

# 大数据处理

## 课程回顾

毛波 & 吴素贞  
厦门大学信息学院

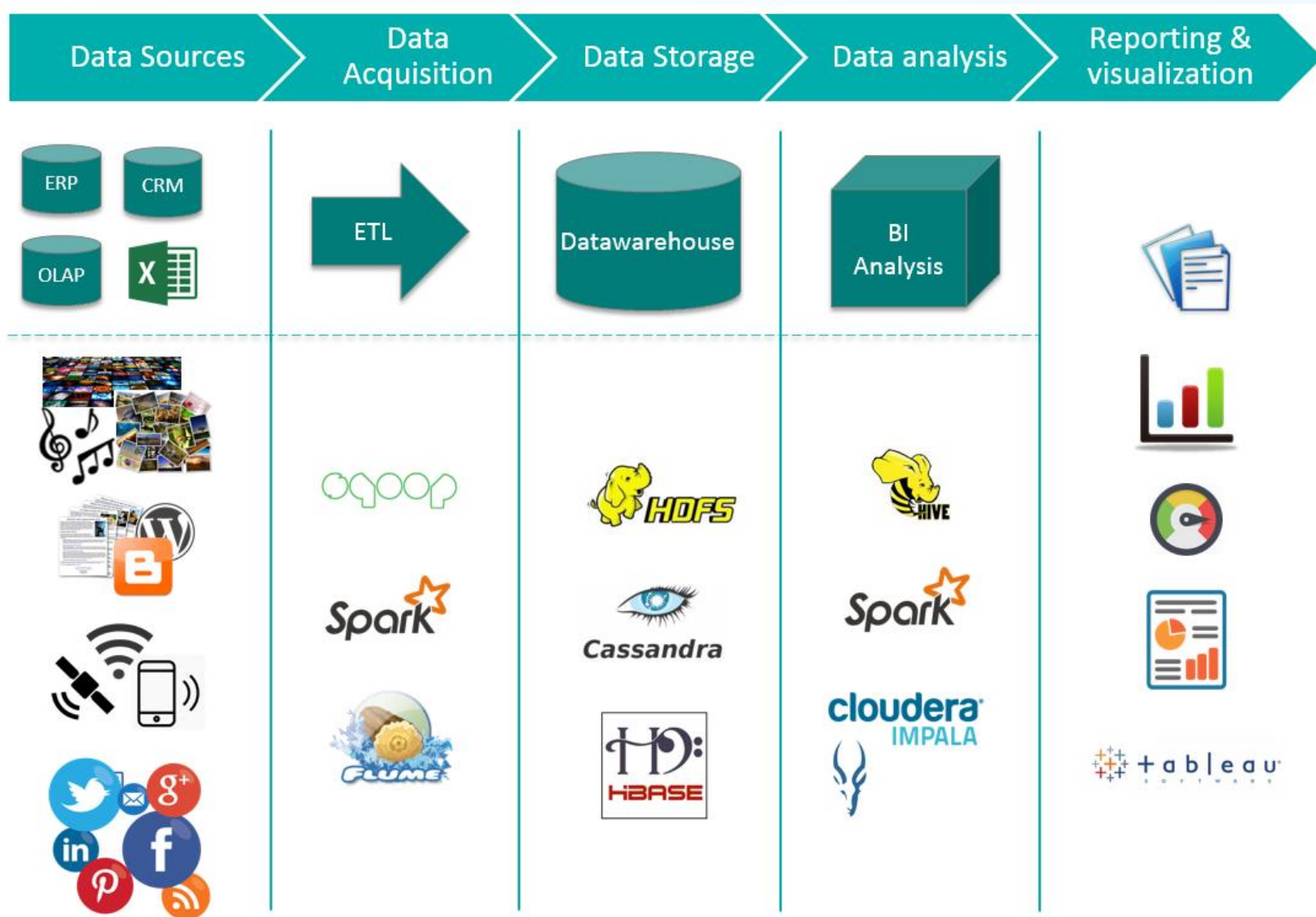
2025-5

# 大数据（Big Data）

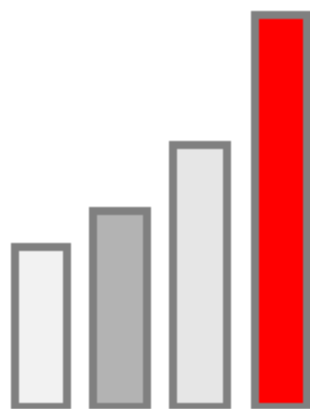
大数据是需要**新处理模式**才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的**信息资产**。（出自研究机构Gartner）



# 大数据（Big Data）生态



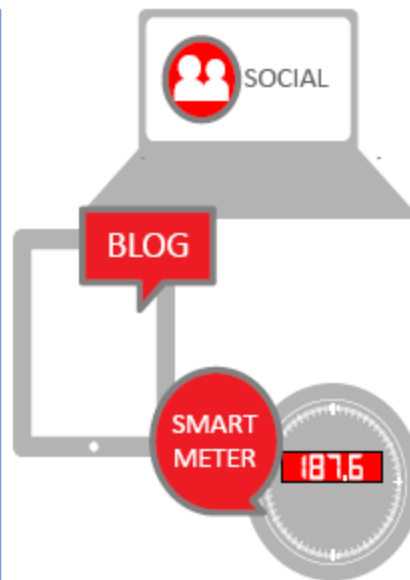
# 大数据特点（4“V”s）



VOLUME  
大量化



VELOCITY  
快速化



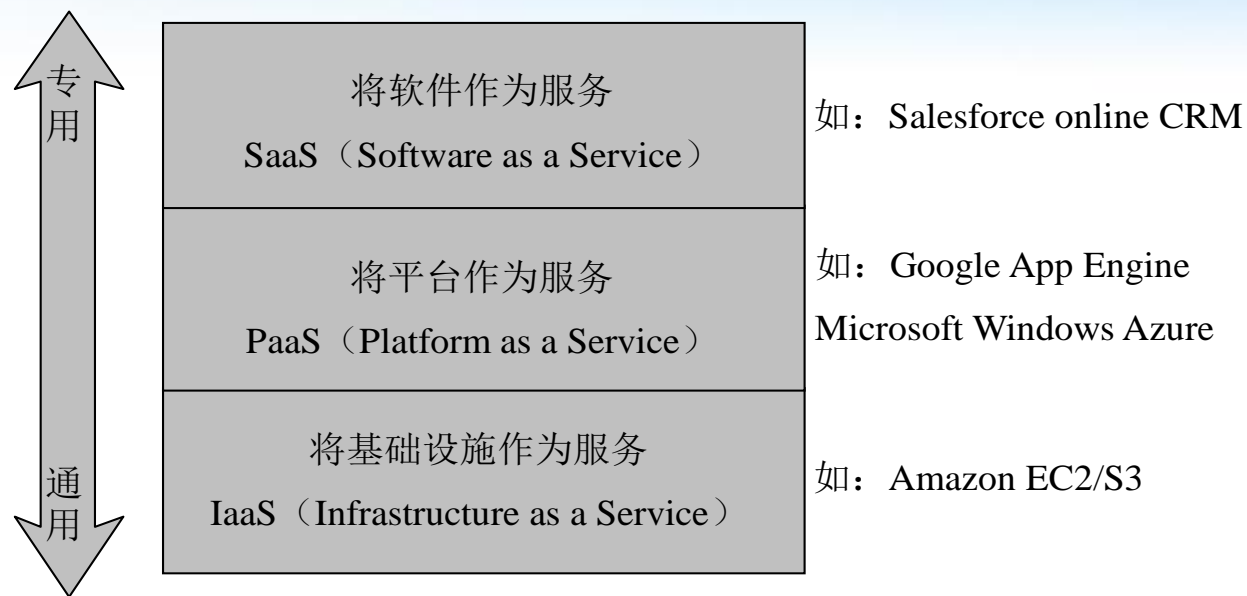
VARIETY  
多样化



VALUE

大数据不仅仅是数据的“大量化”，而是包含“快速化”、“多样化”和“价值化”等多重属性。

# 云计算的服务模型

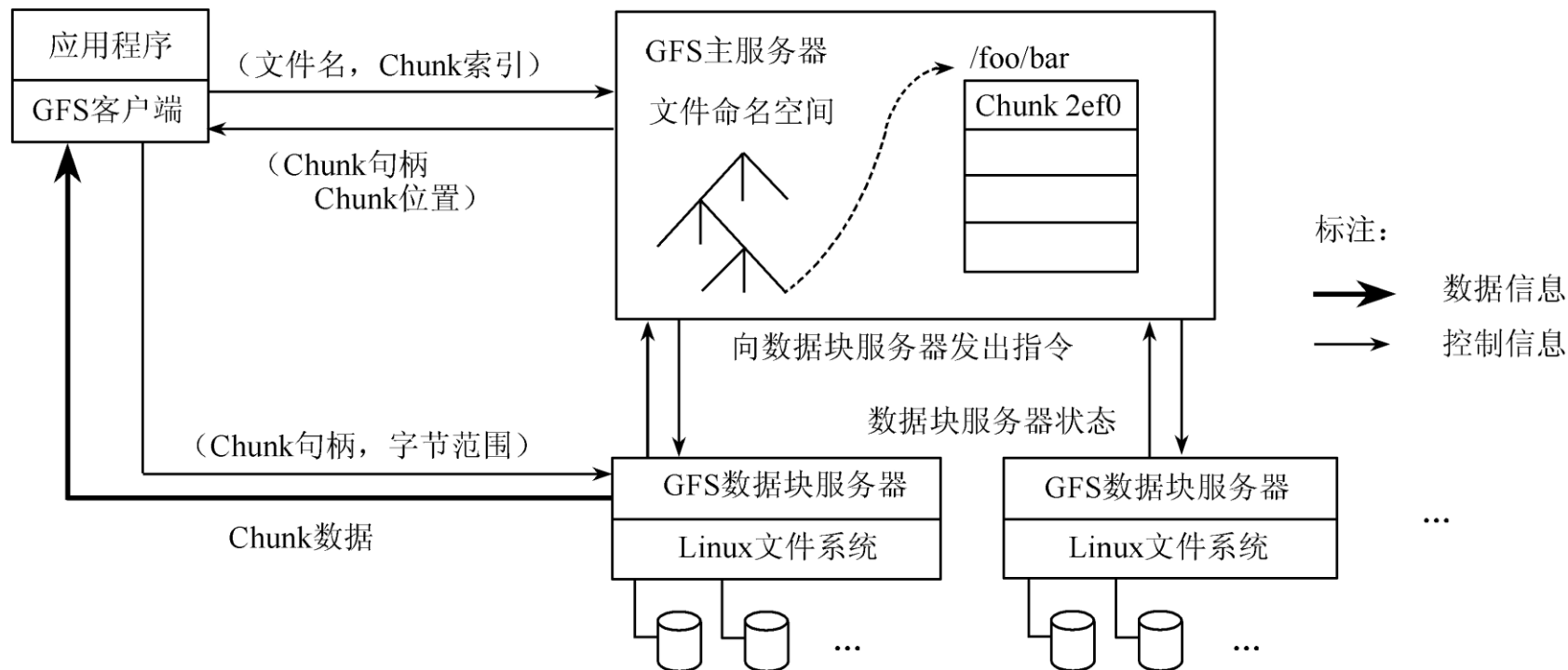


■ 将硬件设备等基础资源封装成服务供用户使用

■ 对资源的抽象层次更进一步，提供用户应用程序运行环境

■ 针对性更强，它将某些特定应用软件功能封装成服务

# Google GFS系统架构

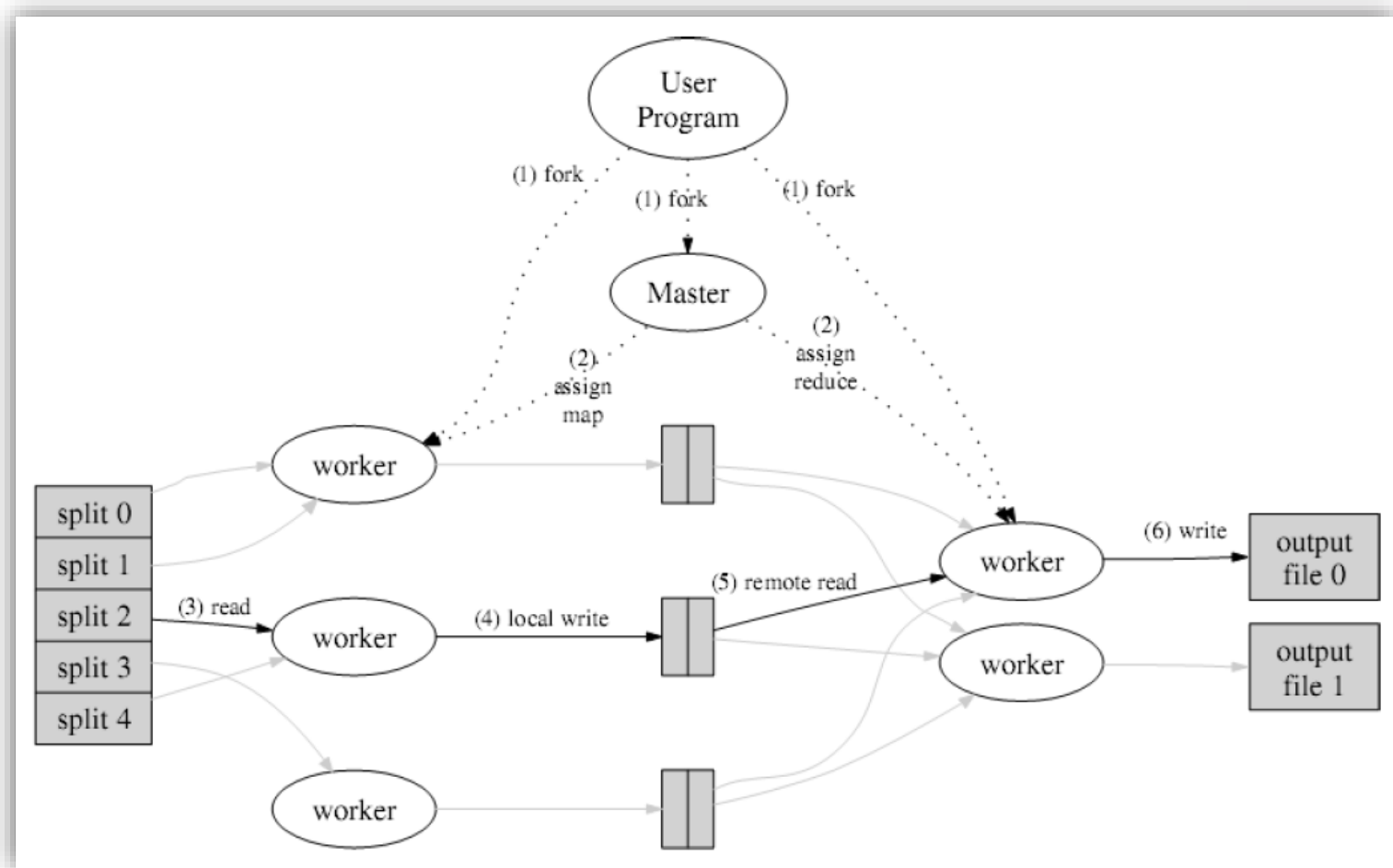


**Client (客户端)**：应用程序的**访问接口**

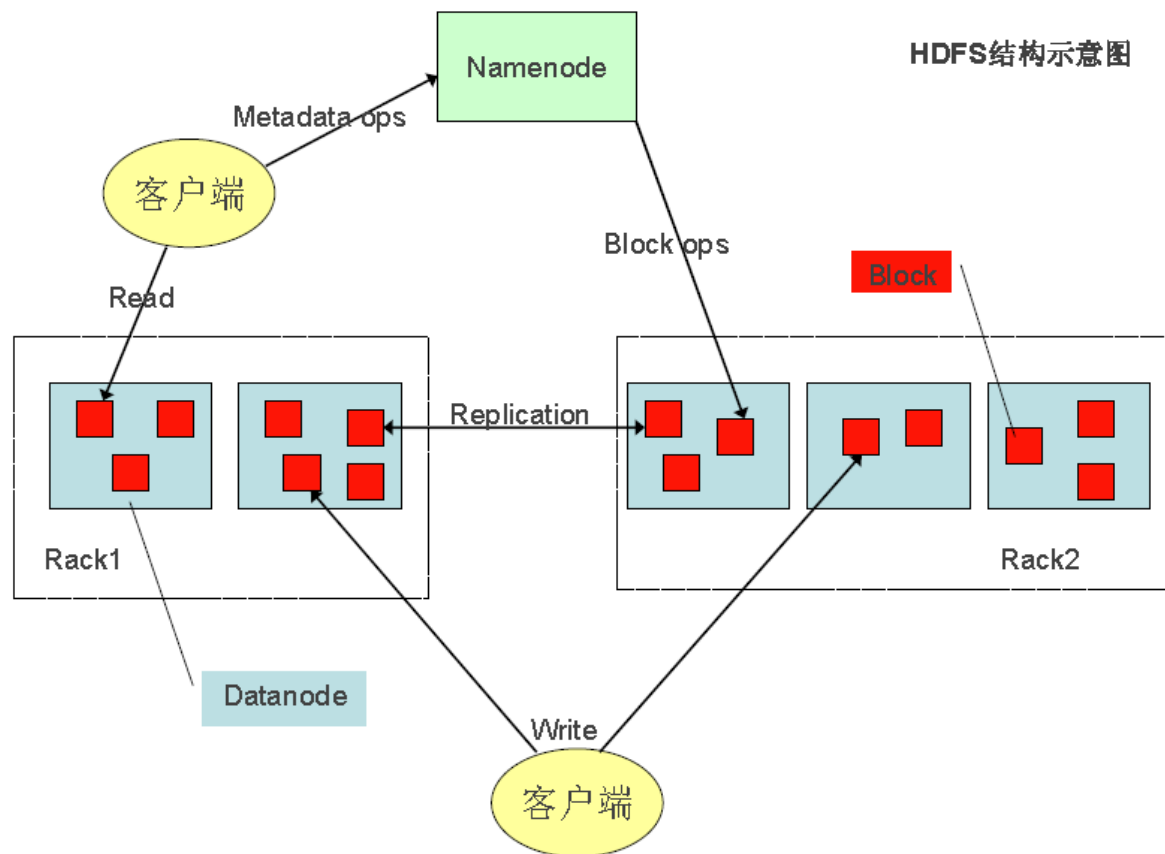
**Master (主服务器)**：**管理节点**，在逻辑上只有一个，保存系统的元数据，负责整个文件系统的管理

**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

# MapReduce架构



# Hadoop体系结构





# Hadoop VS. Google

- 技术架构的比较

- 数据结构化组件: Hbase→BigTable
- 并行计算模型: MapReduce→MapReduce
- 分布式文件系统: HDFS→GFS

*Hadoop 云计算应用*

*HBase*

*MapReduce*

*HDFS*

*Google 云计算应用*

*BigTable*

*MapReduce*

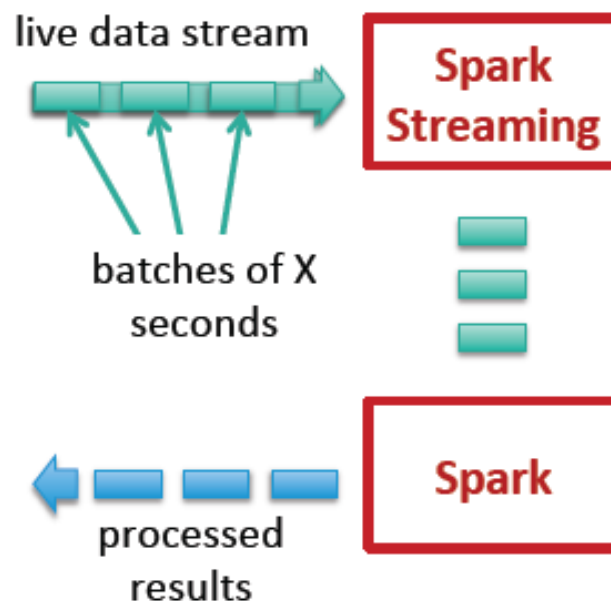
*GFS*

*Chubby*

# Spark实时处理技术

Run a streaming computation as a **series of very small, deterministic batch jobs**

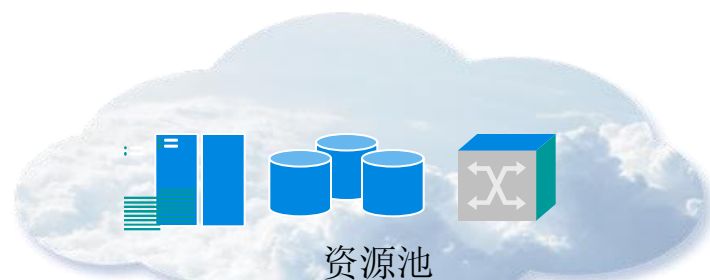
- Chop up the live stream into batches of X seconds
- Spark treats each batch of data as RDDs and processes them using RDD operations
- Finally, the processed results of the RDD operations are returned in batches



# 虚拟化技术的出现

虚拟化技术将物理资源转化为便于切分的资源池，符合云计算的基本条件；

虚拟化给资源以动态调配的能力，符合云计算按需分配的要求；



Amazon采用虚拟化技术提供云计算平台，取得了商业上的成功，虚拟化技术成为云计算的基石；

2006

1960's

IBM推出虚拟化技术，提高了昂贵的大型机的利用率；

1999

VMware公司解决了X86虚拟化问题，推出了X86平台的虚拟机软件，使虚拟化技术开始走向普通用户。

2003

开源虚拟化技术Xen推出，使虚拟化技术的研究和应用更加普及；

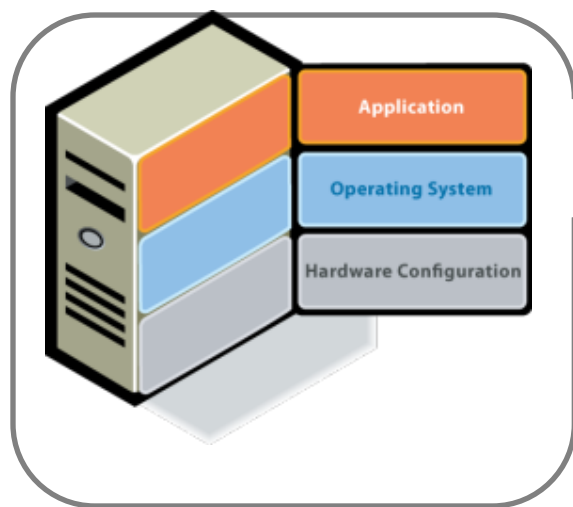
2005

Intel和AMD推出支持虚拟化技术的处理器和芯片组，实现了硬件辅助虚拟化技术；

# 虚拟化技术

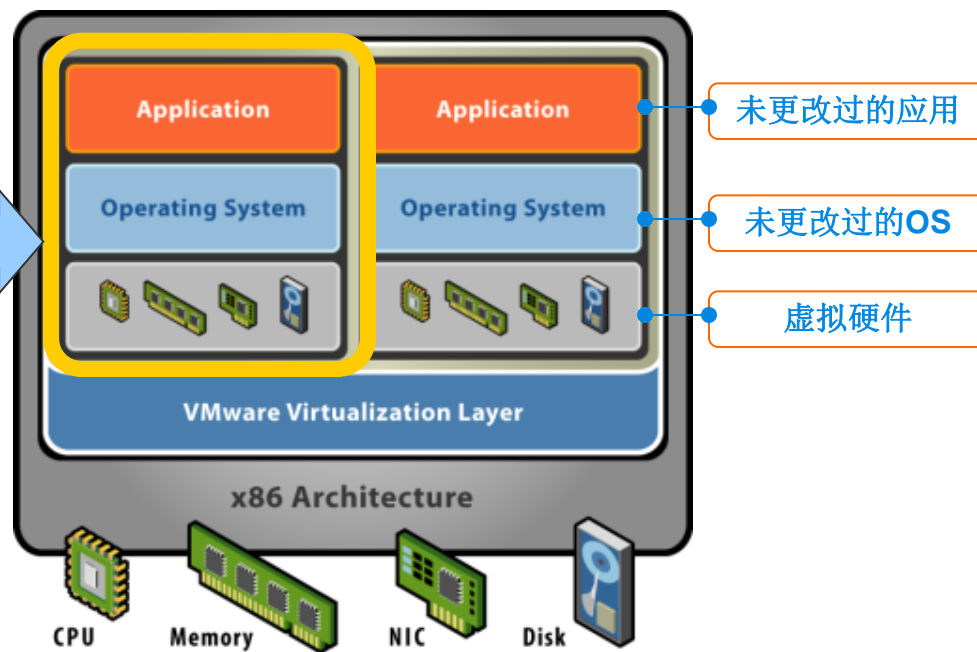
虚拟化将硬件、操作系统和应用程序一同**封装**一个可迁移的虚拟机档案文件中

虚拟化前



- 软件必须与硬件相结合
- 每台机器只能运行单一的操作系统
- 每个操作系统有一个或多个应用程序负载（通常只有一个）

虚拟化后



- 增加虚拟化层
- 裸金属架构
- 每台机器上有多个操作系统和多个应用负载

# 虚拟技术的四大特性

## 分区



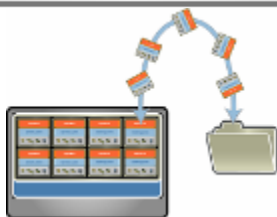
在单一物理服务器上同时运行多个虚拟机

## 隔离



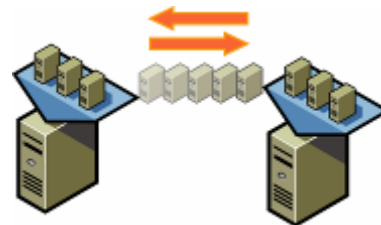
在同一服务器上的虚拟机之间相互隔离

## 封装



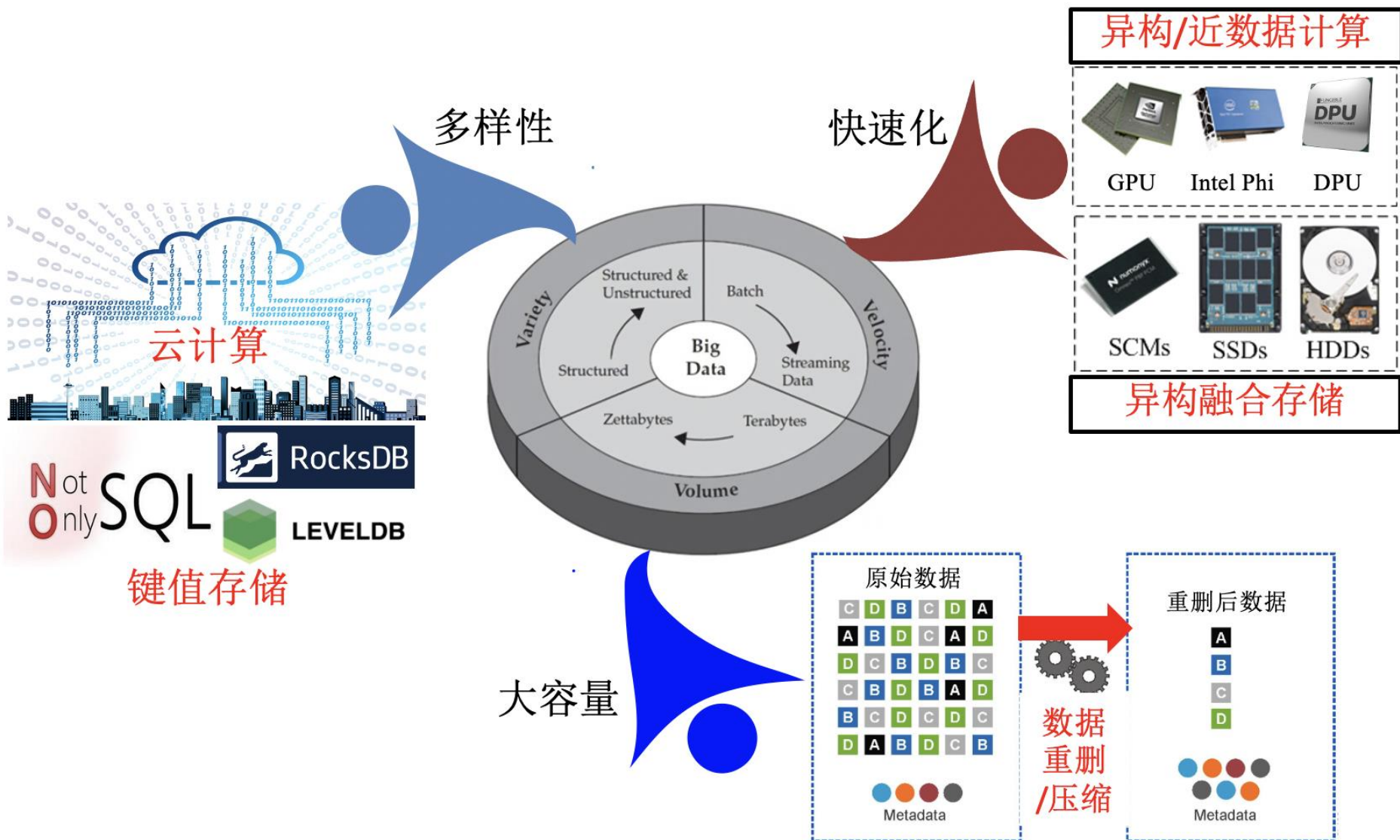
整个虚拟机保存在文件中，而且可以通过移动和复制文件的方式来移动和复制该虚拟机

## 相对于硬件独立

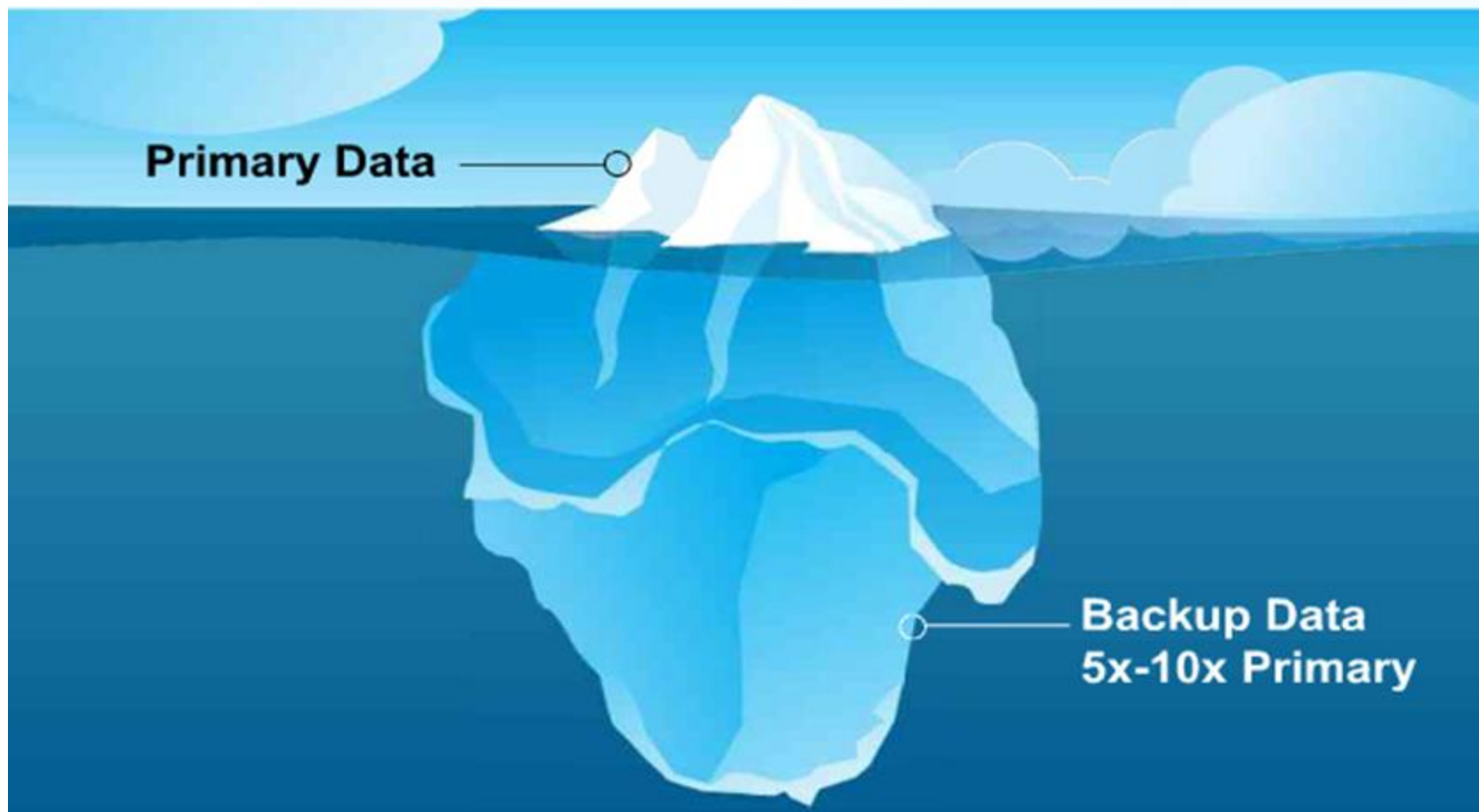


无需修改即可在任何服务器上运行虚拟机

# 专题：应对大数据3V挑战的技术方案

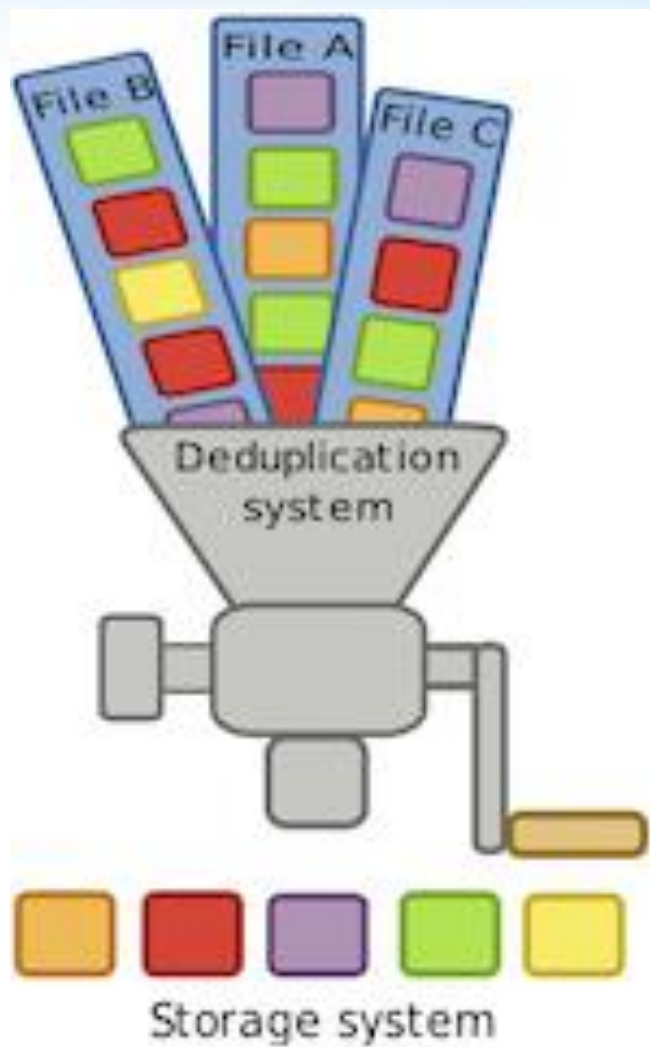


# 重复数据删除技术的出现





# 重复数据删除技术





# 重复数据删除技术的优缺点

- **优势：**
  - 控制数据增长，提高存储利用率
  - 提高网络带宽利用率，减少备份时间
  - 降低成本和能耗
- **劣势：**
  - 需要额外内存和处理资源
  - 降低数据的可靠性
  - 增加了数据恢复的开销

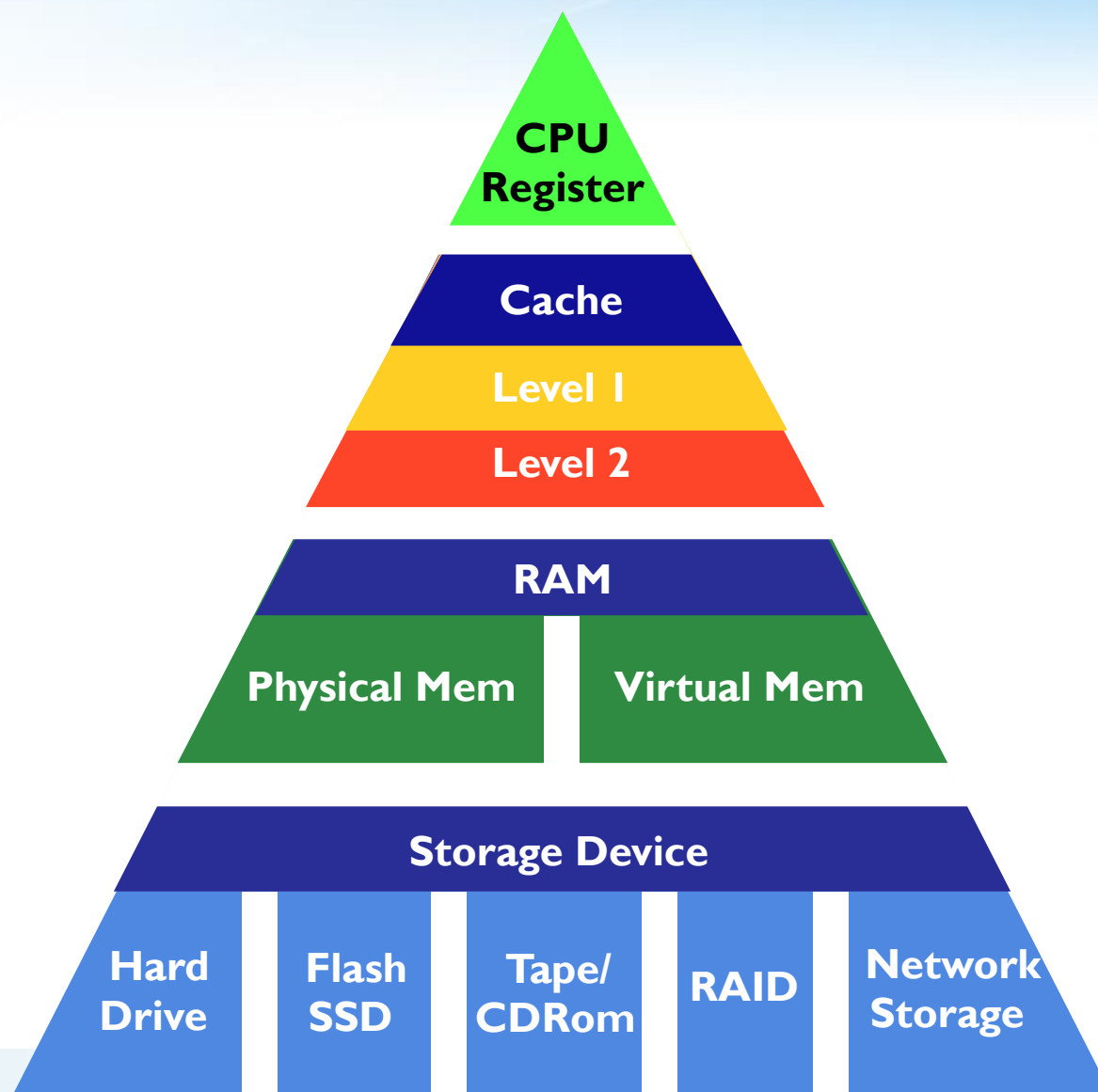
# 固态硬盘等新型存储技术



HDD



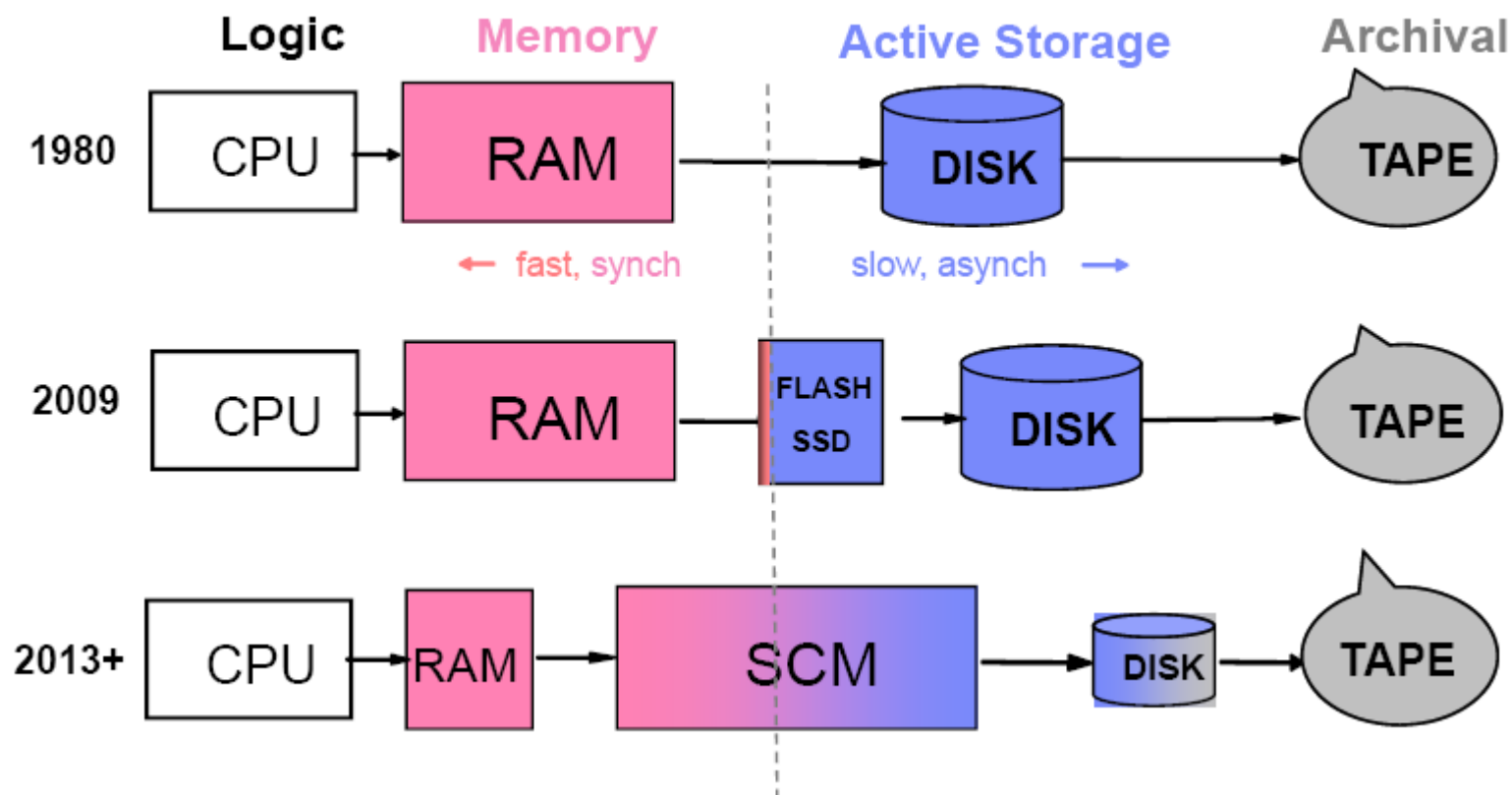
# 固态硬盘等新型存储技术



# RAID比较

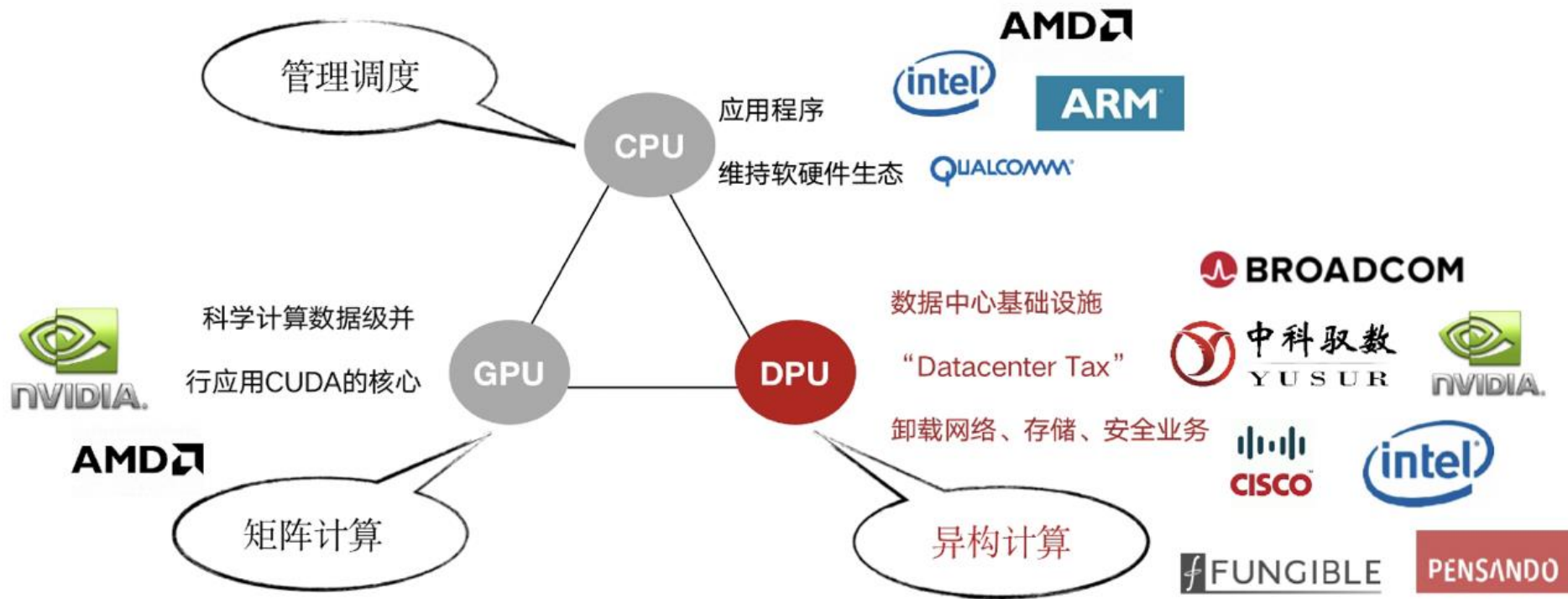
RAID	Min Disks	Storage Efficiency %	Cost	Read Performance	Write Performance
0	2	100	Low	Very good	Very good
1	2	50	High	Better than a single disk	Slower than a single disk
4	3	$(n-1)*100/n$	Moderate	Good for reads	Poor for small random writes
5	3	$(n-1)*100/n$	Moderate	Good for reads	Poor for small random writes
6	4	$(n-2)*100/n$	Moderate	Good for reads	Poor for small random writes

# 固态硬盘等新型存储技术



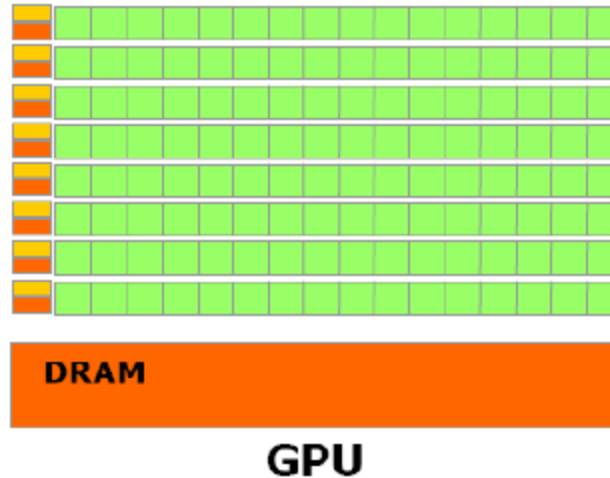
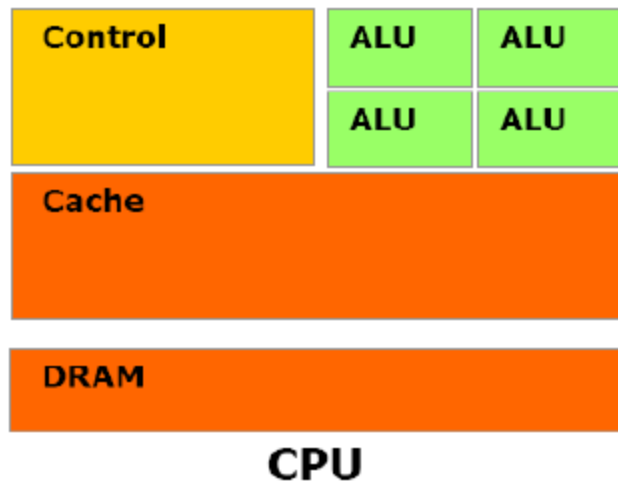
# X PU 的异构计算

(from Nvidia@May, 2020)



# GPU

- CPU: 更多资源用于缓存和逻辑控制
- GPU: 更多资源用于计算, 适用于高并行性、大规模数据密集型、可预测的计算模式。



# SQL vs NoSQL



SQL:  
结构化存储, 固定Schema  
索引  
标准化查询语言  
ACID  
扩展性弱

NoSQL:  
Schema不固定, 可以动态改变  
没有固定查询语言  
BASE (Basically Available, Soft State, Eventually Consistency)  
最终一致性  
可以扩展到很大规模  
高容错性

