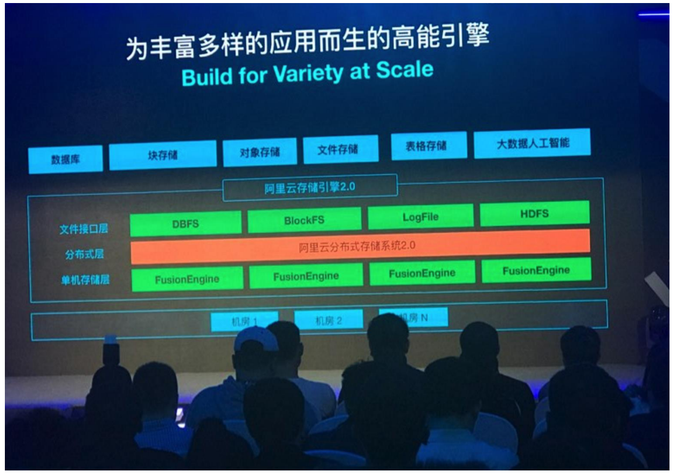
阿里云 | 奥运会全球指定云服务商

基于SPDK的用户态存储引擎：FusionEngine 2.0

阿里巴巴 王正勇/张翼

FusionEngine的诞生

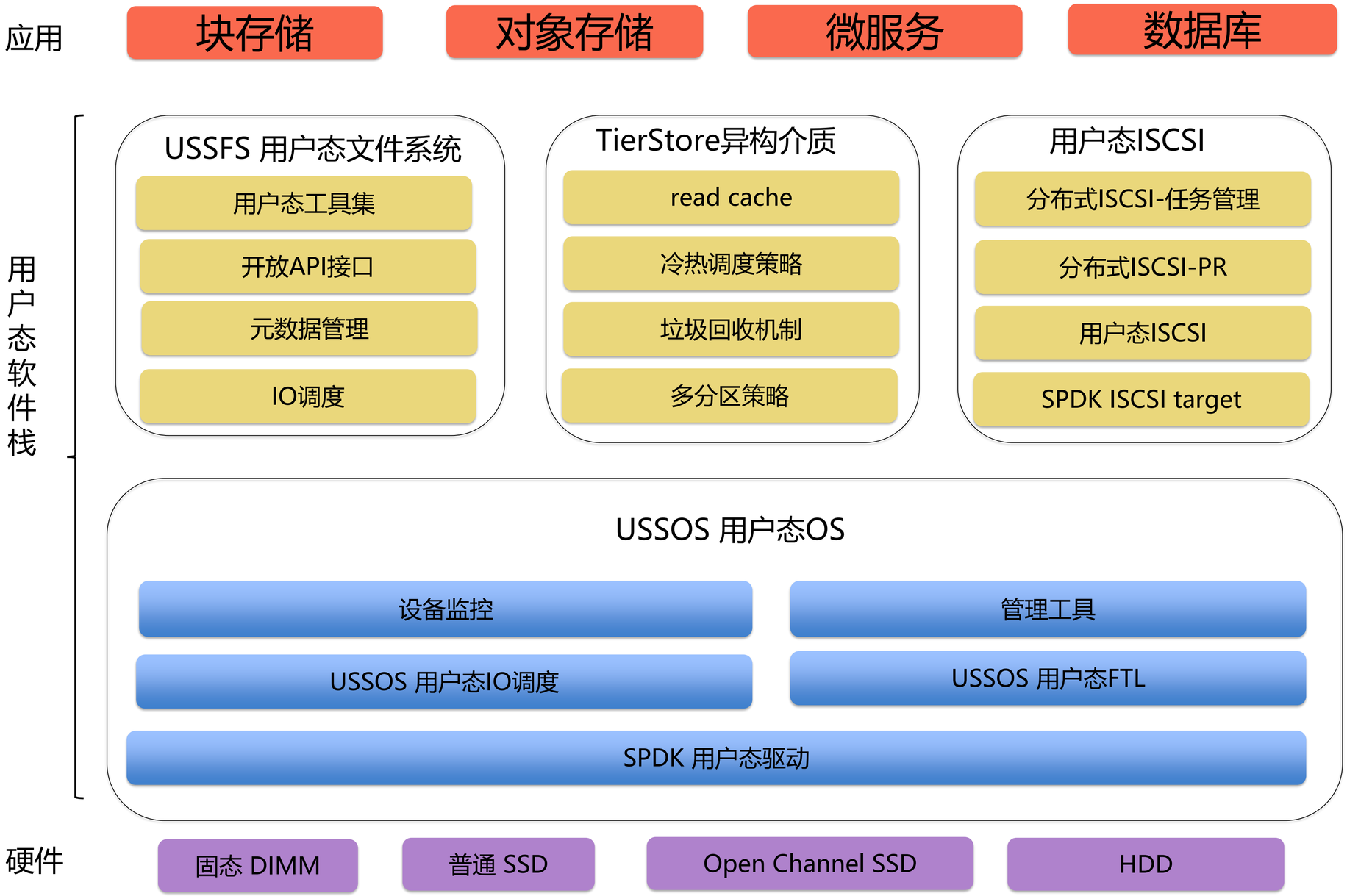


FusionEngine
统一底层存储平台

- 业务诉求：
  - 硬件红利快速获取
  - 订制获得技术壁垒
  - 定位问题迅速可控
  - 高性能,低成本,超稳定

FusionEngine 2.0 总体架构图

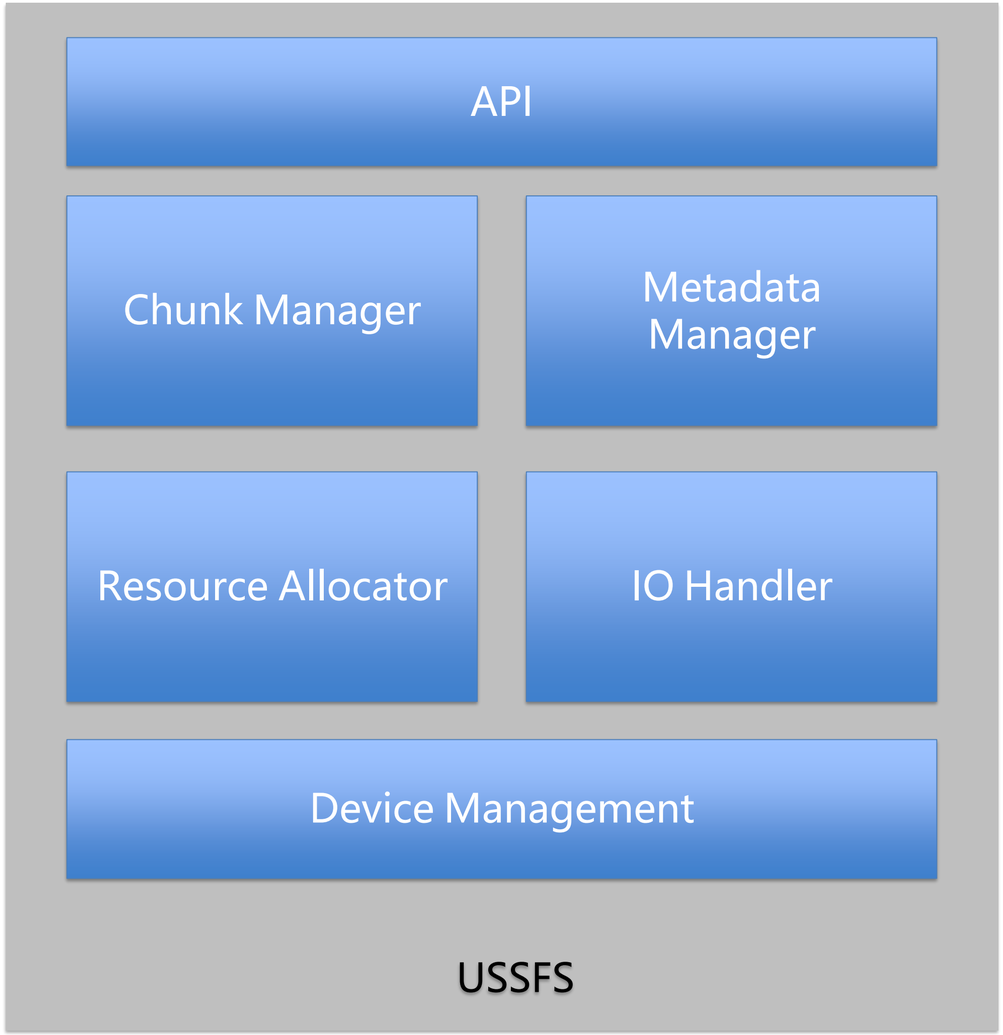
阿里云 | 奥运会全球指定云服务商



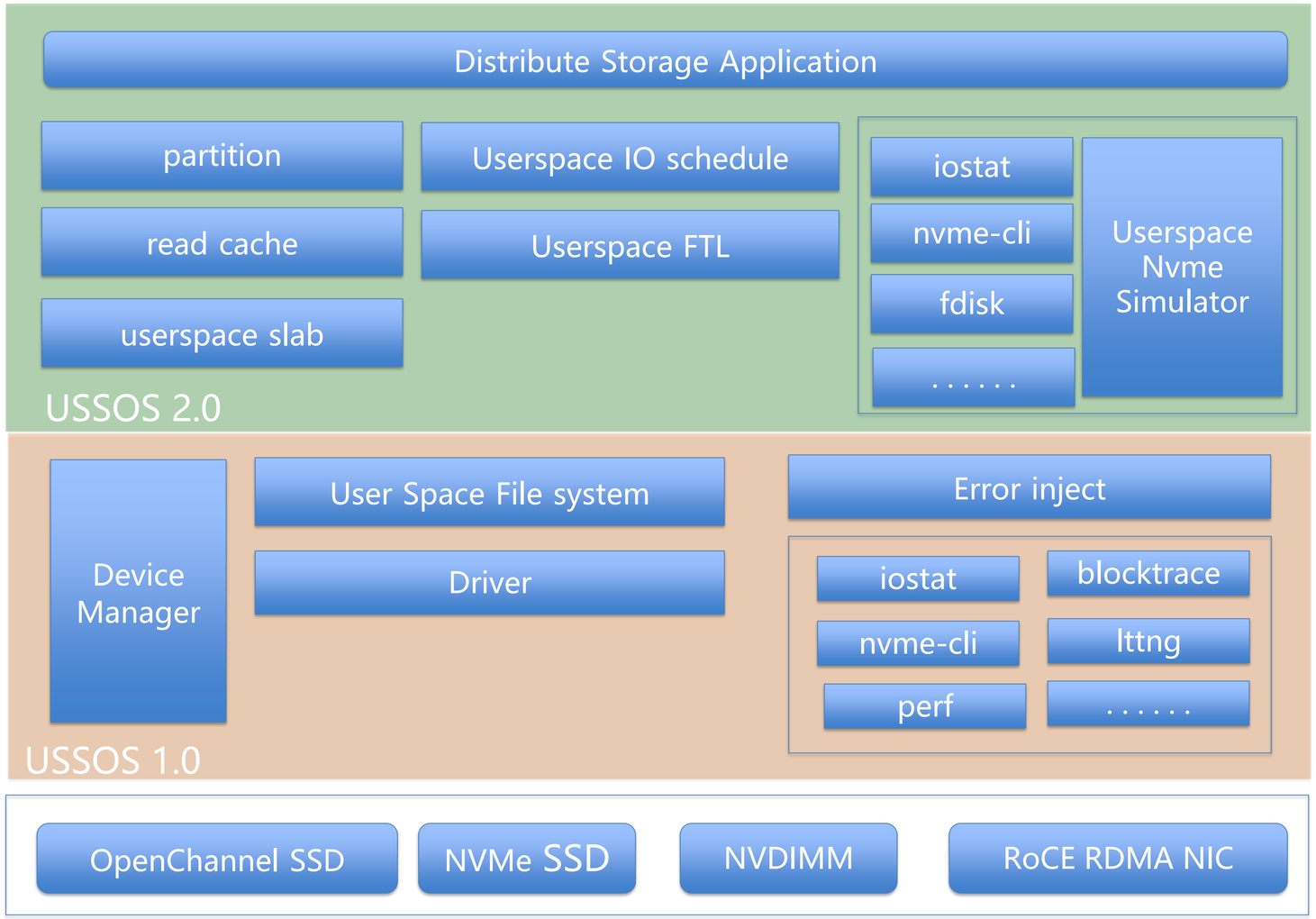
- 降低SSD写放大
- 降低CPU使用率
- 用户态全链路QoS
- 数据管控面分离
- 用户态/内核态定制IO栈支持
- 存储成本降低
- 最新硬件支持

User Space File System（USSFS）

- 建立在裸盘基础之上的本地文件系统
- 将存储介质抽象成一系列Chunk资源
- 以API的方式对外提供服务
- 非日志型文件系统，减少写放大
- 面向高性能存储介质设计，元数据更新频率降低到ext4的1/1000
- 数据与管控路径分离
- 支持Run-to-Complete的线程模型

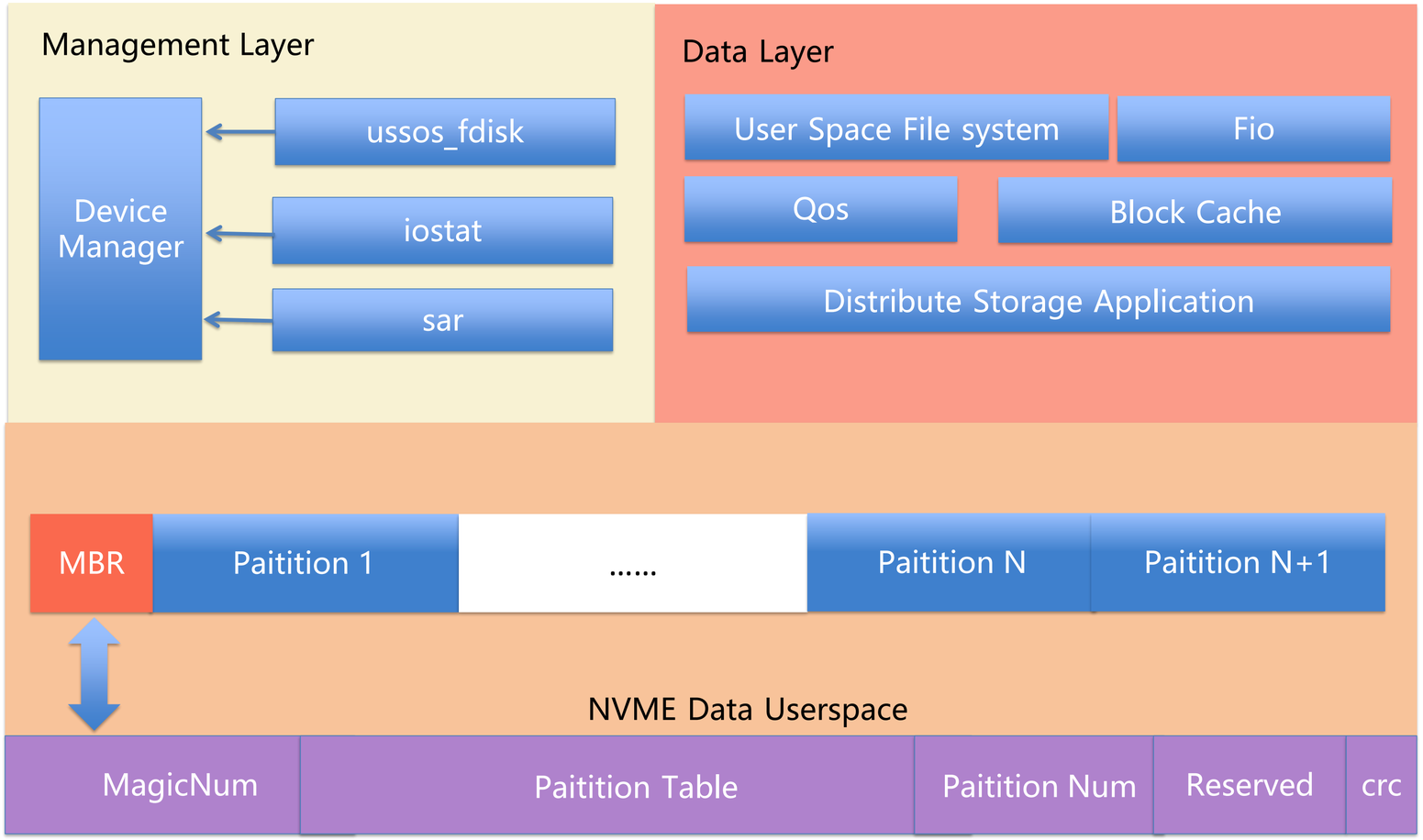


USSOS 2.0



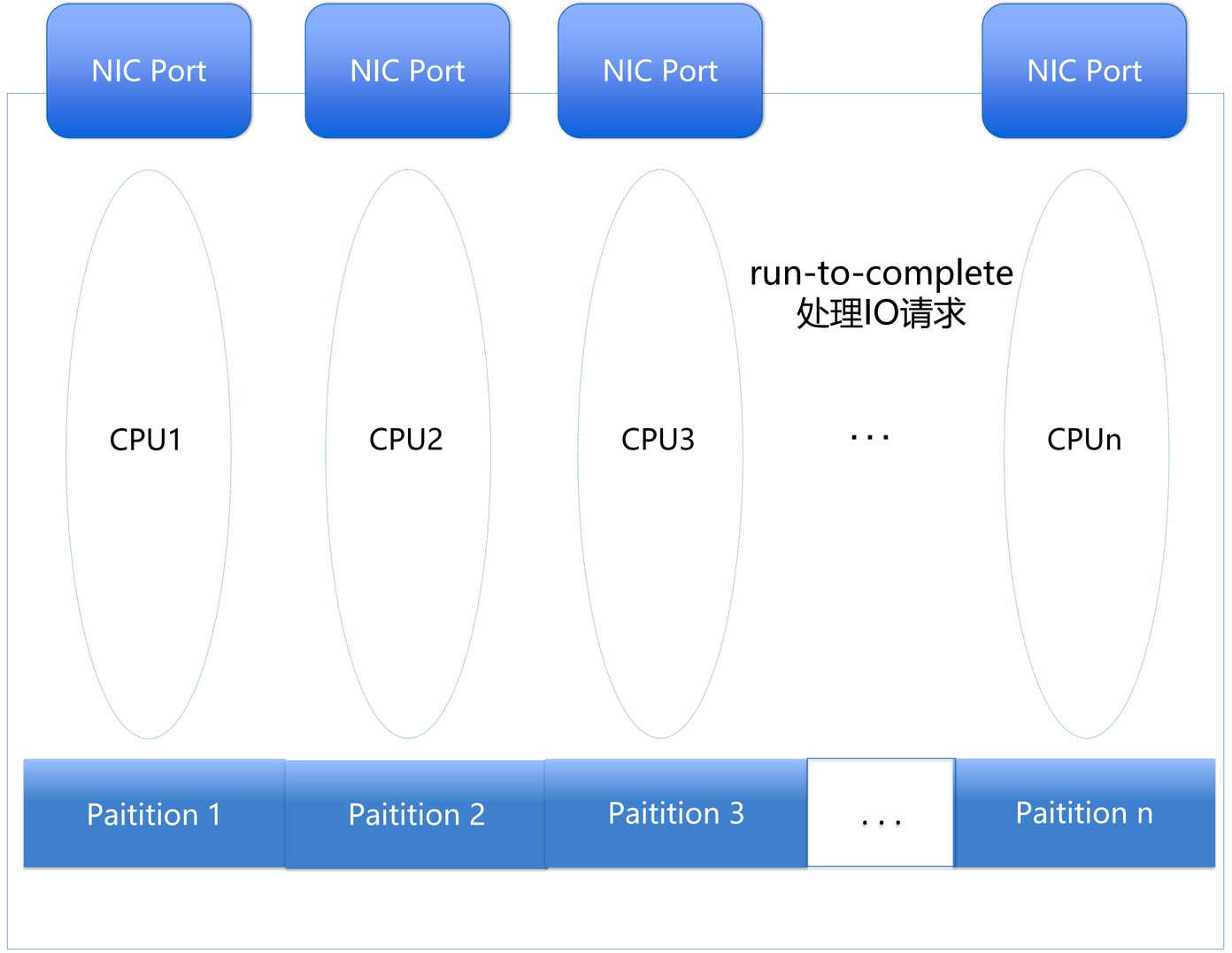
- 裸设备多分区支持
- 全用户态定制化IO调度策略，提升整盘/分区的QoS
- 用户态读缓存机制
- 用户态FTL
- 用户态slab，提升内存分配效率
- 用户态工具增强及支持多分区
- 用户态nvme磁盘模拟器

用户态多分区



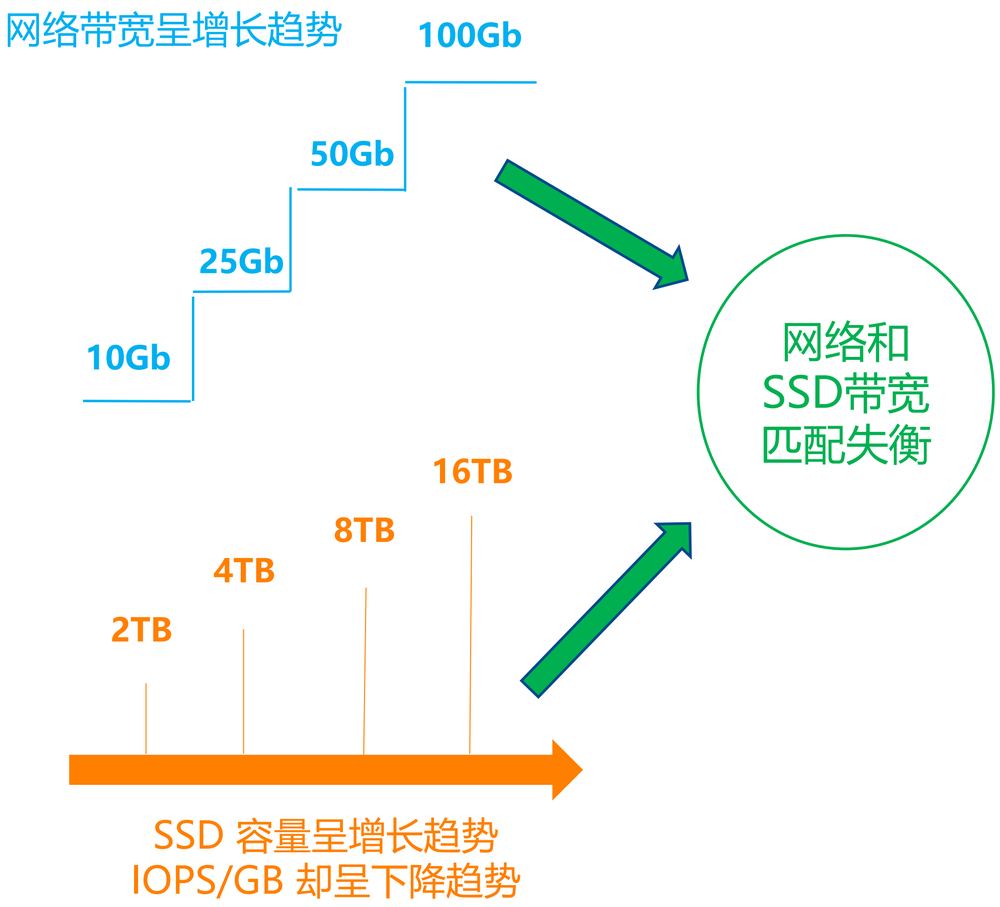
- 基于用户态空间实现
- 支持全盘/分区的QoS
- 主流磁盘工具支持
- 多用户共享单盘
- 资源隔离

多分区 Run-To-Completion Polling 机制



- 降低IO延迟
- 提升处理器IO处理效率
- 发挥高性能介质的能力
- 整盘支持多核 polling

TierStore的项目背景

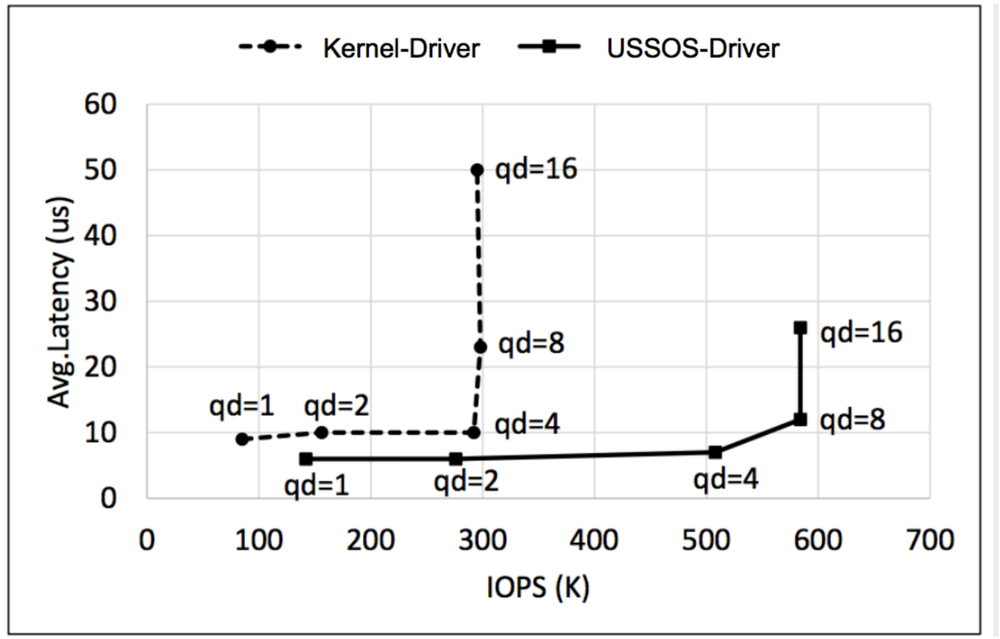


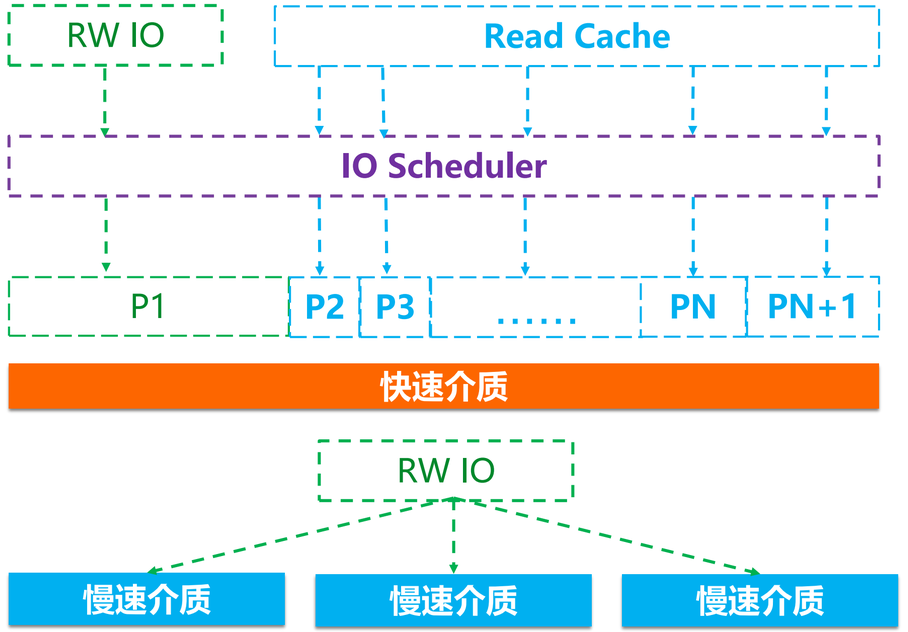
- 公有云按照IOPS/GB方式售卖
- SSD IOPS/GB却呈下降趋势
- 为了匹配网络带宽的递增，服务器中存储的密度同样需要匹配
- 以4T SSD，100Gb网络带宽为例：

  4T SSD： 24 Disks，96TB

  8T SSD： 20 Disks，160TB
- 导致实际成本将会递增

TierStore 用户态线程模型





- 内核态驱动：随着qd增加，iops先增后降
- spdk用户态驱动：高qd下，iops优势明显
- TierStore快速介质池带宽要求极高，基于spdk的用户态驱动才能满足需求

- 充分发挥快速介质的硬件性能
- 读缓存与IO scheduler降低长尾延迟
- 1个core带多块慢速介质，提升CPU使用效率

TierStore 多样化的存储池

阿里云 | 奥运会全球指定云服务商

Storage Class Memory

TLC

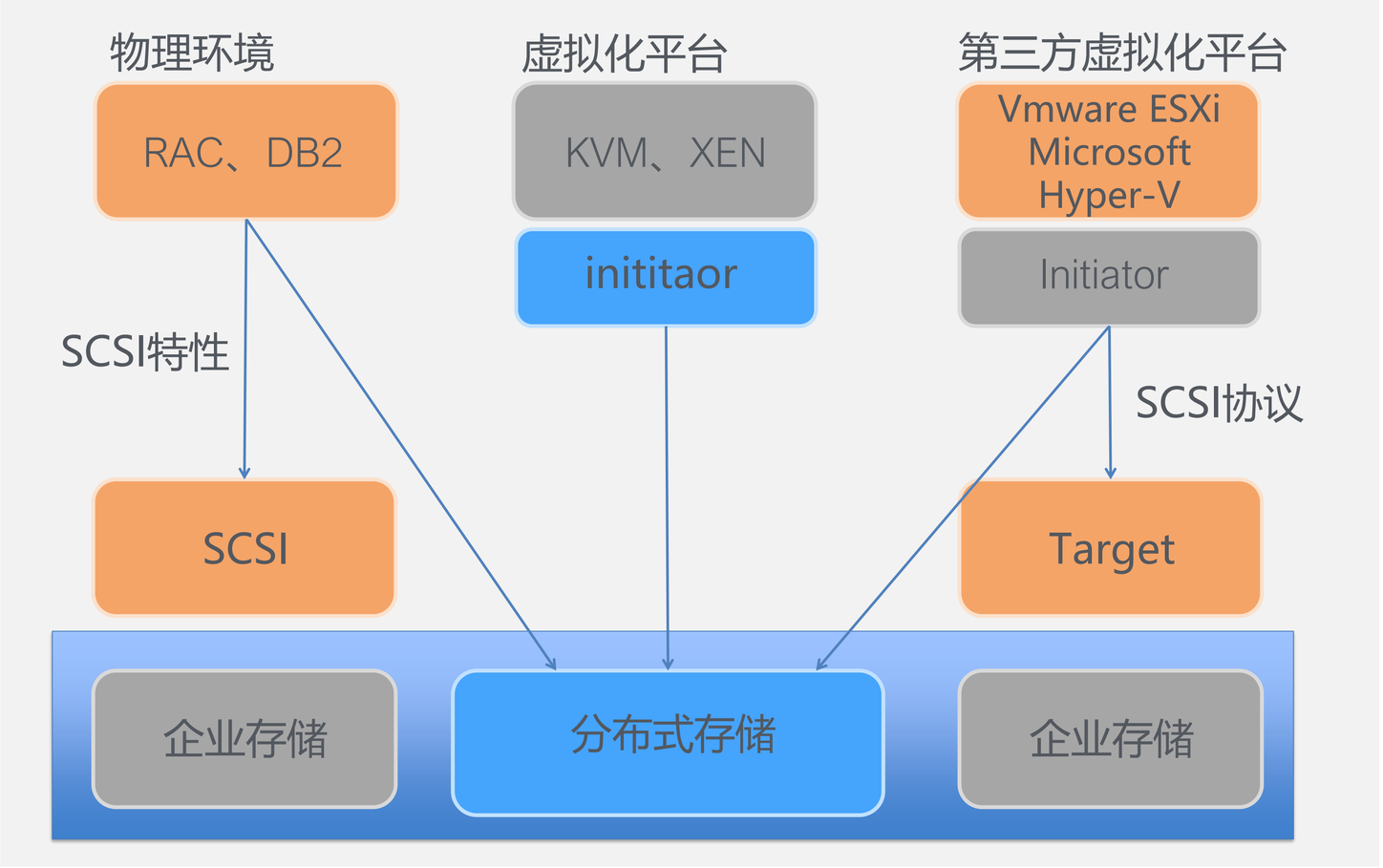
HDD

QLC

快速介质

慢速介质

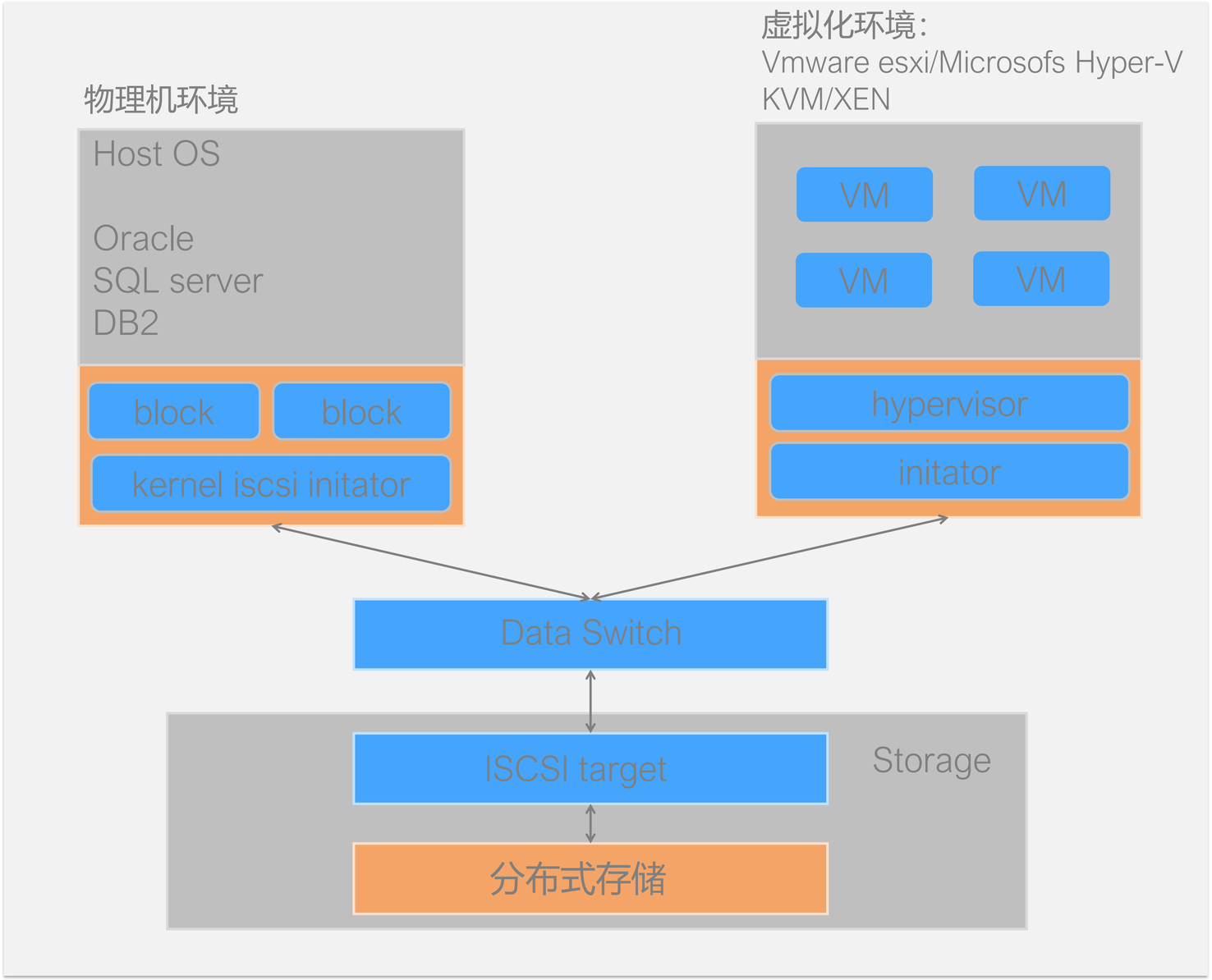
ISCSI



需求

- 传统企业数据中心
- 第三方虚拟化平台
- 虚拟化环境
- 物理环境

总体方案



虚拟化平台和物理环境

- 第三方虚拟化平台
- 通用虚拟化平台
- 物理环境/企业数据中心
- 存储端Target标准ISCSI协议

Target选型

选型

- 高性能
- 社区活跃度高，良好的生态环境
- SCSI协议成熟度
- 更好的结合分布式存储
- 内核态&用户态

备选方案

SCST

- 比较成熟
- Fusion-io公司
- 内核态
- 故障影响整机

LIO

- 比较成熟
- 进入kernel主线
- 线程模型难以发挥多连接并发读写
- 分布式系统下对iSCSI协议的支持

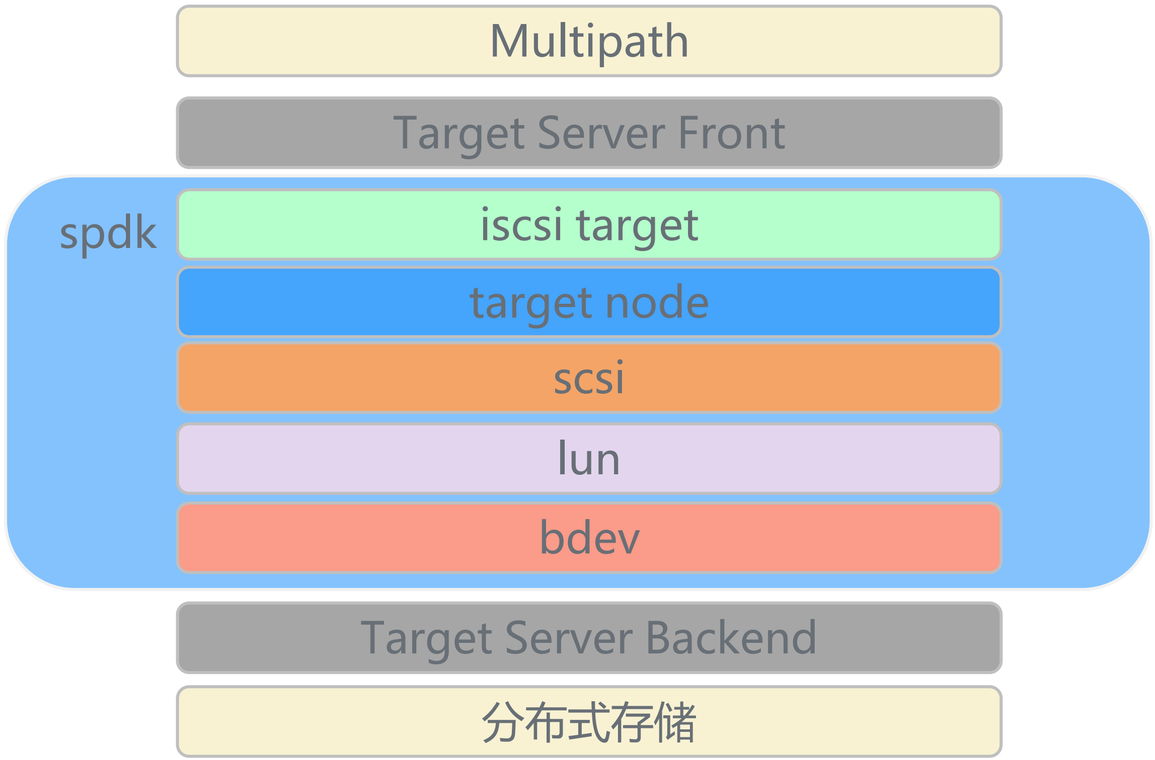
TGT

- 用户态
- 后端存储模块化
- 多线程锁开销大
- 接收主线程瓶颈

SPDK

- intel
- 社区活跃
- 协议支持好
- 用户态
- PM模型

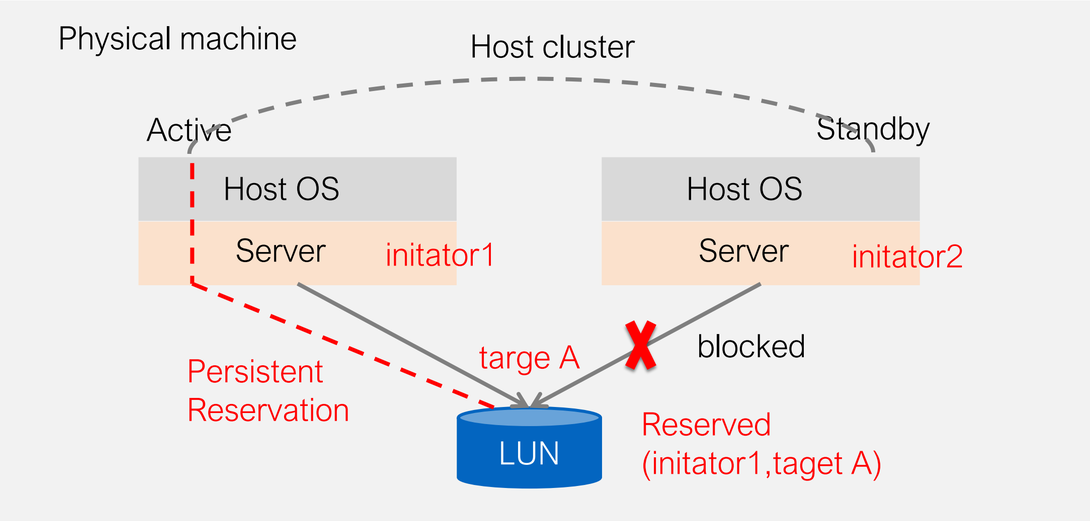
ISCSI架构

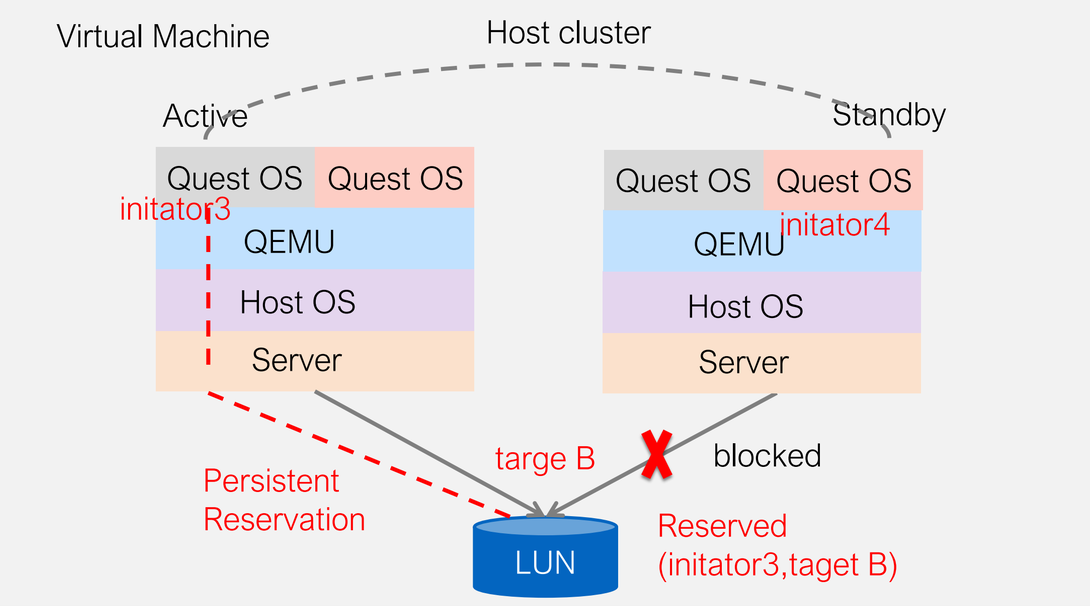


框架特点

- 多路径
- Target server Front维护和管理LUN映射
- iscsi target基于spdk框架实现iscsi协议，支持多主机共享盘
- bdev后端块设备抽象层,连接不同块设备
- Target server Backend后端设备抽象层，连接不同设备和存储协议

共享盘架构

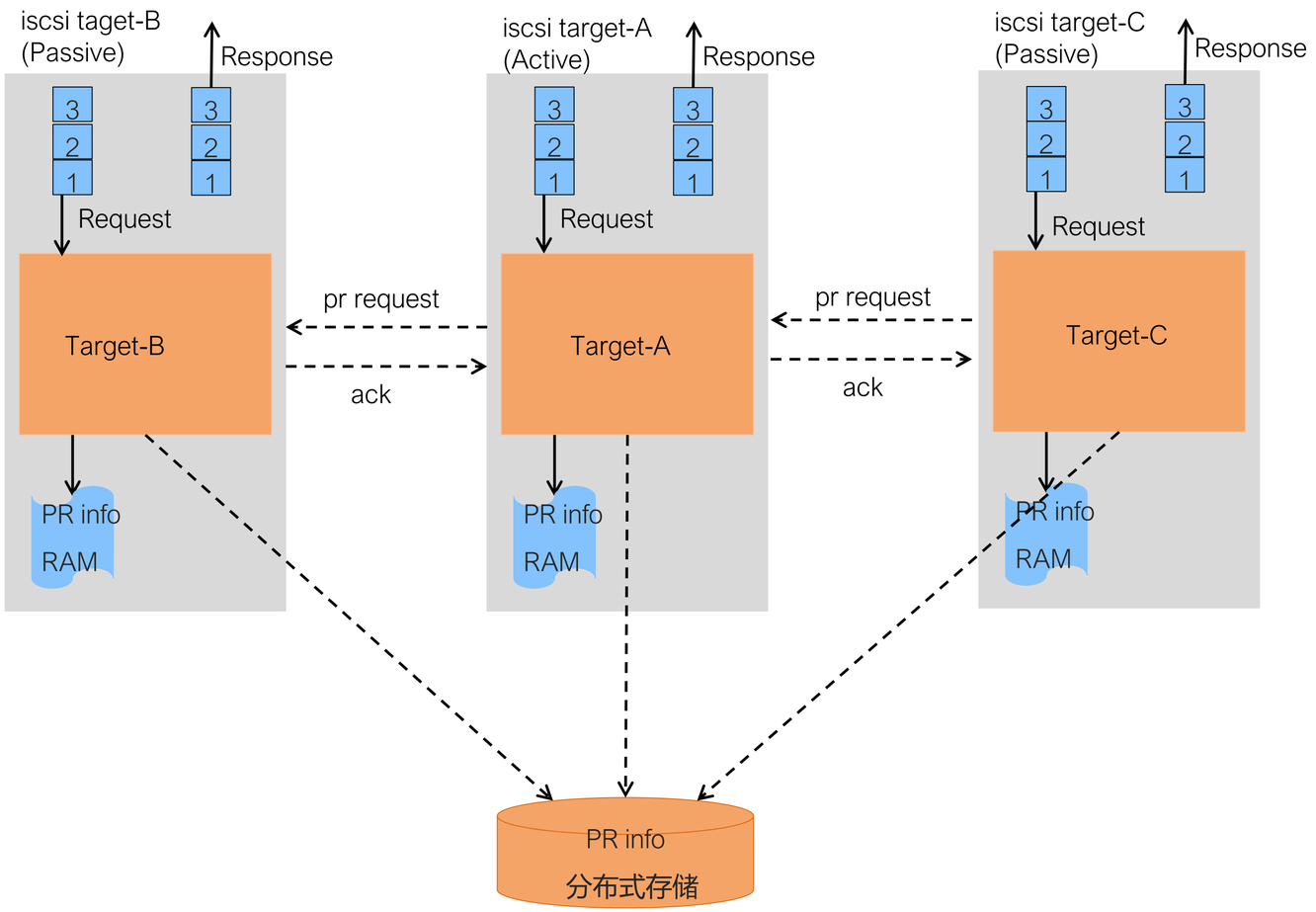




共享盘

- HA集群Guest互斥
- 多主机共享同一个LUN
- LUN支持标准ISCSI协议，提供VM SCSI块设备
- 主机间通过ISCSI Persistent Reservation 3(分布式)对共享盘进行互斥

核心技术点- 分布式PR



特点

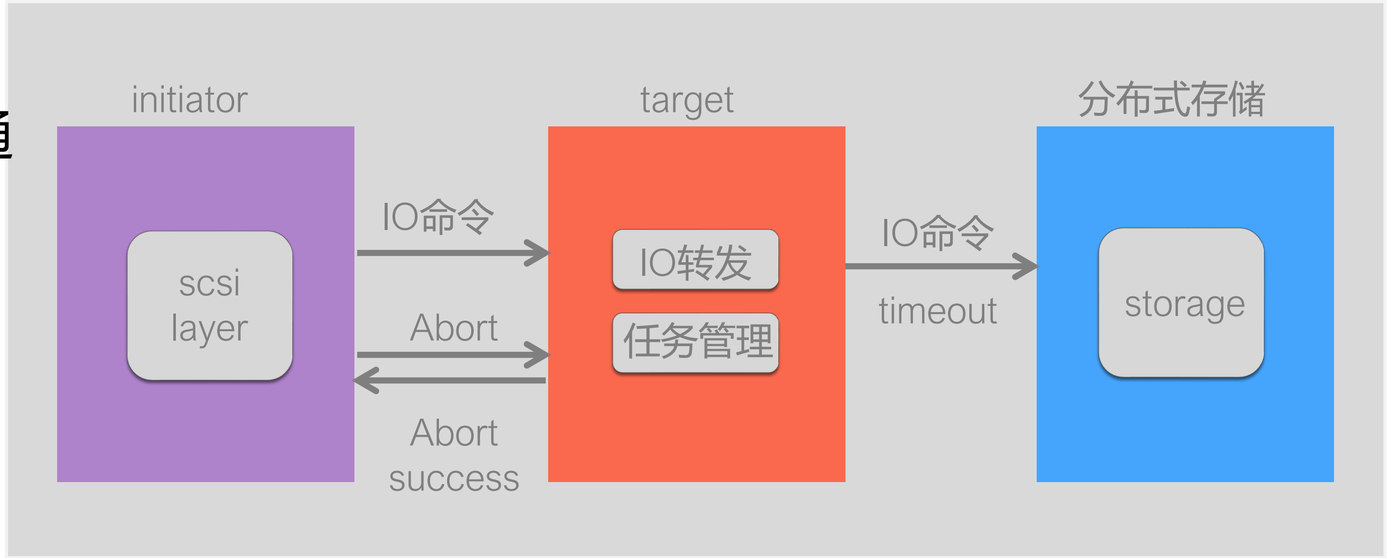
- 存储后端iscsi target多实例
- 实例间共享同一个后端存储盘
- 业务通过多个iscsi target实例进行iscsi persister reservation预留功能进行互斥
- 读写IO通过本地副本信息进行判断
- PR信息持久化在后端分布式存储
- 实例间PR信息一致性通过一致性协议保证

核心技术点- 任务管理

阿里云 | 奥运会全球指定云服务商

任务管理

- 任务管理是SCSI协议特有的
- SCSI任务在未完成之前一直存在，超时应用通过任务管理与后端交互
- 云盘后端超时，会造成vm io hang
- 任务管理配合SCSI中层，可以减轻io hang对应用的影响



协议完善

阿里云 | 奥运会全球指定云服务商

- media changer support
  - cmd set: SMC command set
  - cmds: MOVE MEDIUM/READ ELEMENT STATUS/POSITION TO ELEMENT/INITIALIZE ELEMENT STATUS/RELEASE ELEMENT

- tape drive support
  - cmd set: SSC comand set
  - cmds: FORMAT MEDIUM/LOAD UNLOAD/ERASE/LOCATE/REWIND/VERIFY/WRITE FILEMARKS

联系我们

阿里云 | 奥运会全球指定云服务商

王正勇

张翼

奥运会全球指定云服务商