网易云硬盘系统简介

网易杭州研究院后台技术中心　吴东
新浪微博：@dong_wu

DTCC

2015中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2015

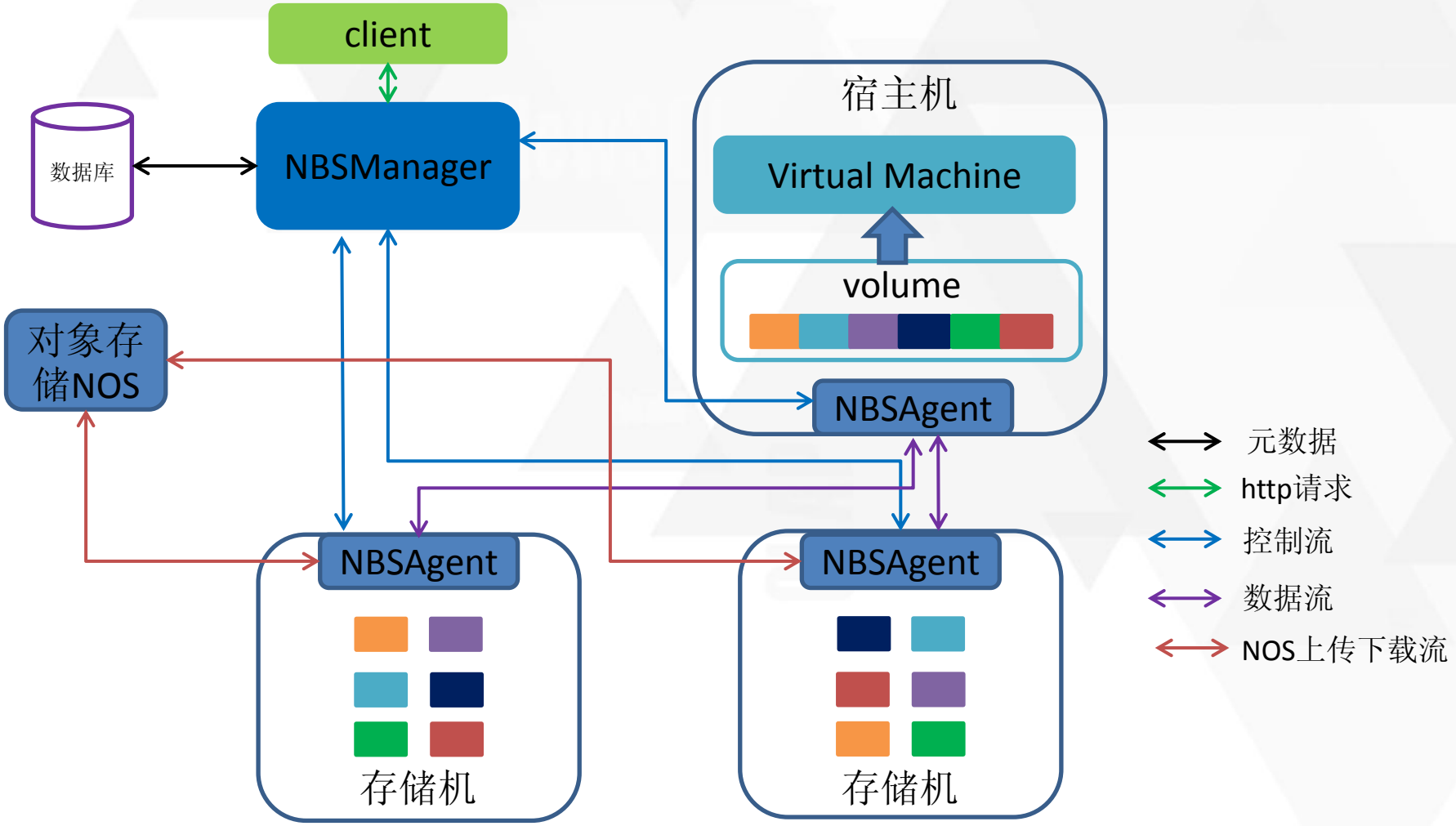
大数据技术探索和价值发现

网易云硬盘（NBS）

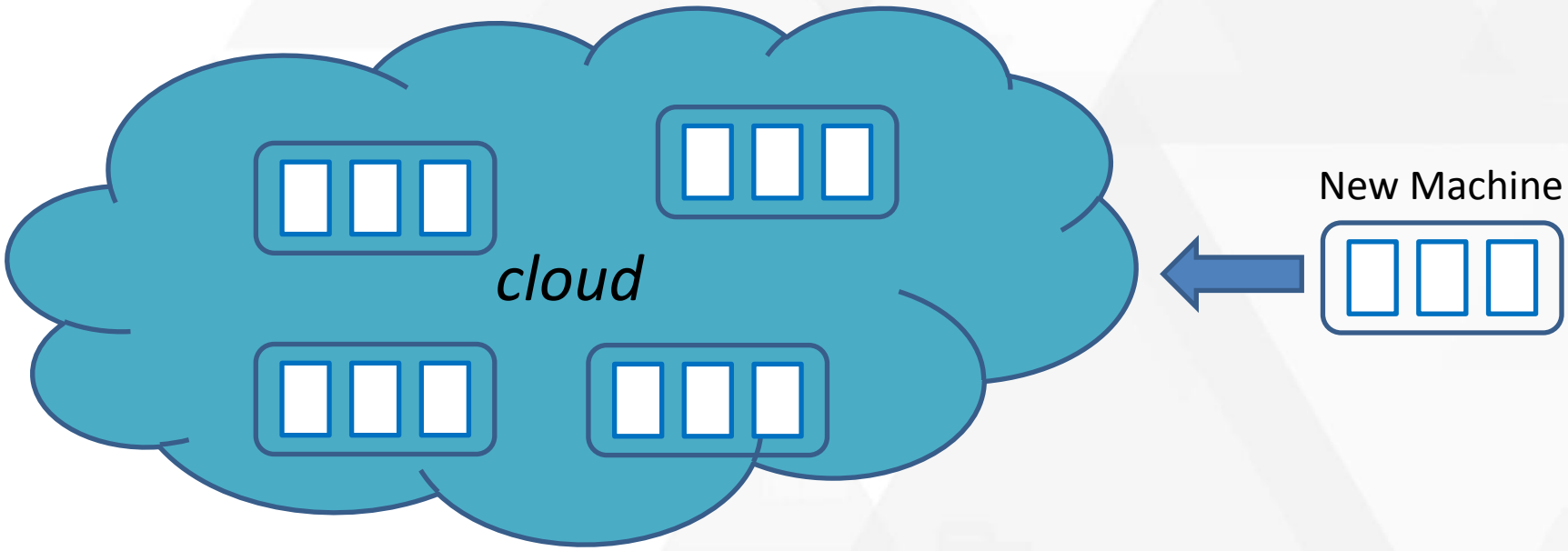
- 网易云硬盘是在传统块存储技术的基础上，基于廉价设备构建可扩展、高可靠、高可用、有QOS保证、易运维的存储系统。
- 公司的关系型数据库服务（RDS）及云搜索（NCS）都是建构在云硬盘之上的，另外还承接了公司多个重要产品的数据存储服务。

云硬盘主要特点

云硬盘系统架构

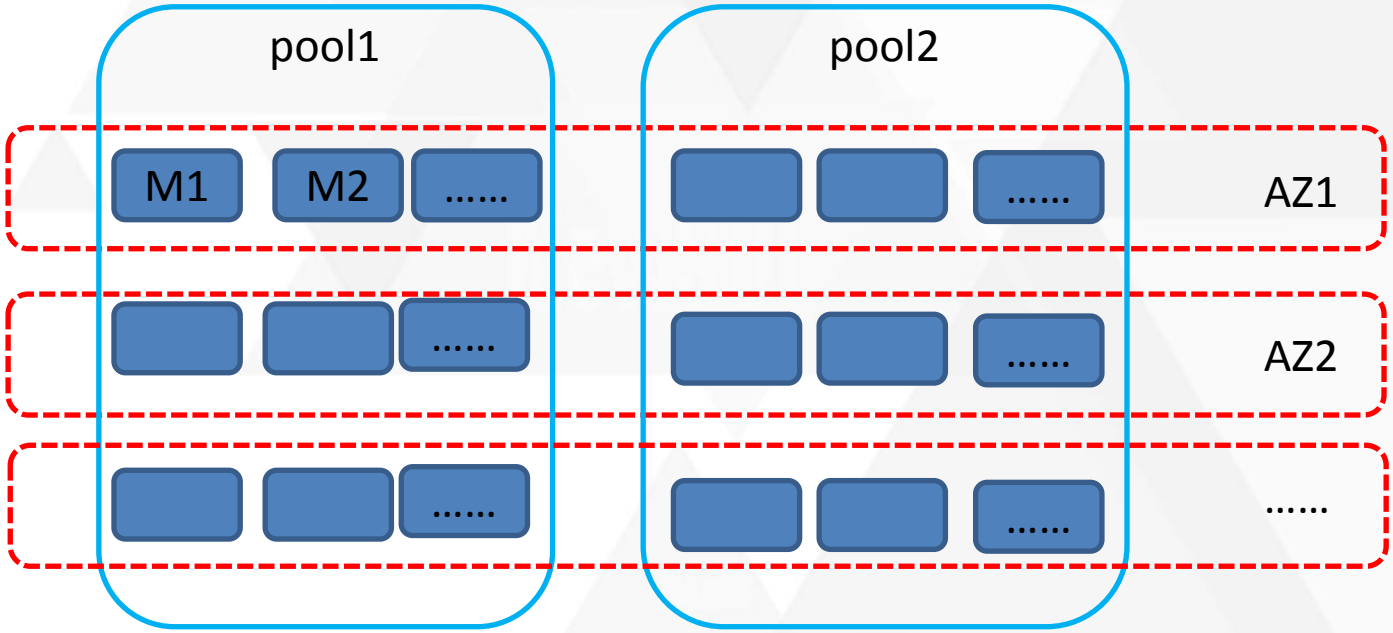

client
数据库
NBSManager
宿主机
Virtual Machine
volume
对象存储NOS
NBSAgent
NBSAgent
NBSAgent
存储机
存储机
元数据
http请求
控制流
数据流
NOS上传下载流

弹性扩展



- 云硬盘弹性扩展：每块硬盘空间大小从1G到3TB，单台云主机上可以挂载多个云硬盘，使得其存储容量扩展到几十TB，并且单块云硬盘支持扩容，可以随时扩容到更大的容量；
- 集群scale out：在整个集群容量负载达到瓶颈时，可以动态地添加新机器，新机器加入后立即可用，并且不会有老机器到新机器的数据拷贝；

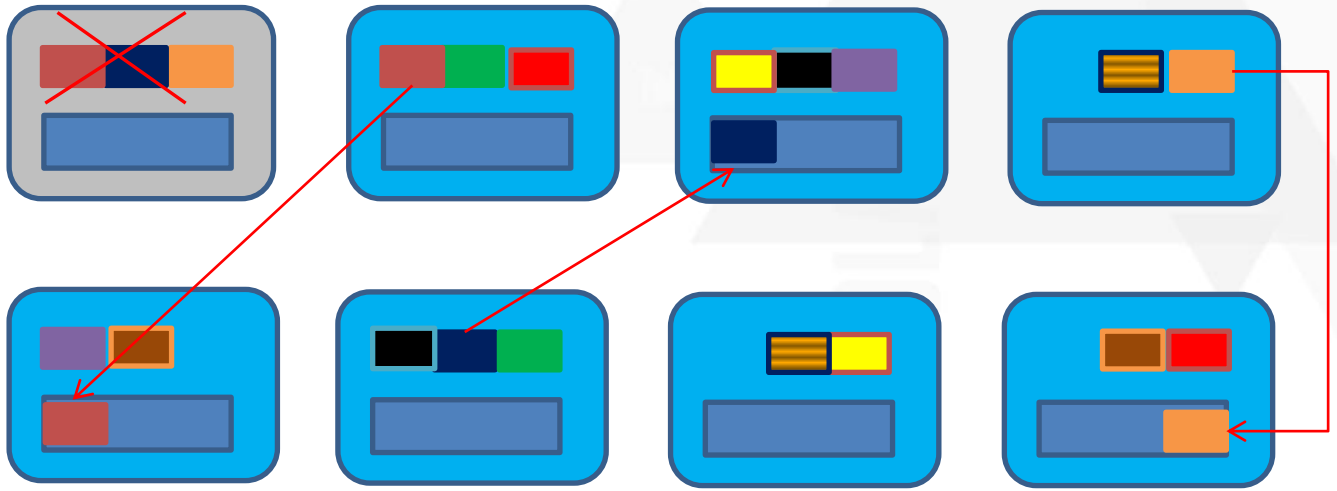
高可靠



- 可用域（Availability Zone）：AZ是一个逻辑概念，一台物理机只属于一个AZ，划分AZ是为了提高容灾性和隔离服务，以及方便运维；
- 存储池（Pool）：存储池是用来做IO隔离用的，一块物理硬盘只属于一个存储池，不同用户可以使用不同的存储池，存储池间的IO相互独立，这样就能避免用户间IO的影响；
- 多副本分布：一个云硬盘由多个chunk组成，每个chunk是多副本的，分布在不同AZ下机器的物理硬盘上；

高可靠

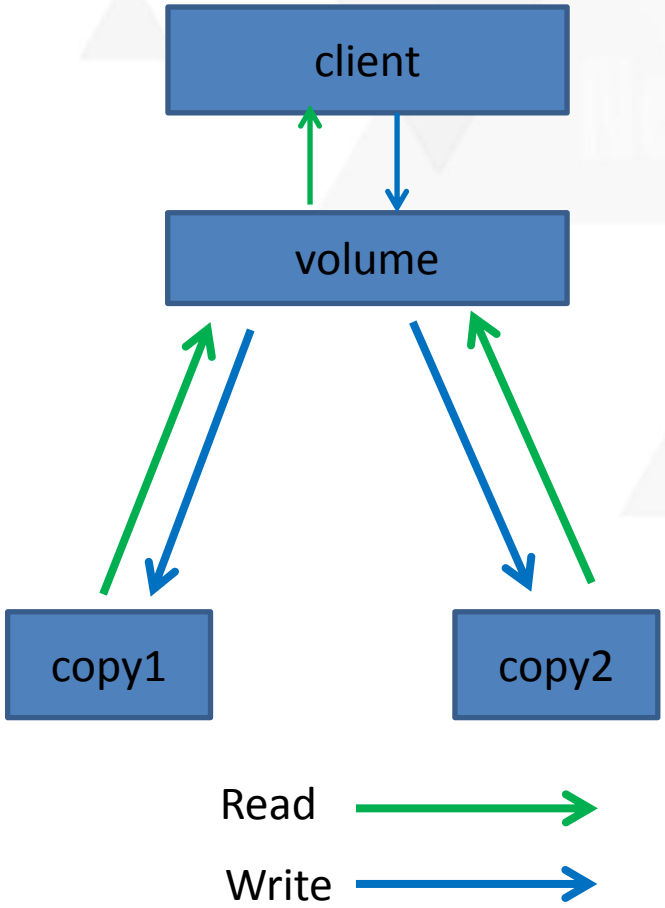
- 故障自动检测和自动恢复：心跳、卷、磁盘、机器等状态的自动汇报与检测机制，故障之后还能进行自动恢复；
- 并行恢复：一块物理盘坏掉之后，可用由分散到多台机器上的多块盘进行并行的恢复；机器坏了之后，也能由分散到多台机器的副本进行恢复；



故障容错恢复机制

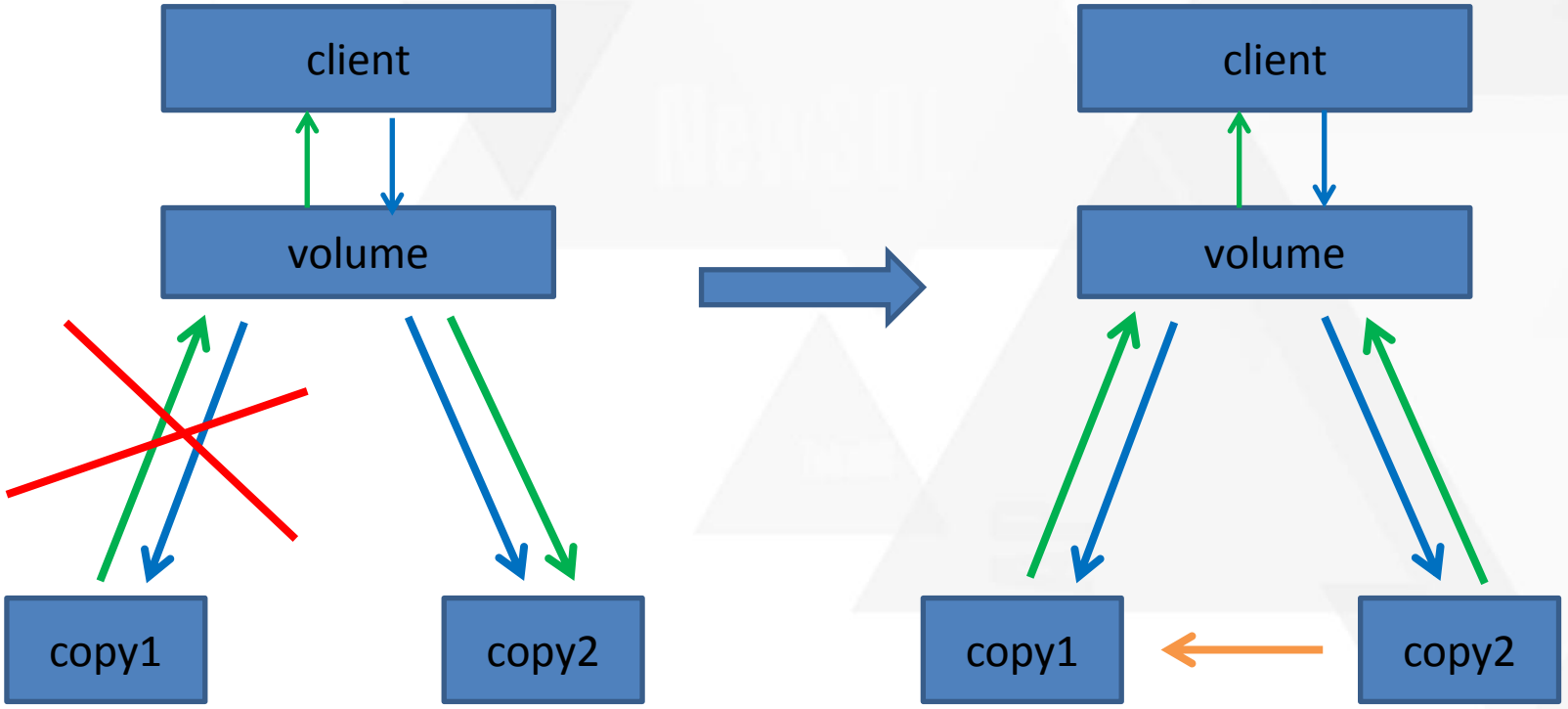
- 故障发现
  - 检测+汇报机制
- 故障分类与处理
  - 机器故障
    - 机器重启，数据增量同步
    - 机器坏，硬盘插到备机上，数据增量同步；
    - 机器坏，无备机情况，数据全量同步
  - 网络故障
    - 网络瞬断，容忍10s的io hang时间，超时后认为副本降级
    - 长时间中断，根据影响时间确定是否进行全量同步
  - 硬盘故障
    - 落到这块盘上的副本都坏了，新分配副本进行全量同步

强一致性



- 并行写多副本
- 从一个副本读

强一致性



Read

Write

Sync

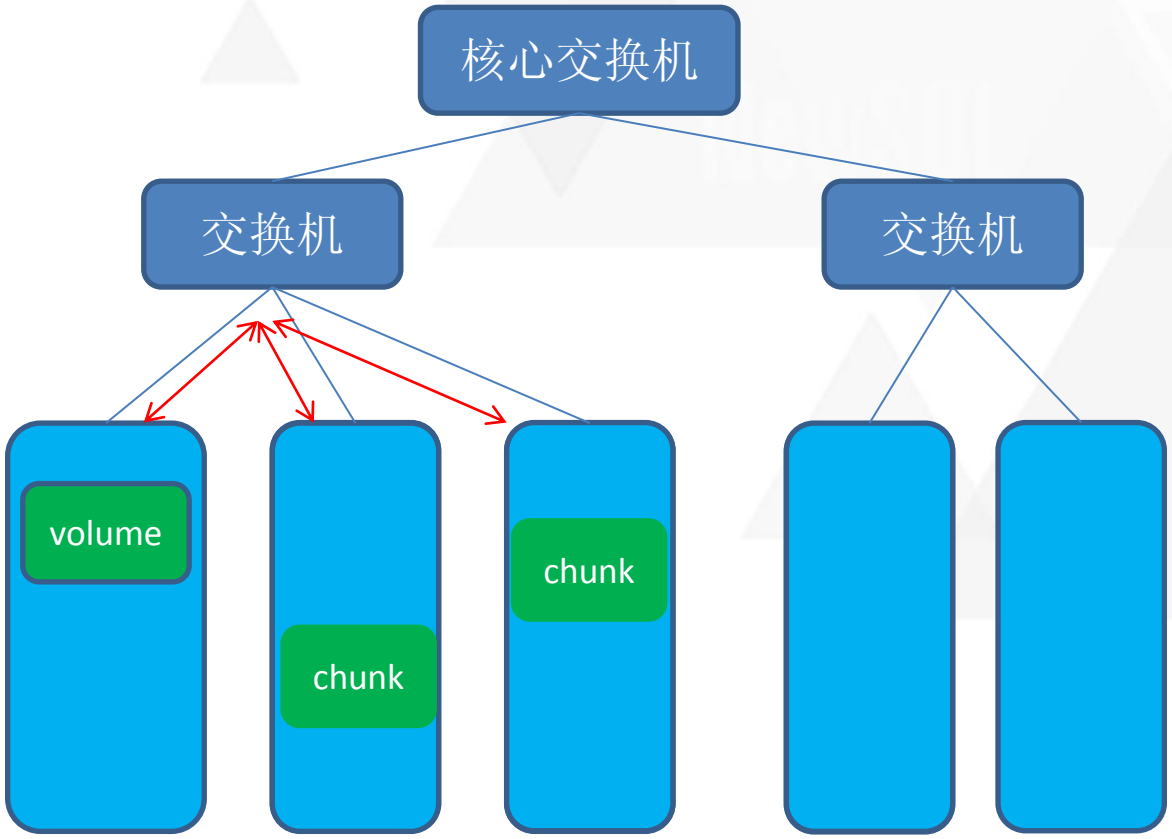
- IO过程中有副本异常，会进行数据同步

高性能

云硬盘（2副本为例）

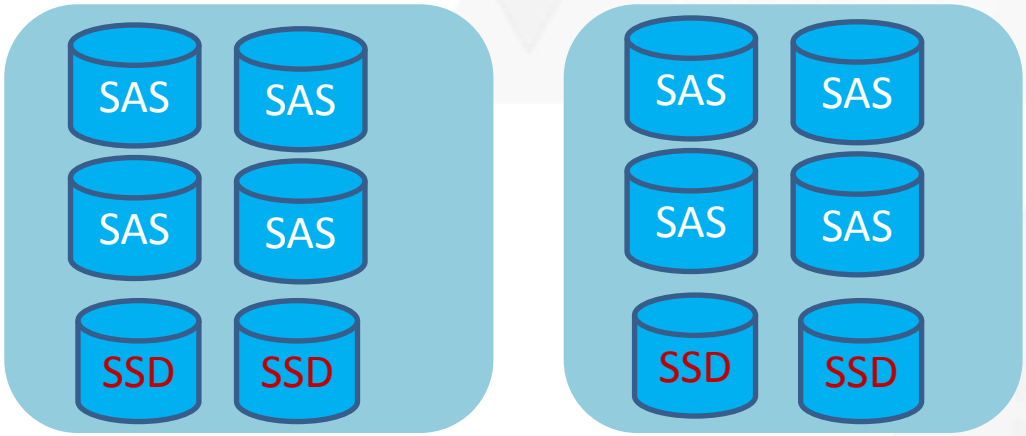
- 数据分片（chunk）：云硬盘采用分片的机制，一块云硬盘被切成多片，不同分片分布到不同的磁盘上；
- 高并发度：多个chunk分布到不同的机器和磁盘上，并发IO提高性能；
- RAID卡写缓存：存储服务器带RAID卡写缓存，提高写性能；

高性能



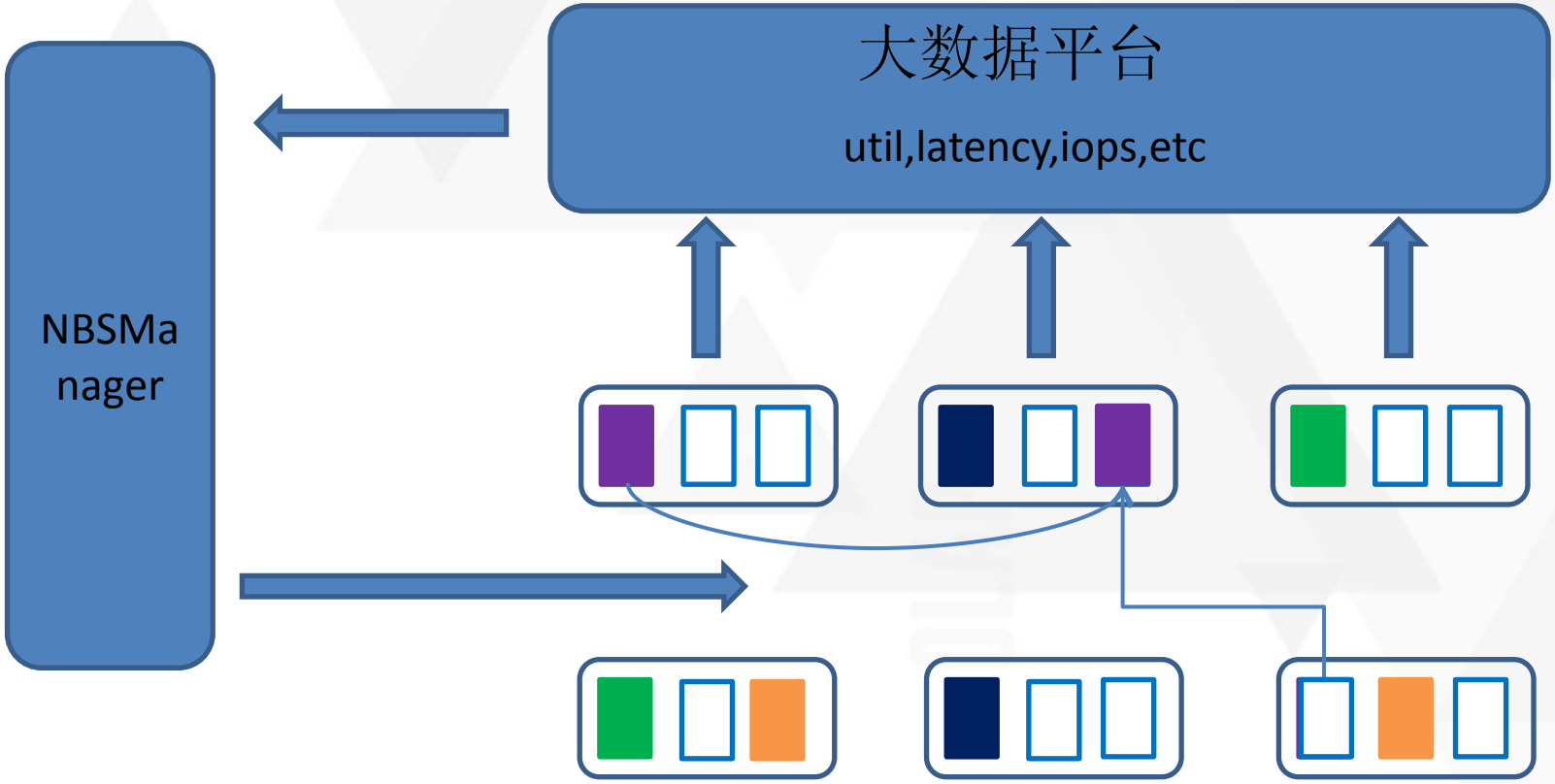
- 集群（Cluster）
  - 两个堆叠的接入层交互机称为一个集群
- Cluster就近原则
  - 云硬盘就近分配到与云主机相同的集群中，让数据流尽量在同一个交换机内交互

高性能

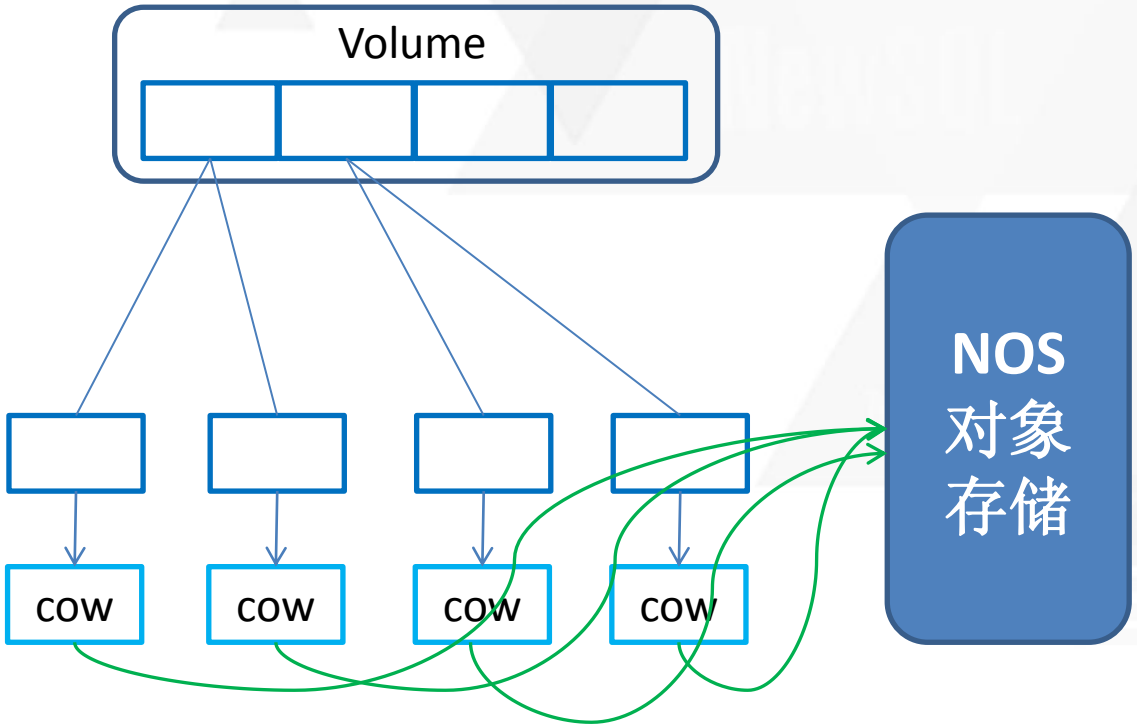


- 多介质支持
  - 机械盘和SSD混合部署，可以使用SSD型盘来满足更高性能的需求

动态负载均衡



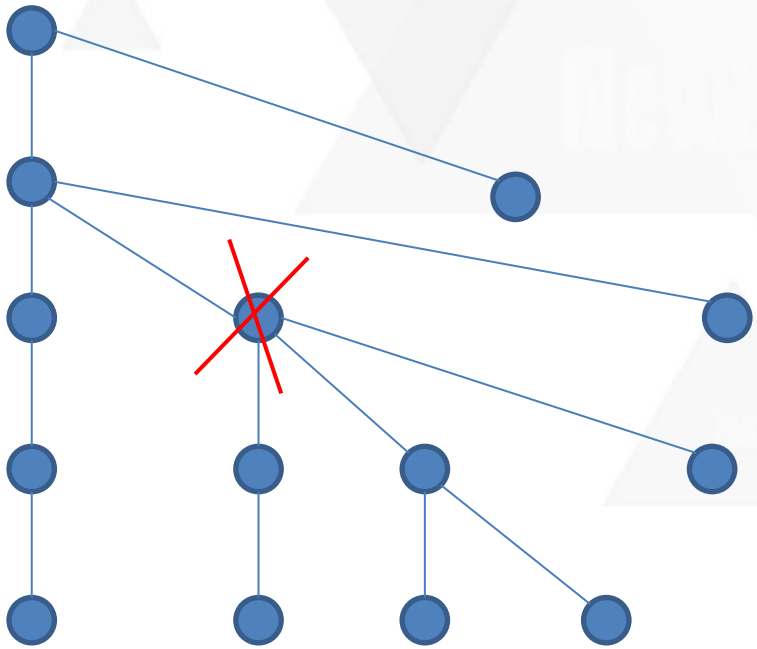
快照技术



快照特点

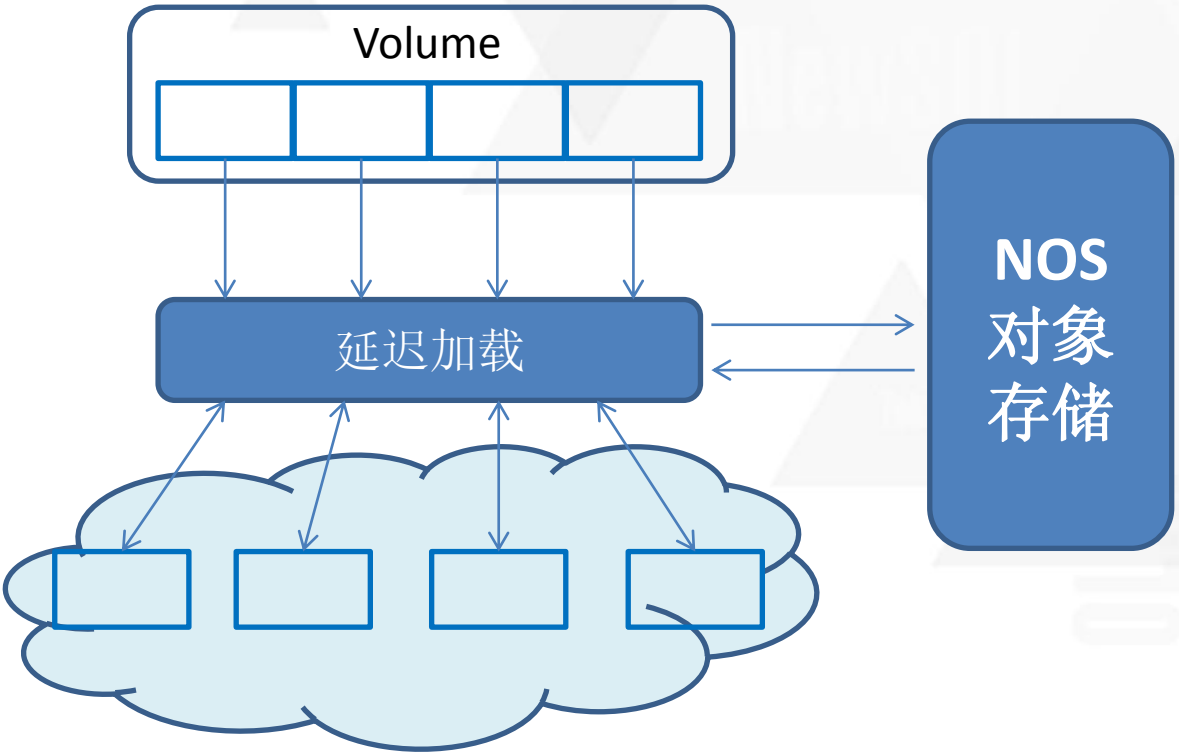
- 在线快照
- 并发上传
- 增量快照
- 延迟加载

全量与增量快照



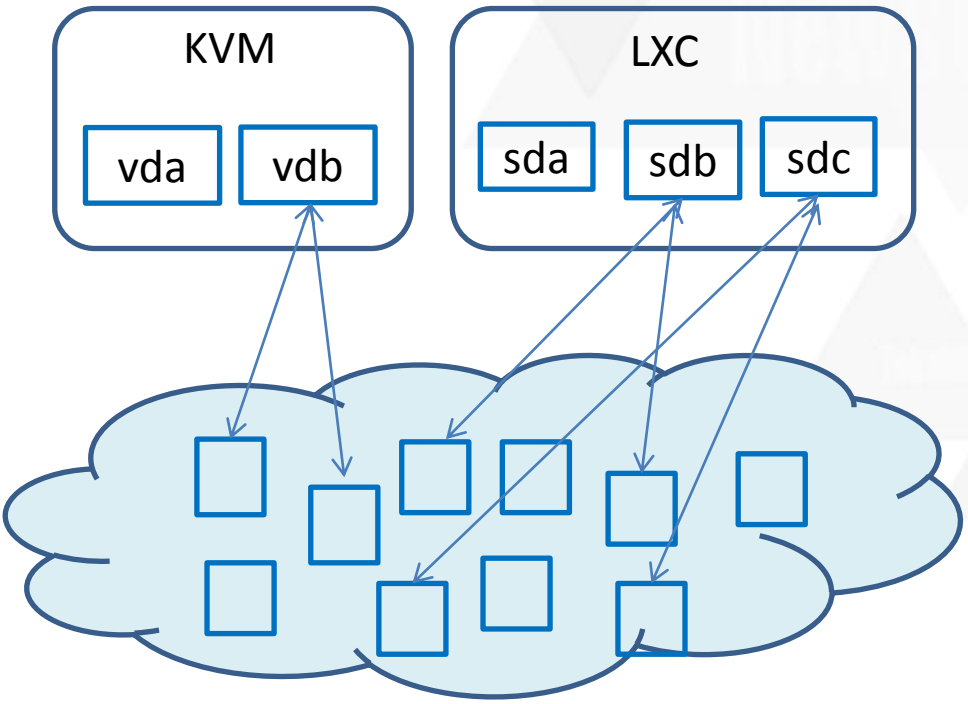
- 树形快照
  - 多条快照链，共用一个全量快照点
- 快照逻辑删除
  - 只有快照的数据块没有被引用了才能真正删除，否则就只是逻辑删除

延迟加载



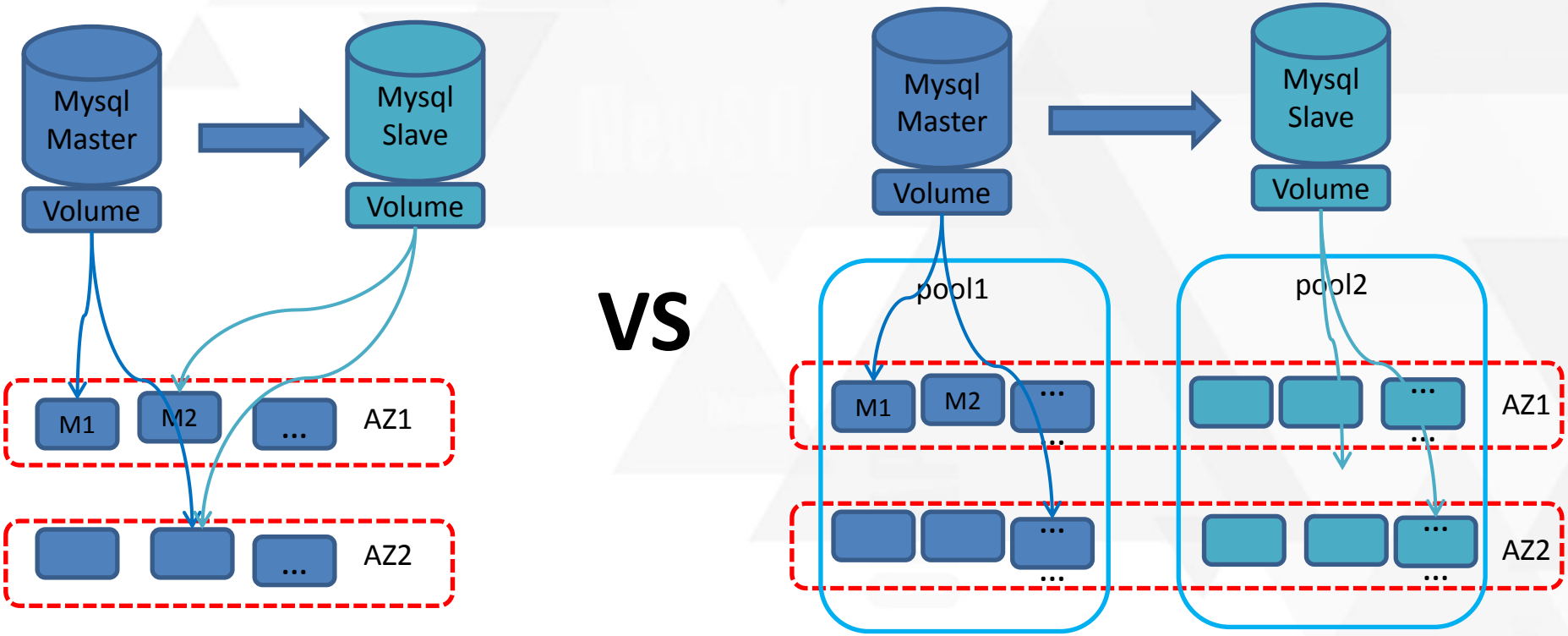
- 按需加载
  - 从快照恢复的卷不必等到全部下载完成就可以使用
- 后台任务下载
  - 后台顺序下载数据块，避免每次IO都去访问NOS

应用场景



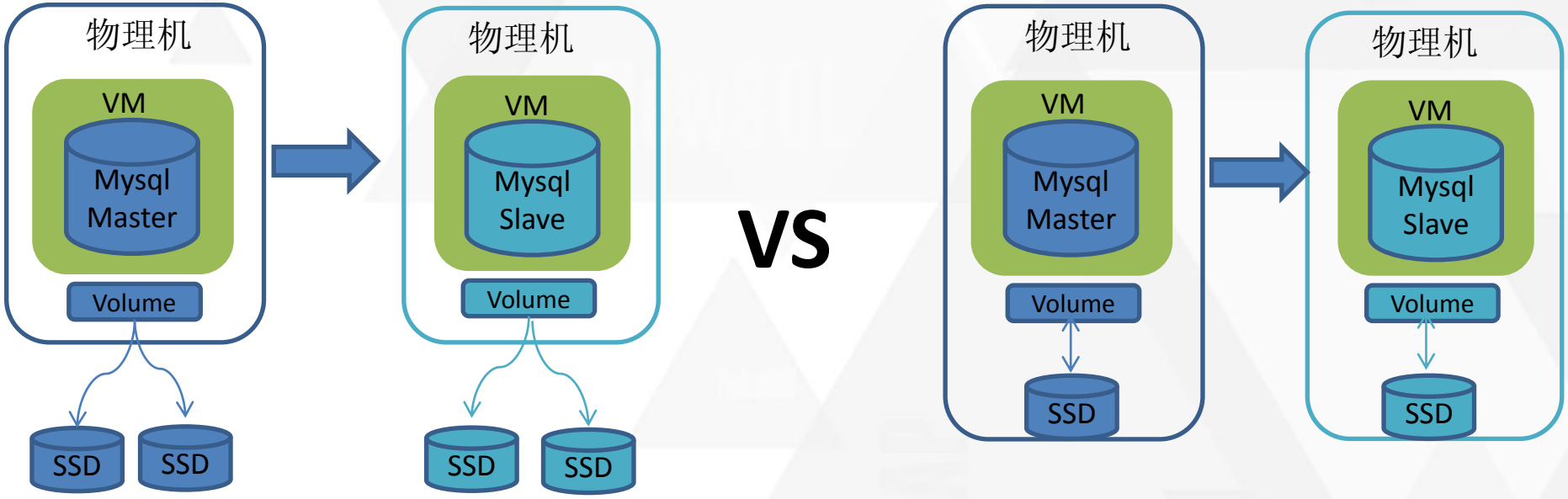
- 与KVM和LXC结合
  - 可以将云硬盘挂载到KVM或者LXC虚拟机中使用
  - 可以将云硬盘从一台虚拟机卸载，然后挂载到另外一台虚拟机使用，数据不会丢失
- 每台虚拟机支持挂载多块云硬盘
  - 将每块云硬盘分别创建文件系统mount然后使用
  - 将多块云硬盘使用软raid或者device mapper组成一块逻辑设备然后再创建文件系统mount后使用

数据库应用场景



- 高可靠：主从分不同存储池，严格意义的云硬盘4副本，可靠性更高，左边使用的主从的盘有可能落到相同的机器上，甚至相同的硬盘上；
- 高可用：左边的一台机器挂了，有可能主从都影响了，而右边是严格意义的主从分离。

数据库应用场景



- 本地单副本SSD卷，由数据库主从提供高可靠，SSD卷提供高性能，本地卷减少了网络开销，并且可用降低多副本ssd的成本开销。


ChinaUnix
THANKS