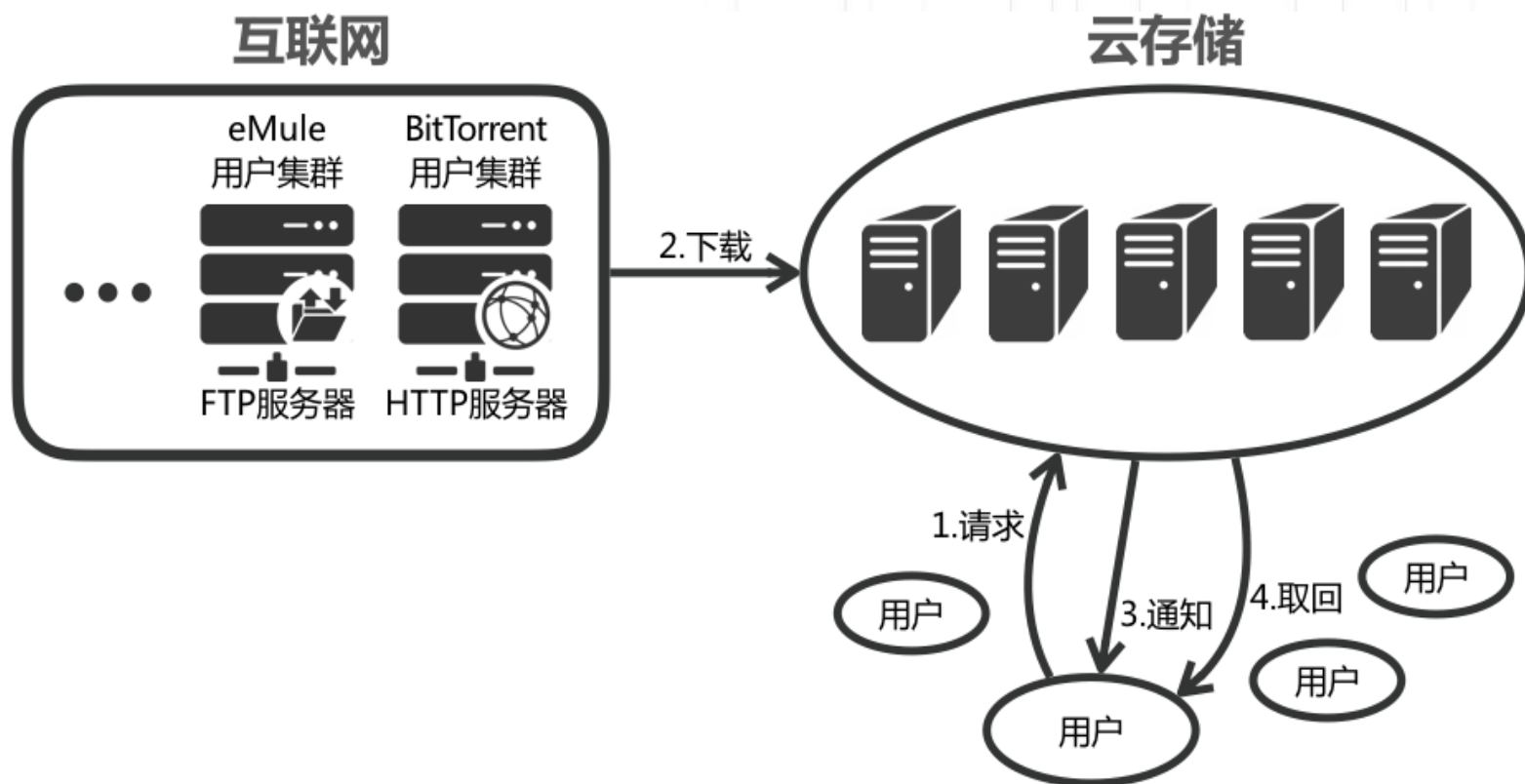
无论何时何地，用户使用任意终端设备，其存放在云端的数据都会被自动地同步到（该用户的）所有在线设备和其他共享用户的设备上。





# 云下载

87



# 本节内容

88

- 1 云计算生态系统
- 2 服务器、操作系统和网络
- 3 虚拟化
- 4 云存储与云下载
- 5 “云物联”的展望

# 云计算展望



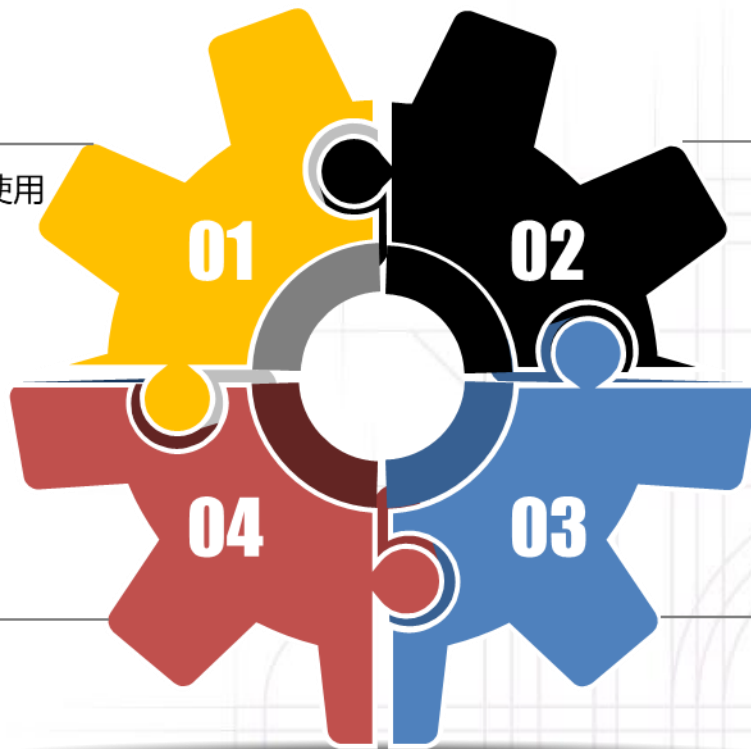
## 行业的使用

IT行业：经营中如水电般默认使用  
传统行业：逐步接受云计算



## 云安全成为挑战

无法预见的信息风险  
无法控制的潜在威胁



## 云计算的部署



公有云：长期的主流  
私有云：大企业不可缺，部署要求高  
混合云：将大行其道，少有单纯公有云

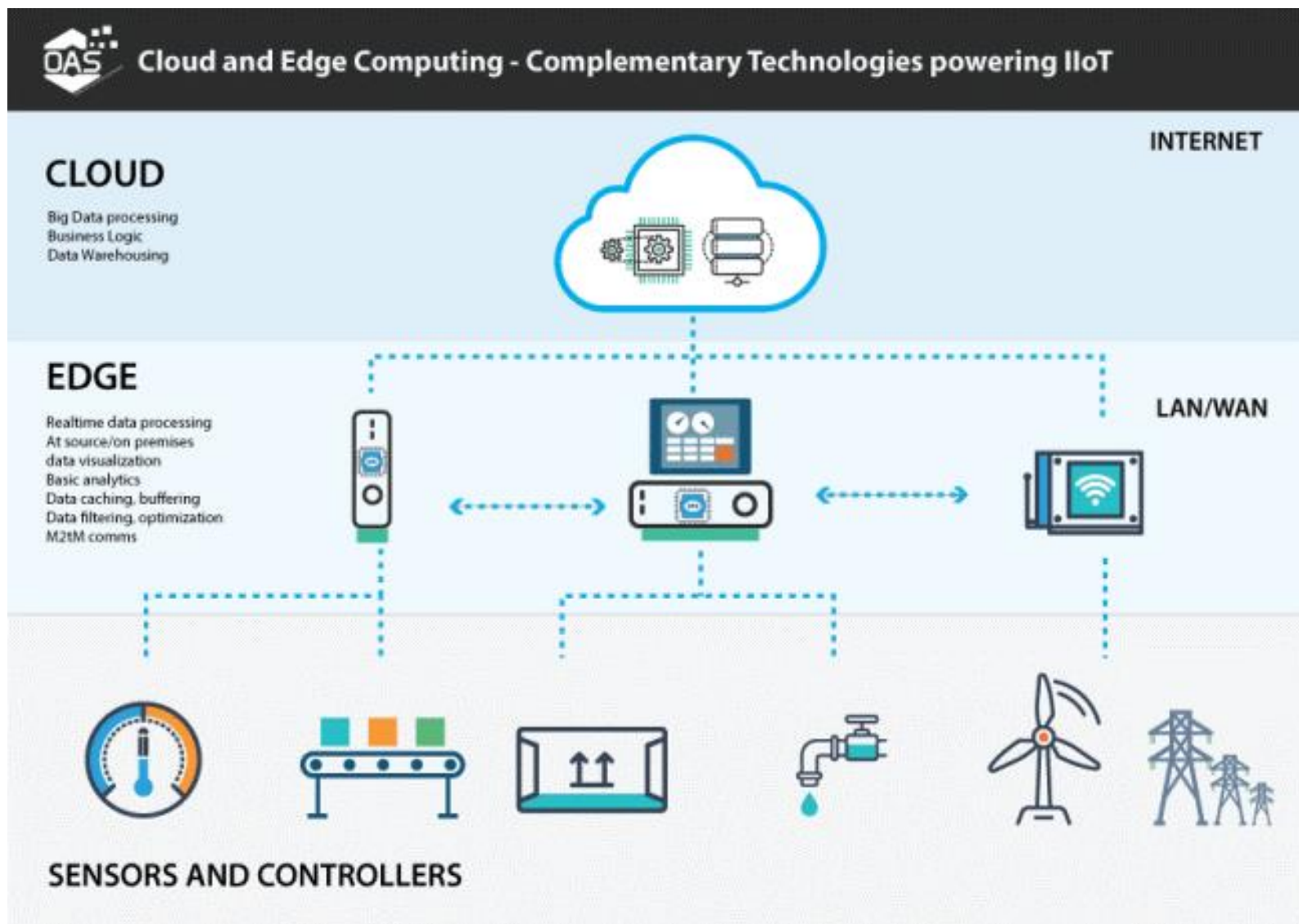
## 更加现实的应用



脱离底层技术讨论，聚焦应用实现  
万物互联：成为物联网技术的重要依托

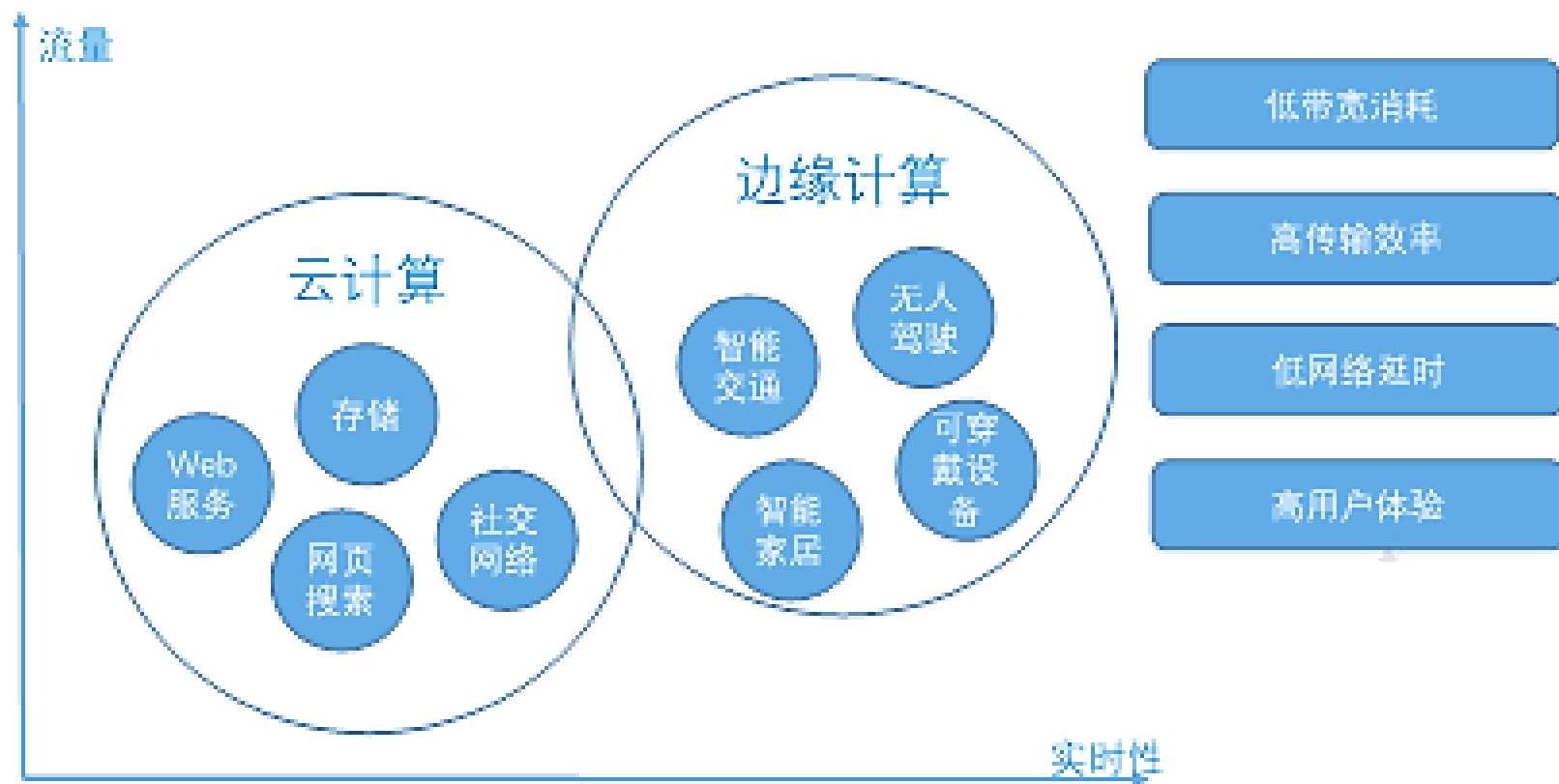
# 边缘计算的兴起

90



# 边缘计算的兴起

91



# 边缘计算的兴起

92

- 云计算把握整体，边缘计算更专注局部，边缘计算是云计算的补充和优化
- 边缘计算的特点

- ▣ 分布式和低延时计算

- 边缘计算聚焦实时、短周期数据的分析，能够更好地支撑本地业务的实时智能化处理与执行

- ▣ 效率更高

- 由于边缘计算距离用户更近，在边缘节点处实现了对数据的过滤和分析，因此效率更高

- ▣ 更加智能化

- AI+边缘计算的组合出击让边缘计算不止于计算，更多了一份智能化

- ▣ 更加节能

- 云计算和边缘计算结合，成本只有单独使用云计算的39%

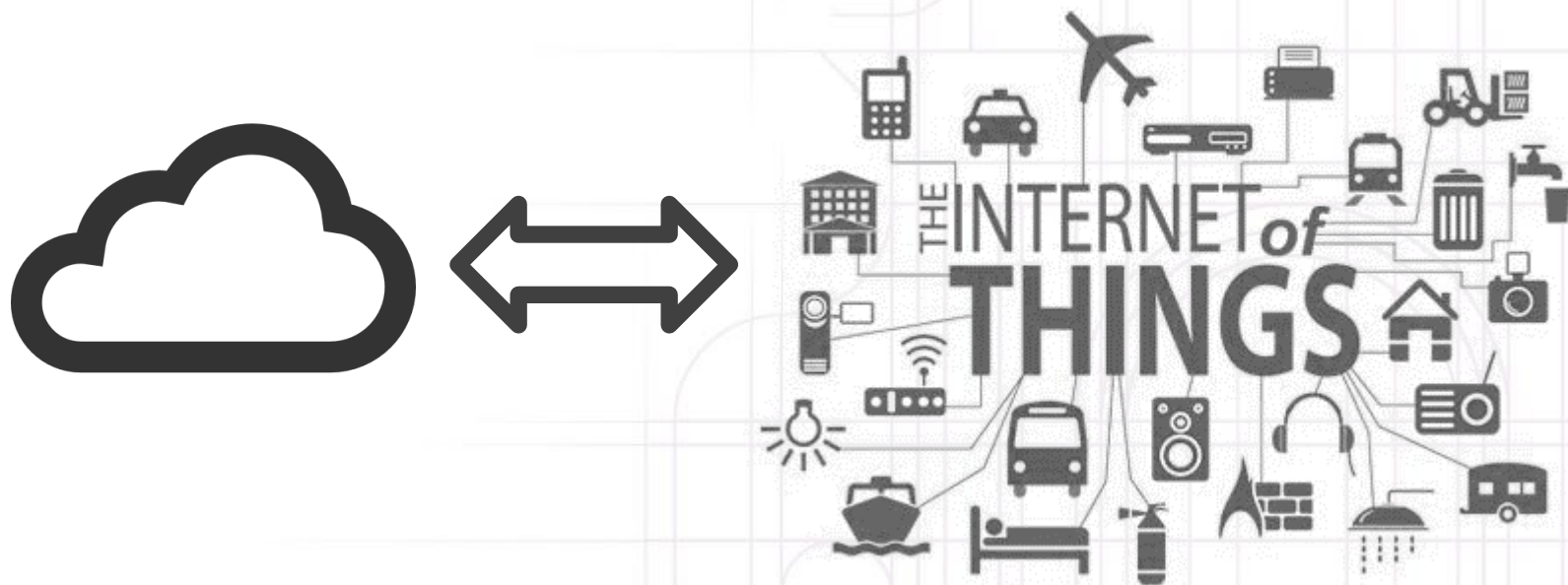
- ▣ 缓解流量压力

- 在进行云端传输时通过边缘节点进行一部分简单数据处理，进而能够设备响应时间，减少从设备到云端的数据流量

# “云物联”的展望

93

- 云计算背景下，用户可以灵活租用云计算服务、避免基础设施投资，将资金和时间用于为客户提供更好的物联网服务。
- 在云计算的强大支持下，21世纪的物联网必将更加普及和高效。



# 本节小结

94

介绍了云计算的相关内容。首先介绍了云服务的硬件，操作系统，网络等底层支持。之后讲云计算的关键：虚拟化。然后介绍其应用如云存储。最后与物联网结合，展望“云物联”。