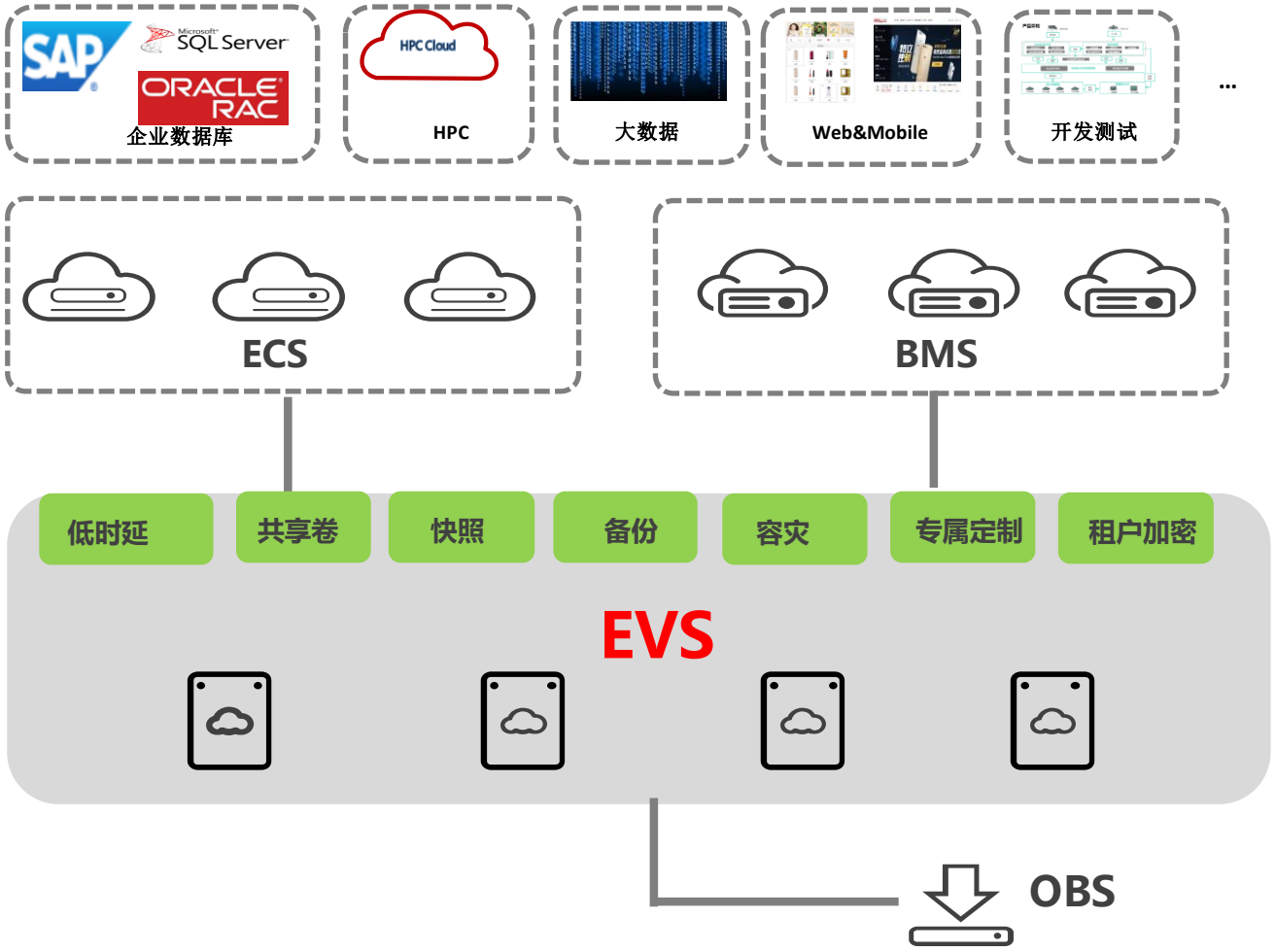
HUAWEI

企业级稳定低时延、高性能分布式存储架构剖析

张志乐　　华为云存储服务资深架构师

LEADING NEW ICT

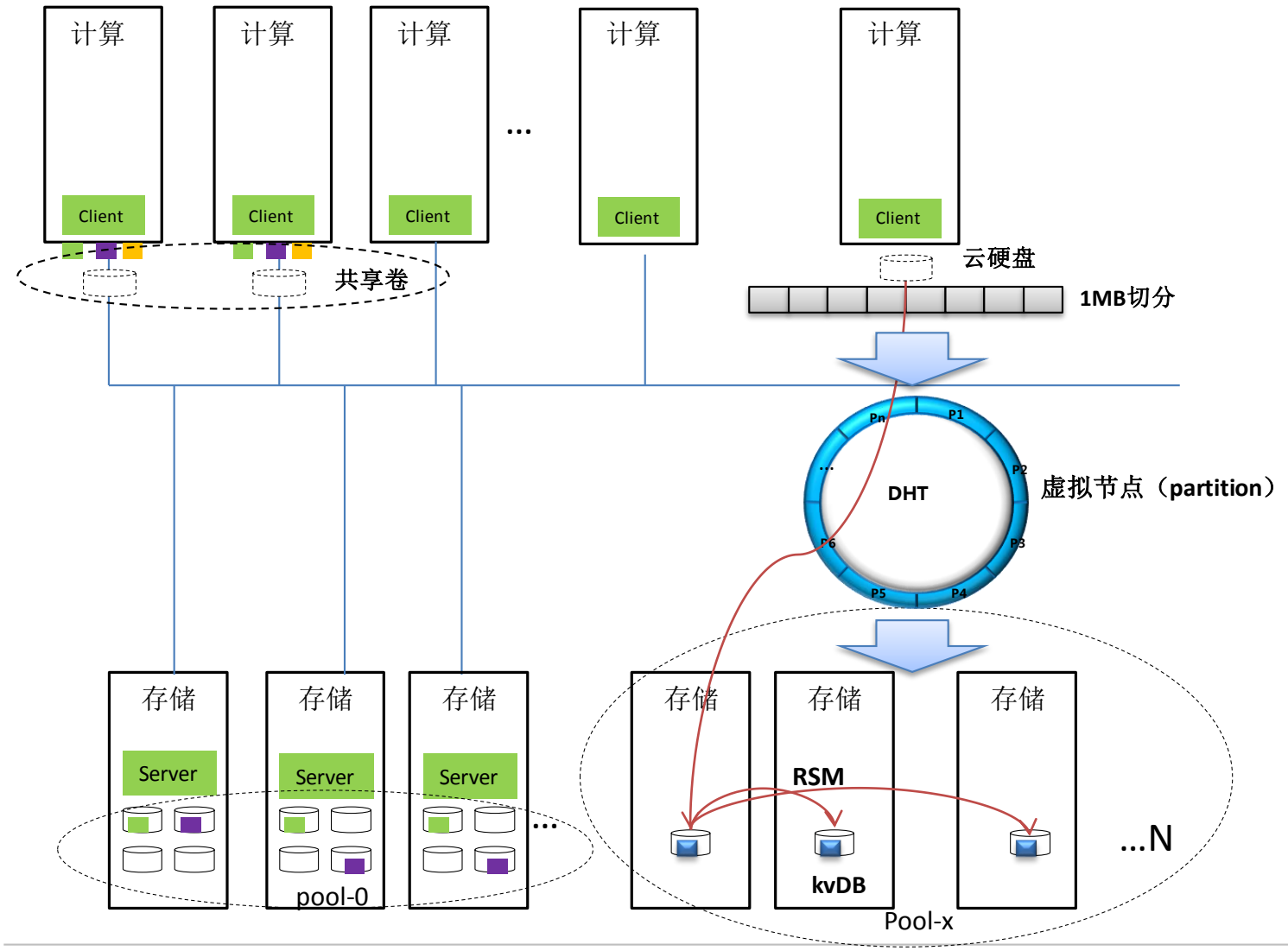
华为云EVS存储服务



多种SLA满足不同企业应用

云硬盘类型	性能	场景
·普通云硬盘	1000 IOPS, 90MB/s	普通虚机、开发测试
·高性能云硬盘	3000 IOPS, 150MB/s	Web、企业OA、大型开发测试
·超高性能云硬盘	20000 IOPS, 350MB/s	中小型数据库、NoSQL
·超高性能云硬盘（优化I型）	30000 IOPS, 1GB/s	大中型数据库、高性能HPC、关键应用

高并发弹性存储池架构



1. **Native的计算、存储分离架构：**
   - Client部署在计算节点上，无状态；
   - Server部署在存储节点上；
   - 计算、存储按需扩展；

2. **Partition, Pool, KV数据块：**
   - 整个集群划分多个存储池pool
   - Pool为故障域
   - 存储池使用Partition管理数据分区
   - 每个Partition分区多个KV数据块

3. **DHT数据寻址和路由：**
   - 卷被分割成很多1MB block
   - Client使用DHT将数据打散到不同的partition
   - Partition被映射到不同的Server/Disk上

4. **RSM强一致性复制协议：**
   - 自研强一致性复制协议，Client->主->备，IO效率高；
   - Partition有主备副本关系，服务器/磁盘没有副本关系；

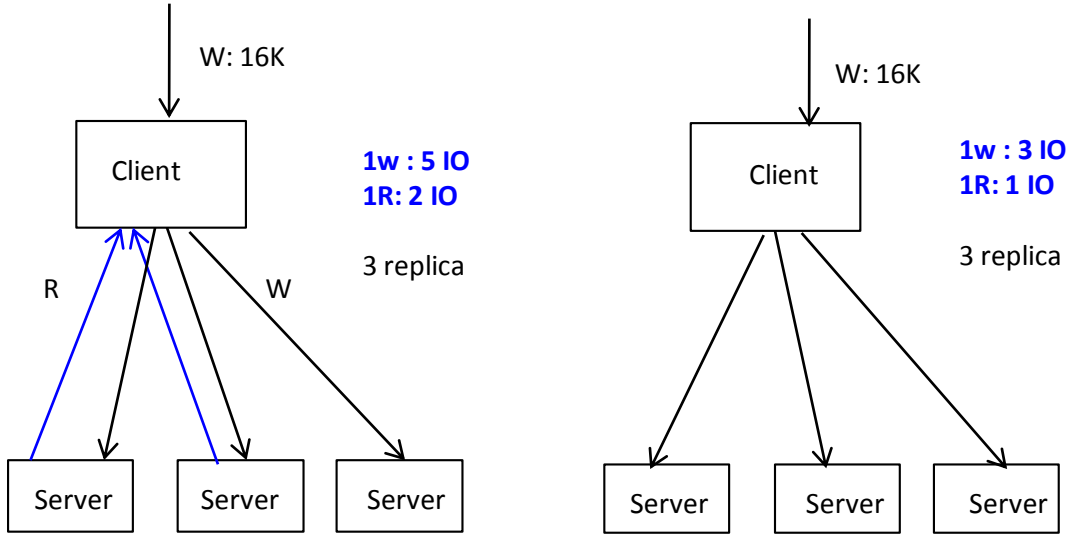
5. **支持RoCE、Infiniband、TCP/IP多协议组网：**
   - 支持RoCE组网，低时延
   - 支持普通TCP/IP组网
   - 支持Infiniband组网，高吞吐

6. **支持A-A并发读写访问模式和共享卷：**
   - 支持A-A并发读写共享卷
   - 无集中Target、最高32万IOPS共享卷；
   - 支持SCSI预留锁，运行企业legacy应用；

稳定时延关键技术—复制和数据路由

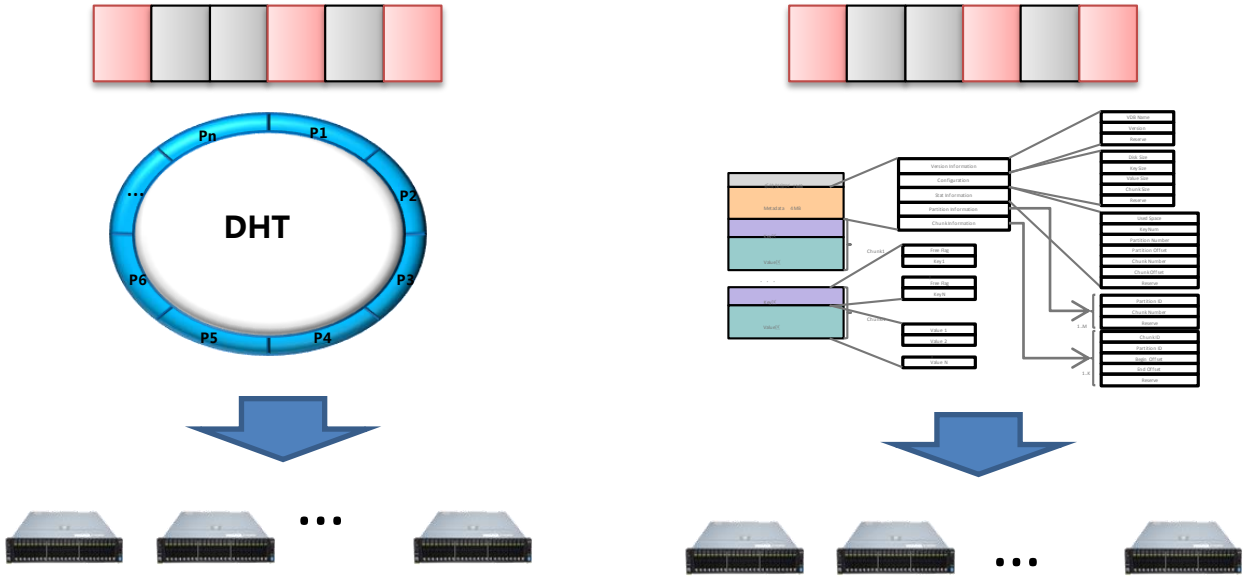
最终一致性 vs 强一致性

(e.g. NRW) (e.g. RSM, Raft)



1. IO效率和读写时延：
   - NRW=322, 1 write， 5 disk IO； 1 read，2 disk IO；
   - RSM， 1 write，3 disk IO; 1 read, 1disk IO;
2. 状态同步：
   - Gossip；
   - MDC集中式心跳
3. 容错能力：
   - 2F+1 tolerant F failure，3份拷贝容错1份；
   - F+1 tolerant F failure，3分拷贝容错2份；

DHT路由 vs 元数据路由



1. 读写时延：
   - DHT Hash计算，路由表全内存；
   - 不命中时存在元数据查找;
2. 数据量与性能的关系：
   - 数据量增多，路由信息量不变；
   - 数据量增多，元数据增多，可能导致性能的损失；
3. 路由信息备份和应急：
   - 路由表信息少，便于备份和应急；
   - 元数据较复杂去做备份和应急；

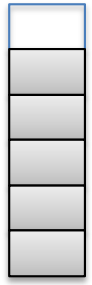
对于高性能块存储而言，强一致性复制协议的IO效率更高，DHT方式时延更稳定。

稳定时延关键技术—介质处理

原理：SSD垃圾回收

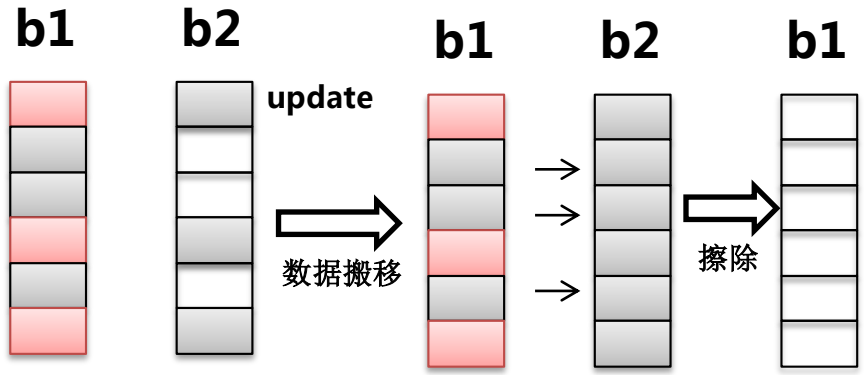
空间管理

block
(128~256pages
)



Page (4/8kB)

按照page分配写入空间
数据只能写入空白页
写入模式为append only
空间回收只能按照block粒度

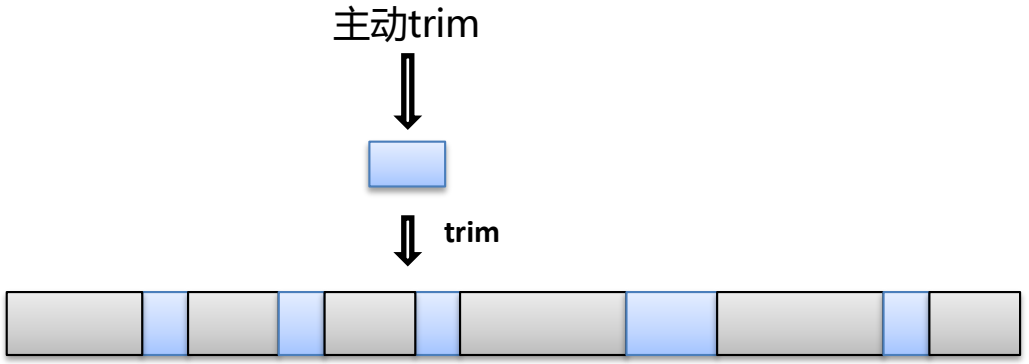


1. 垃圾回收会引入写放大，影响寿命
2. 大量垃圾回收会导致SSD性能下降，**时延增大10倍！**

手段一：多流分离

1. 不同IO模型的多流分离： 元数据小IO读写、数据IO读写分离；
2. 小IO流频繁的擦写，不会导致整个SSD全局空间的频繁擦写；

手段二：无效空间免搬移



1. 删除空间，软件主动Trim回收；
2. SSD GC时不会再进行搬迁，有利于寿命和性能；

稳定时延关键技术—网络处理

IO卡顿对应用的影响

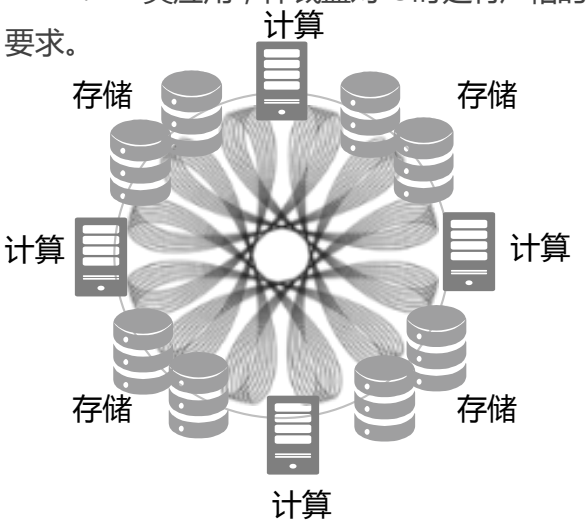
时延敏感型应用对IO卡顿零容忍

1. Oracle RAC OCR卷IO超时（15秒），数据库实例直接退出，只能人工恢复。

2. SAP HANA数据库，IO超时120未返回，数据库异常退出。

3. 数据库日志IO，通常30秒超时后，数据库事务异常，数据库实例直接退出。

4. HA类应用，仲裁盘对IO时延有严格的要求。

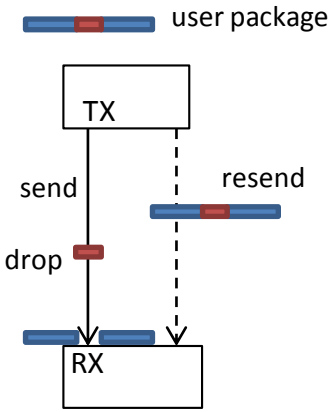


网络拥塞问题

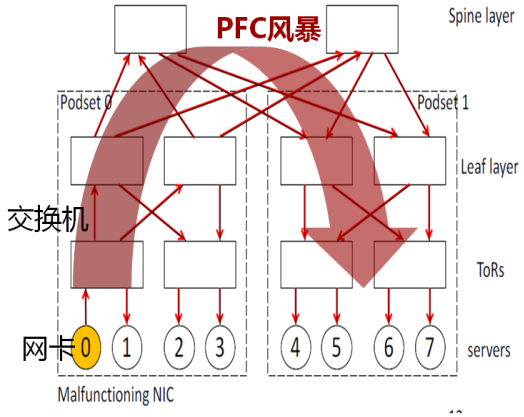
RDMA网络拥塞

1.相对于tcp，rdma协议栈对于网络拥塞容忍度低

2. 大规模集群下，多对多数据传输，问题突出

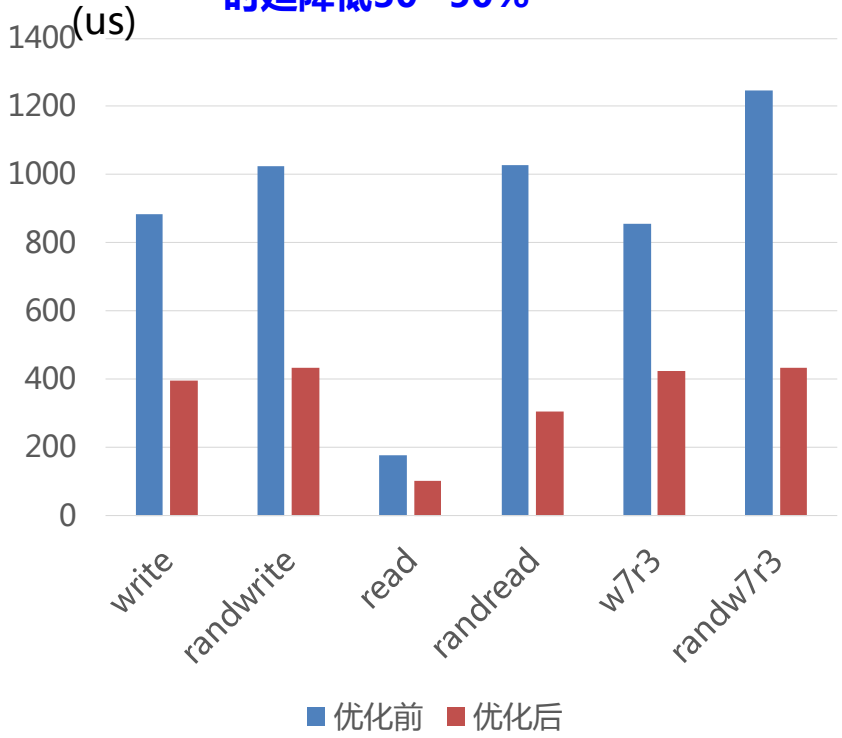


PFC风暴



RDMA网络拥塞检测控制效果

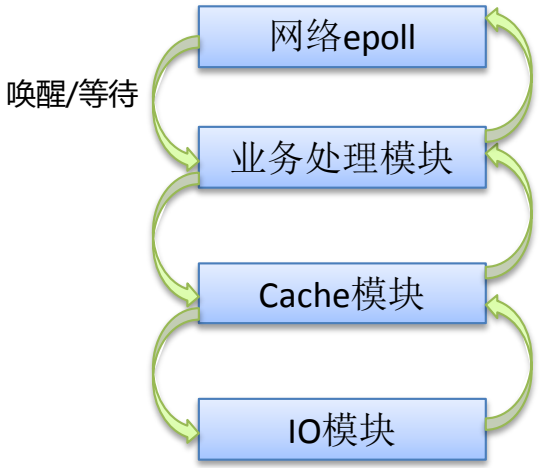
时延降低30~50%



RDMA网络拥塞控制前后时延性能对比

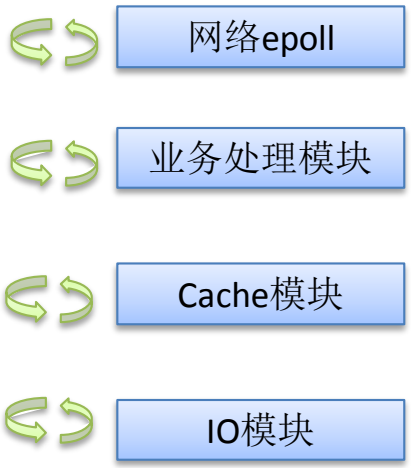
稳定时延关键技术—软件和代码调度

线程调度：异步线程pipeline



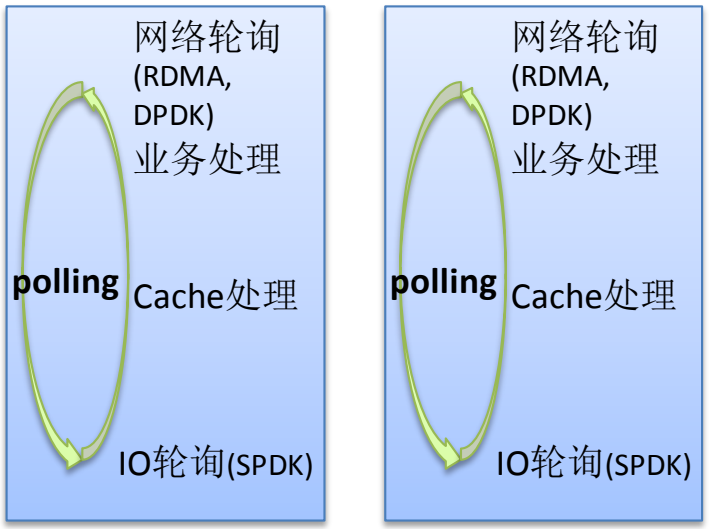
4轮等待唤醒，耗时40us~200us

智能唤醒



4次线程上下文切换，耗时5us~20us

线程调度：轮询



无上下文切换、无中断、零等待

智能线程调度

预唤醒：IO处理入口，唤醒所有处理线程，降低线程切换时间。

智能等待：线程调度和业务处理逻辑深度融合，根据IO并发度和处理时长自动触发线程等待，避免cpu无效消耗。

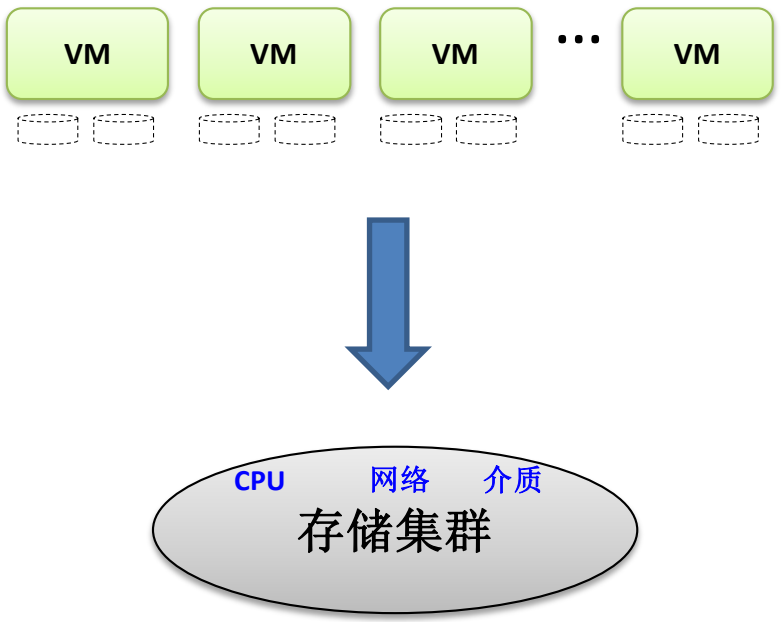
Polling Mode：

网络：通过RMDA轮询或者DPDK轮询，节省内核到用户态的拷贝和切换；

磁盘：通过SPDK，减少中断和切换；

稳定时延关键技术--精细化QoS

为什么要QoS



QoS控制策略和机制

策略一： 按容量 每GB IOPS，每TB带宽

- HDD、SSD性能不同，每GB IOPS有区别；
- 顺序和随机有区别；

策略二： Max控制

- 最大IOPS、MBPS控制；

策略三： Burst

- 小容量(e.g. 10GB系统盘)，突发IO需要Burst(如虚拟机启动瞬时1000+iops）；
- Burst时间控制(30 Minutes、闲时配额积累)；

对云盘的评测误区

- **本地盘 VS 云硬盘**
- **无QoS控制 VS 有QoS控制云硬盘**
- **不同类型云盘对比**
- **不同大小云硬盘对比**

多租户隔离、SLA保证

- **CPU/网络/介质资源的隔离：**在无限的资源池中使用有限的合法资源。
- **租户资源的公平调度；**

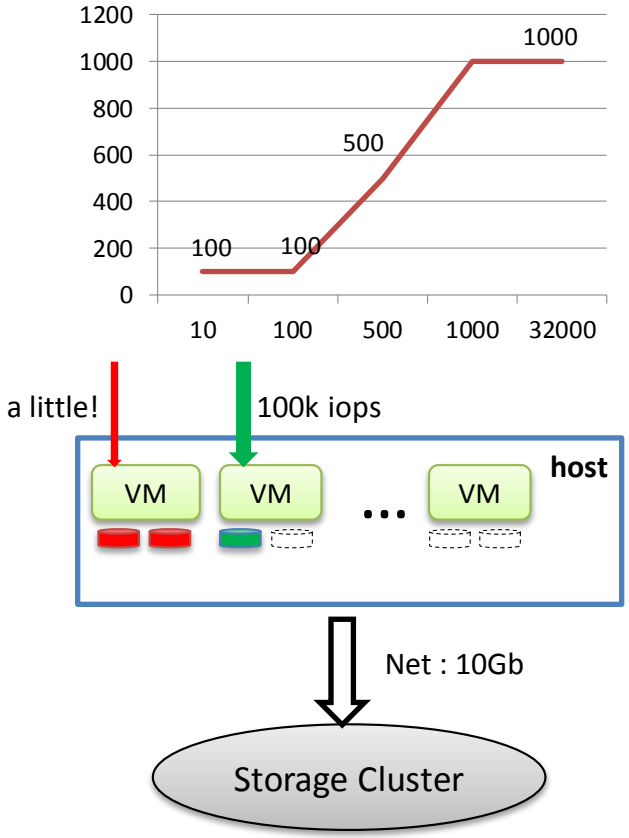
智能ms级QoS控制

- **在接近QoS上限时，<5%波动；**
- **精细化QoS：**按照容量、云盘类型提供精细化的QoS策略，为每个应用提供稳定性能

EVS云盘性能演示

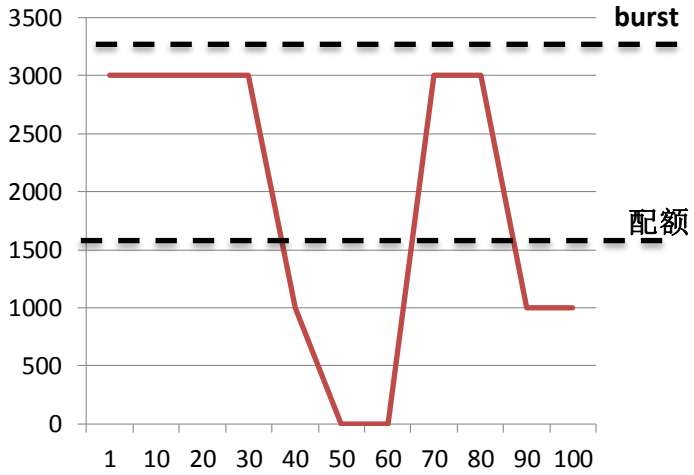
QoS控制

普通云盘QoS



Burst演示

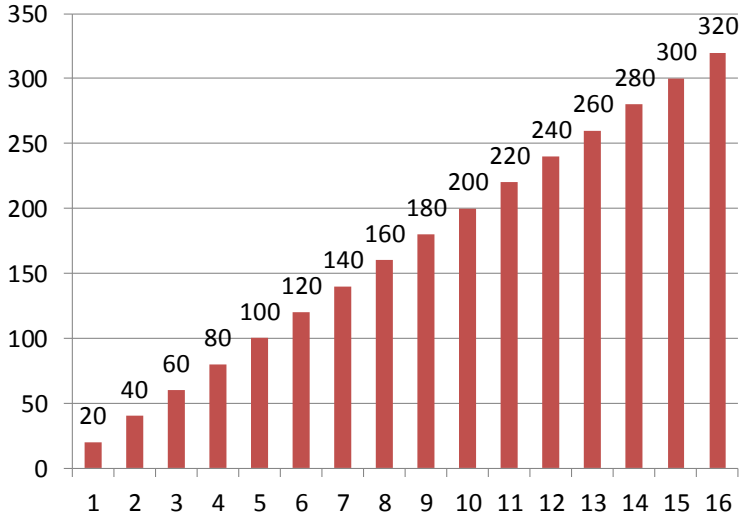
高性能云盘burst



持续30分钟Burst！

共享云硬盘

超高性能共享云盘



共享云盘性能线性增长！