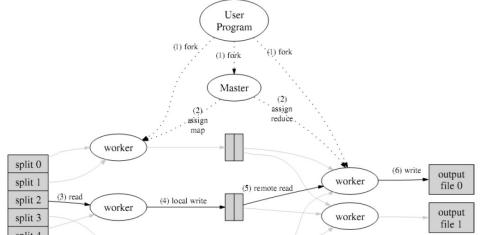
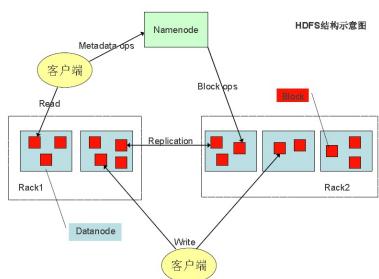


MapReduce架构



Hadoop体系结构



Hadoop VS. Google

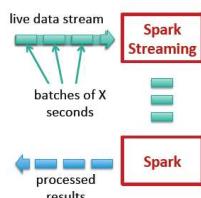
- 技术架构的比较
 - 数据结构化管理组件: Hbase→BigTable
 - 并行计算模型: MapReduce→MapReduce
 - 分布式文件系统: HDFS→GFS



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；



2006
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础；

1960's

1999

2003

2005

IBM推出虚拟化技术，提高了昂贵的大型机的利用率；

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；

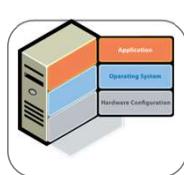
开源虚拟化技术Xen推出，使虚拟化技术的研究和应用更加普及；

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

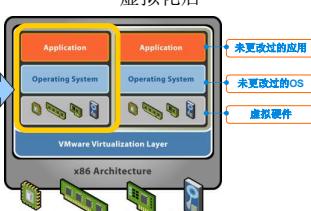
虚拟化将硬件、操作系统和应用程序一同封装一个可迁移的虚拟机档案文件中

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

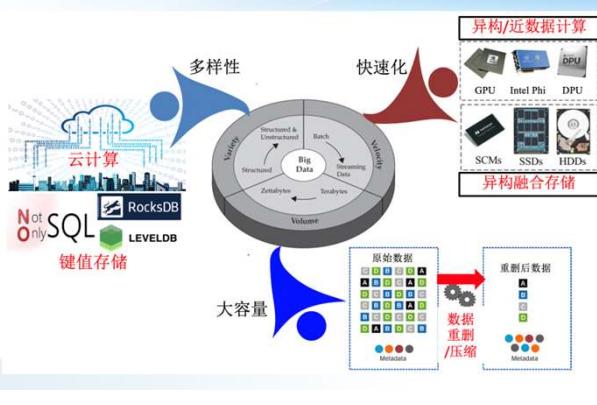


- 增加虚拟化层
- 裸金属架构
- 每台机器上有多个操作系统和多个应用负载

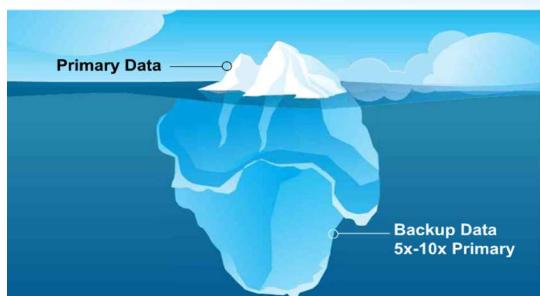
虚拟技术的四大特性



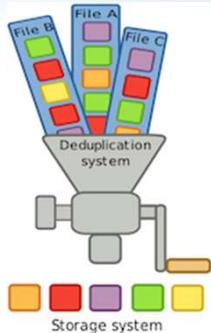
专题：应对大数据3V挑战的技术方案



重复数据删除技术的出现



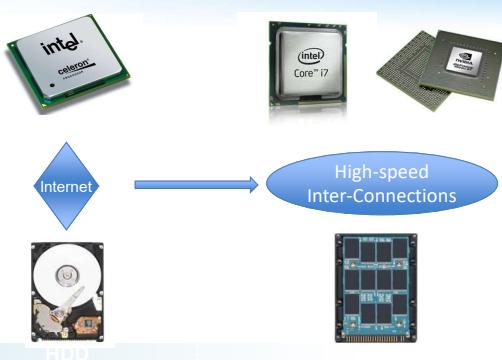
重复数据删除技术

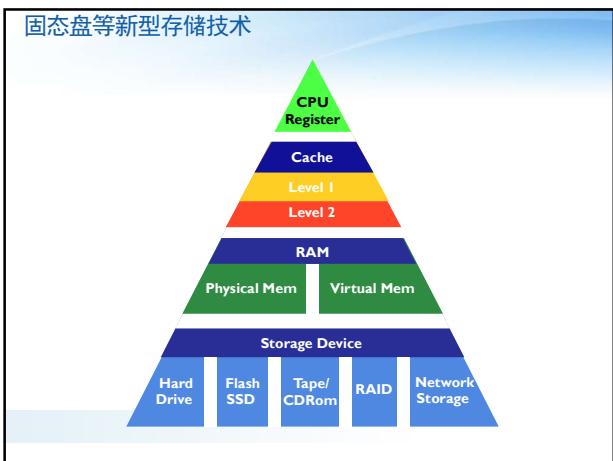


重复数据删除技术的优缺点

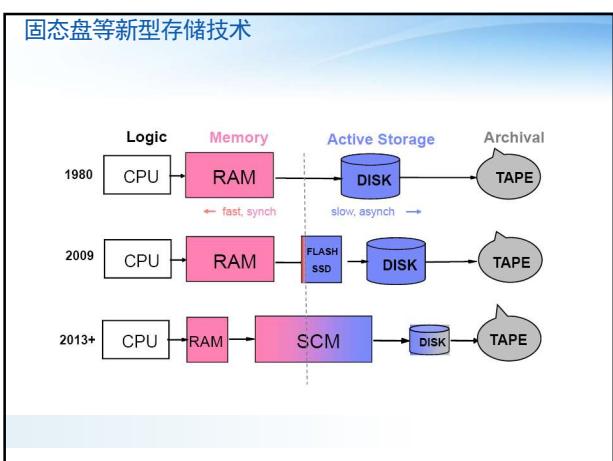
- **优势:**
 - 控制数据增长，提升存储利用率
 - 提高网络带宽利用率，减少备份时间
 - 降低成本和能耗
- **劣势:**
 - 需要额外内存和处理资源
 - 降低数据的可靠性
 - 增加了数据恢复的开销

固态盘等新型存储技术

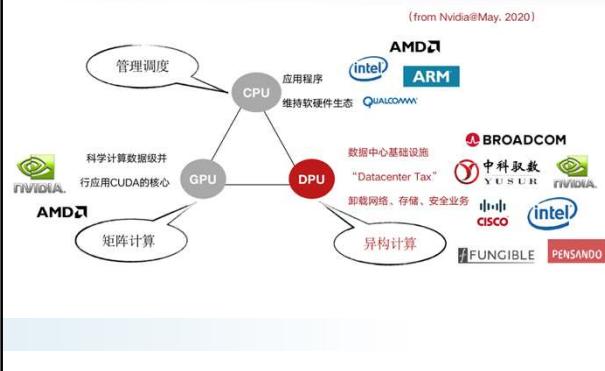




RAID	Min Disks	Storage Efficiency %	Cost	Read Performance	Write Performance
0	2	100	Low	Very good	Very good
1	2	50	High	Better than a single disk	Slower than a single disk
4	3	$(n-1)*100/n$	Moderate	Good for reads	Poor for small random writes
5	3	$(n-1)*100/n$	Moderate	Good for reads	Poor for small random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

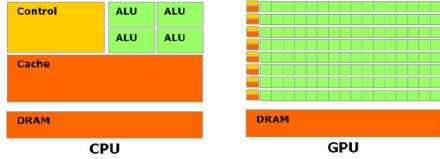


X PU 的异构计算



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

