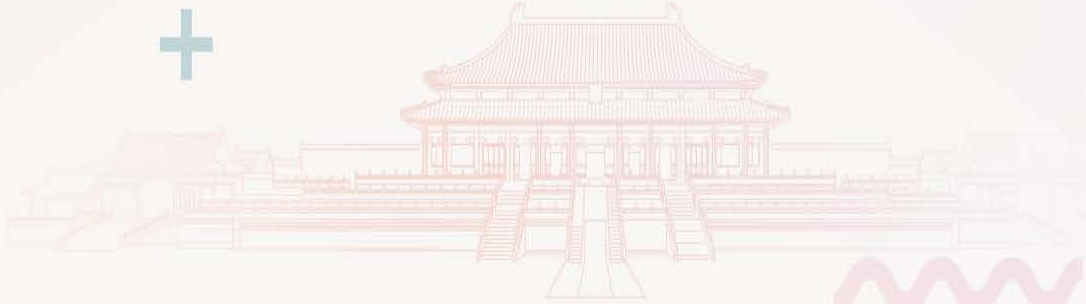
CEPHALOCON APAC 2018
THE FUTURE OF STORAGE
22-23 March 2018 | BEIJING

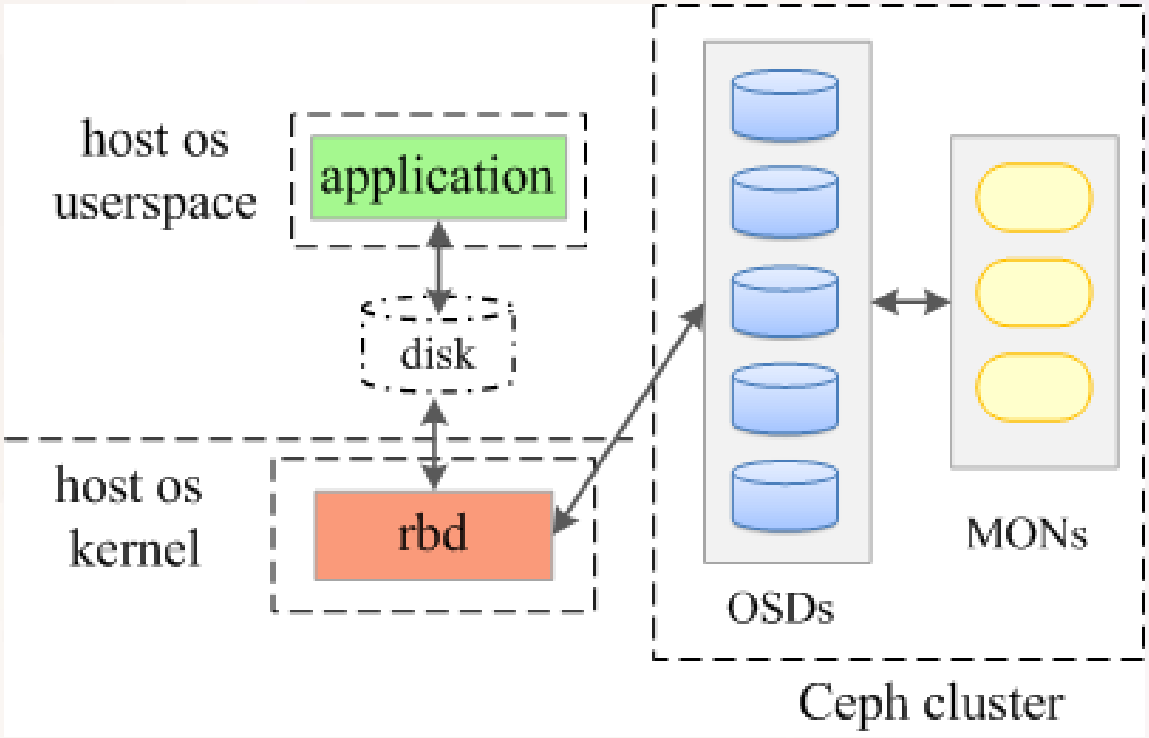
rbd-nbd的设计实现及应用

汪 黎



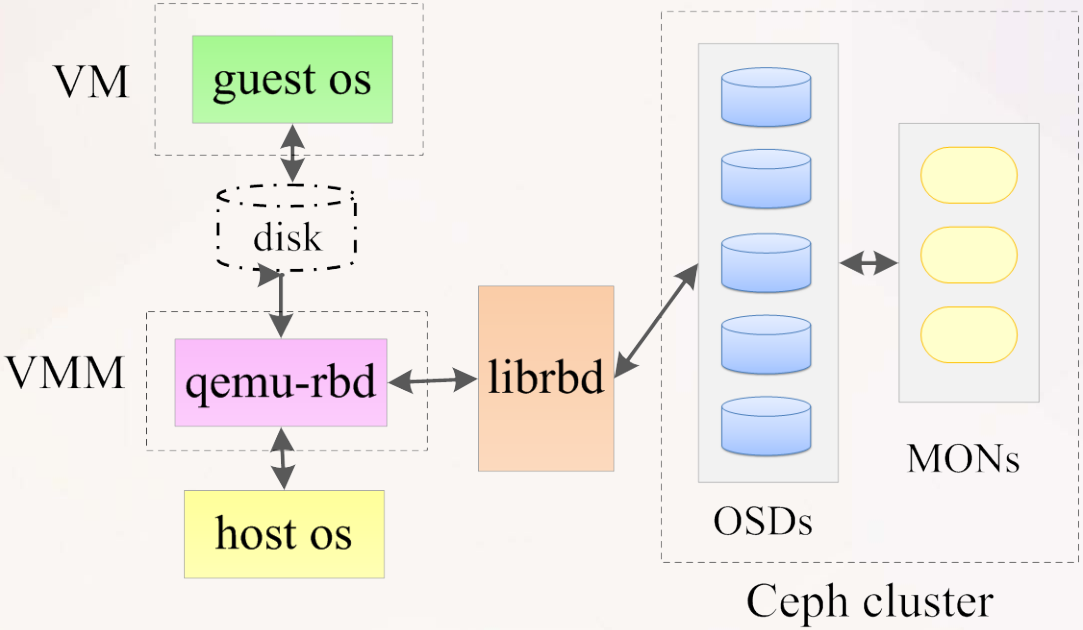
Ceph Block Solution

- rbd
  - 在host上导出一个block device
  - 内核实现



Ceph Block Solution

- qemu-rbd
  - qemu的一个block driver, 在vm中导出一个磁盘
  - guest os对该磁盘的访问请求会传递给qemu-rbd, qemu-rbd调用librbd转化为Ceph的访问
  - Bypass host kernel disk IO stack



Need Another Solution

- rbd
  - 需要在kernel实现librbd所有功能, 由于应用比qemu-rbd少,开发比librbd要滞后
  - 内核模块出错影响域大
  - 内核模块的灵活性和可移植性均不是很好
  - 在非x86体系结构的平台上测试不够充分
- Container Scenario
  - No qemu, thus no qemu-rbd
- VM live backup
  - described later

传统NBD工作流程

- nbd kernel module
  - 导出块设备文件/dev/nbdx
- nbd client
  - 创建一个socket与nbd server建立连接
  - 对/dev/nbdx调用NBD_SET_SOCK ioctl接口将该socket信息传入kernel
  - 调用NBD_DO_IT ioctl接口，进入nbd kernel module的代码, 创建一个kernel thread，然后在内核监听socket, 收取处理nbd server响应
- nbd kernel thread
  - 接收块设备访问请求，然后通过socket将请求发送给server
- nbd server
  - 将数据访问请求转化为对服务器本地磁盘或文件的访问
  - 将响应通过socket发送给nbd client

NBD

- nbd在host导出一个block device
- 对该device的所有访问请求都转发给用户空间程序处理
- nbd提供一个块设备驱动框架，可支持在用户空间实现块设备，类似FUSE支持在用户空间实现文件系统
- Can we implement a Ceph block solution based on nbd

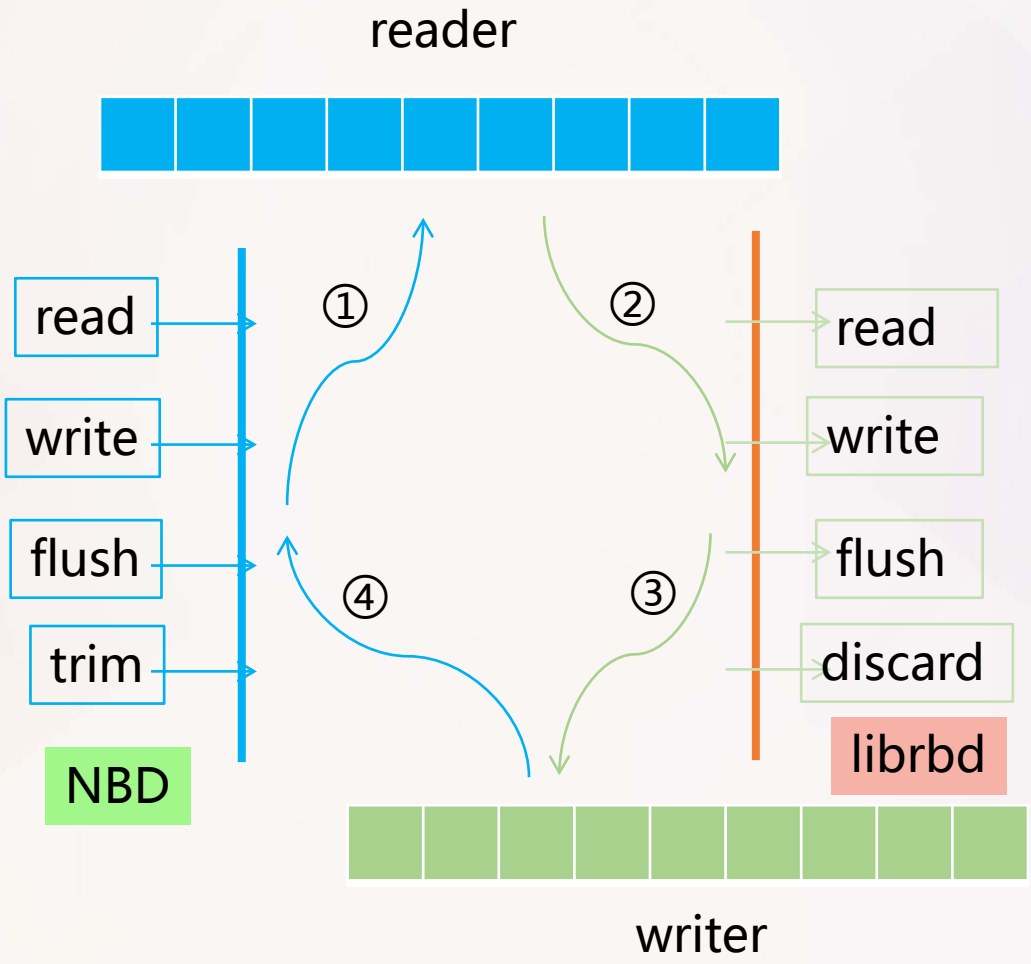
rbd-nbd

- 在用户空间实现一个rbd-nbd程序，即nbd client, 启动时创建一个socketpair, 两个socket分别为csock, dsock. 将csock利用NBD_SET_SOCK ioctl接口传递给nbd kernel module
- nbd kernel module创建nbd kernel thread在队列一上睡眠
- nbd client创建nbd server, nbd server创建两个线程, reader, writer. reader监听dsock, writer在队列二上睡眠
- nbd client调用NBD_DO_IT ioctl接口进入kernel监听csock

rbd-nbd

- 应用对/dev/nbdx的访问请求由nbd kernel module放到队列一，唤醒nbd kernel thread
- nbd kernel thread将请求通过csock发送给reader
- reader将请求放到队列二,调用librbd将请求发给ceph
- librbd将响应放到队列三, 唤醒writer
- writer将响应通过dsock发送给nbd client
- nbd client处理响应

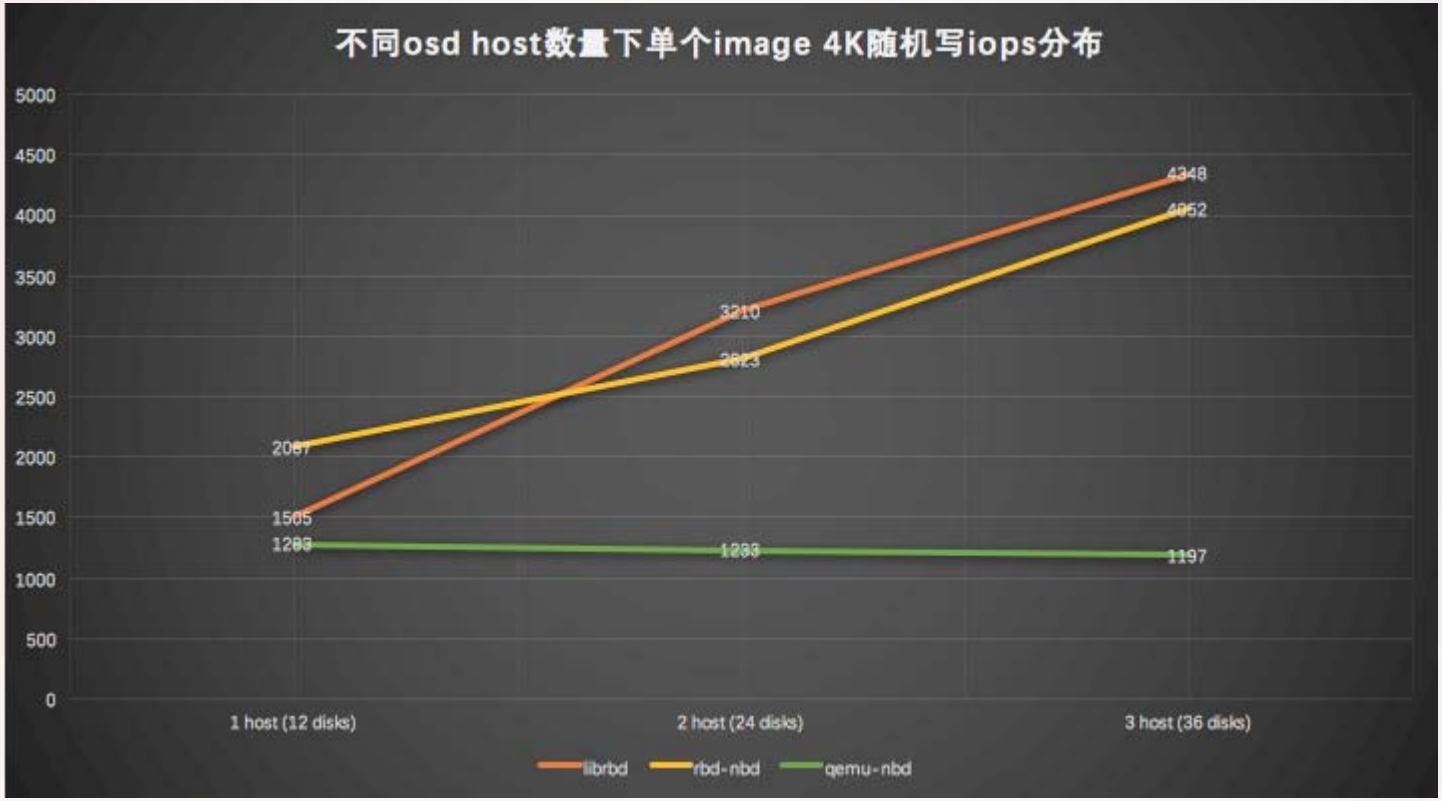
rbd-nbd


reader
read
write
flush
trim
NBD
①
②
③
④
read
write
flush
discard
librbd
writer

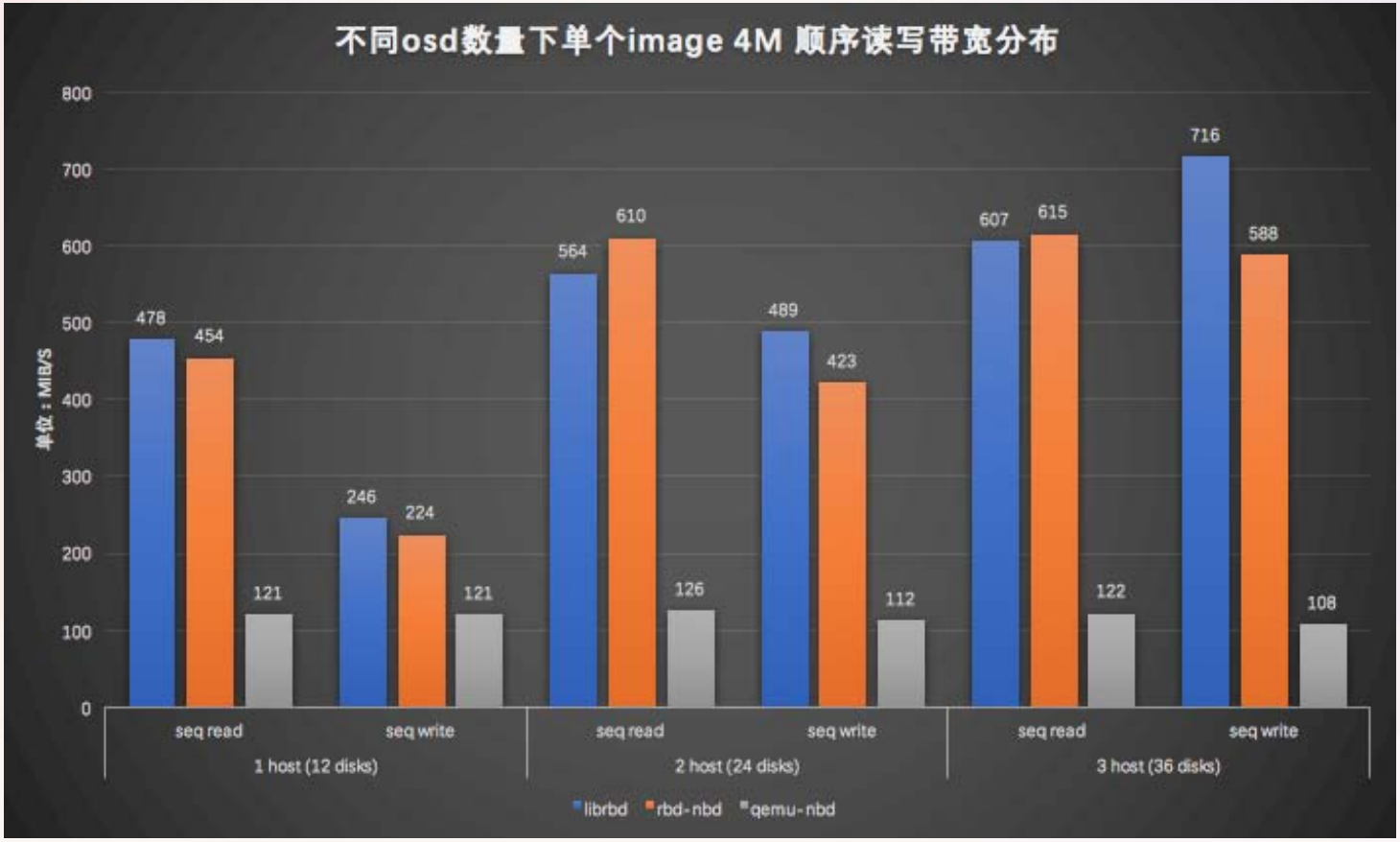
性能测试

- 环境
  - CPU: 2 Intel Xeon E5-2630 v4 @ 2.2G
  - Memory: 128GB
  - 12*8T SATA HDD
  - fio
    - ioengine: 测试librbd为rbd, 测试qemu-nbd, rbd-nbd为libaio
    - iodepth: 256
  - Ceph
    - Luminous v12.2.2

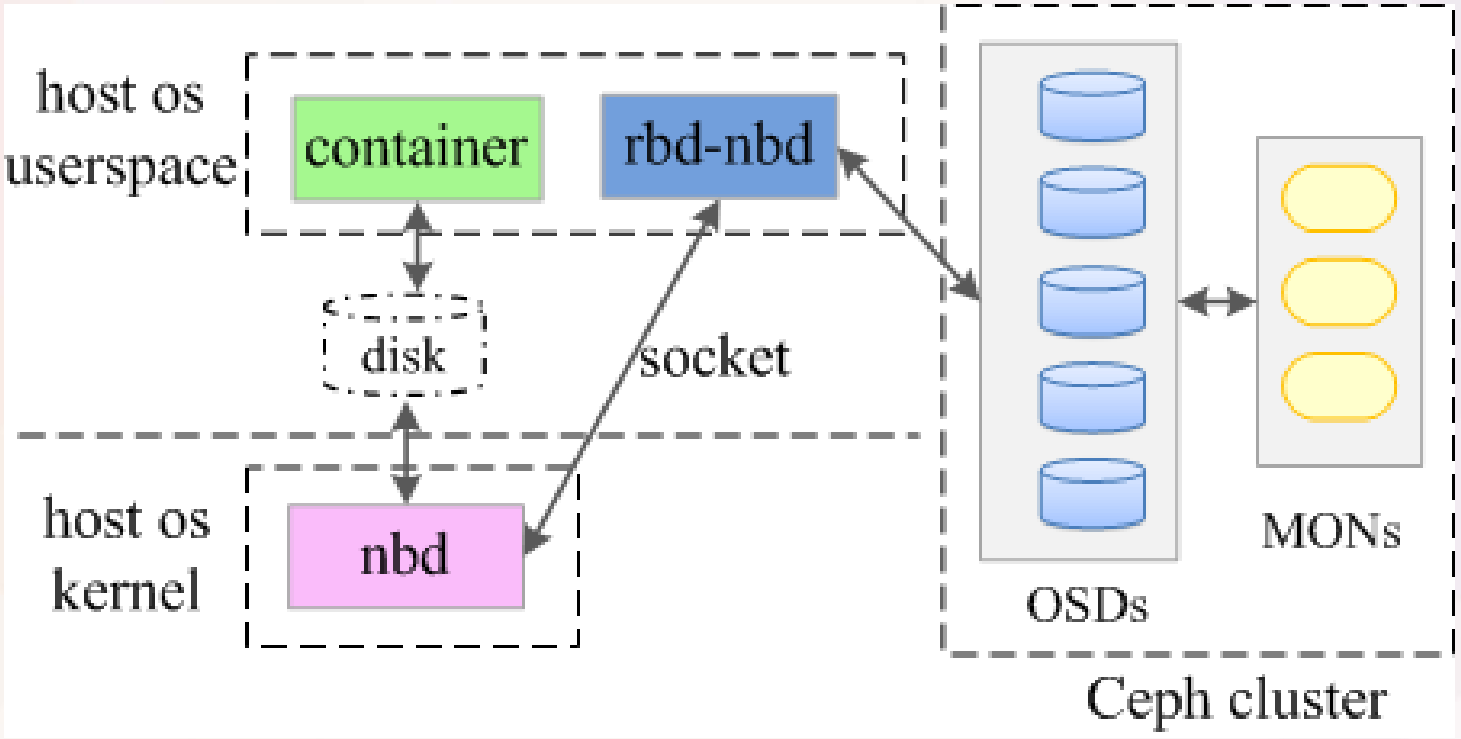
Random Write



Sequential Read Write



Container Scenario



VM Live Backup

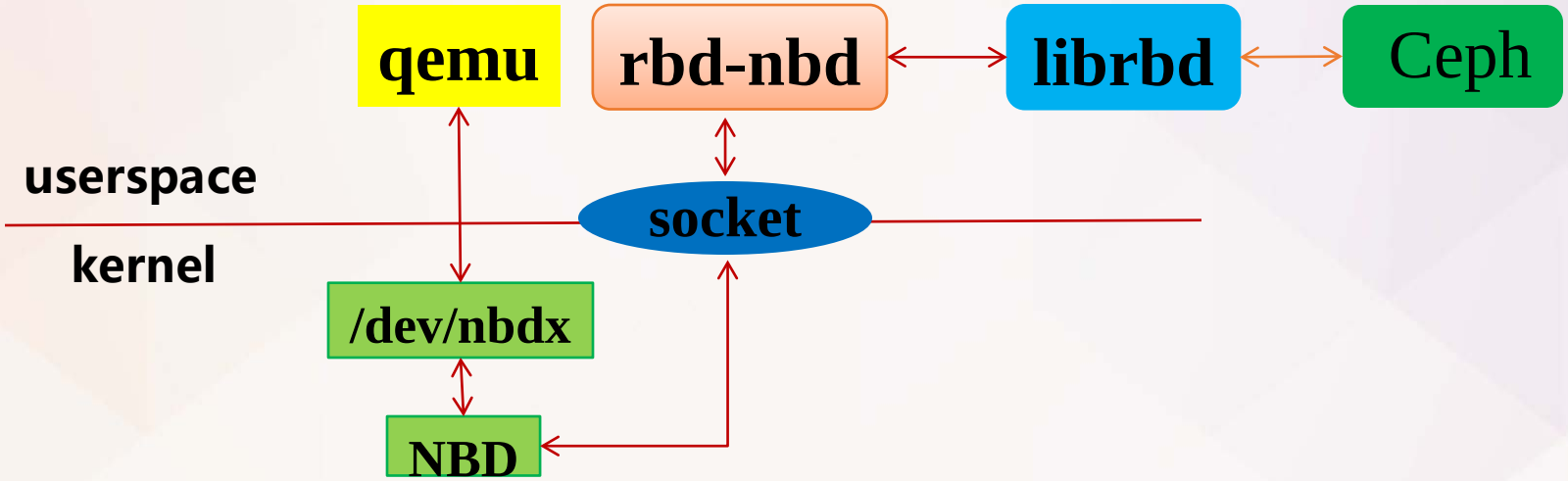
- Qemu Drive Mirror
  - 将vm的本地镜像实时同步到Ceph
  - 可作为vm本地盘的备份，以及快速热迁移的支持
- qemu-rbd的问题
  - 稀疏问题
    - qemu-rbd对qemu看到的是一个raw格式的image，且没有实现支持稀疏查询的回调，所以通过qemu-rbd备份到Ceph的image，再拷贝回本地不是稀疏的
  - 紧耦合问题
    - 跟qemu进程紧耦合，如果网络或ceph服务出问题，drive mirror会hang住，影响用户vm的关闭，重启，快照等功能。要解决只能kill qemu, 这会使用户vm重启

rbd-nbd的优势

- 支持稀疏
  - 对qemu看到的是本地/dev/nbdx文件, 不感知Ceph的存在, 可以把/dev/nbdx文件指定为qcow2格式, 这样写到Ceph上的是qcow2格式内容, 拷贝回本地时也是qcow2格式的, 是稀疏的
  - 由于qcow2有元数据, 创建rbd的时候size需要比磁盘容量稍大, 一般元数据在1GB以内
  - rbd是thin provisioning的,所以指定更大的size不会浪费空间

rbd-nbd的优势

- 松耦合
  - 跟qemu进程松散耦合，两者通过socket通信，如果网络或Ceph服务出问题，只需kill rbd-nbd进程，qemu会返回eio



Conclusion

- rbd-nbd
  - 与rbd的性能接近
  - 用户空间实现, 出错影响域小
  - 利用librbd, 功能强大稳定
  - nbd代码简单, 在kernel已有很长时间, 良好的稳定性
  - 良好的可移植性
  - 简单易于维护

谢谢!