UCLOUD 优刻得

基于SPDK的UDisk全栈优化

块存储团队

杨昱天

大纲

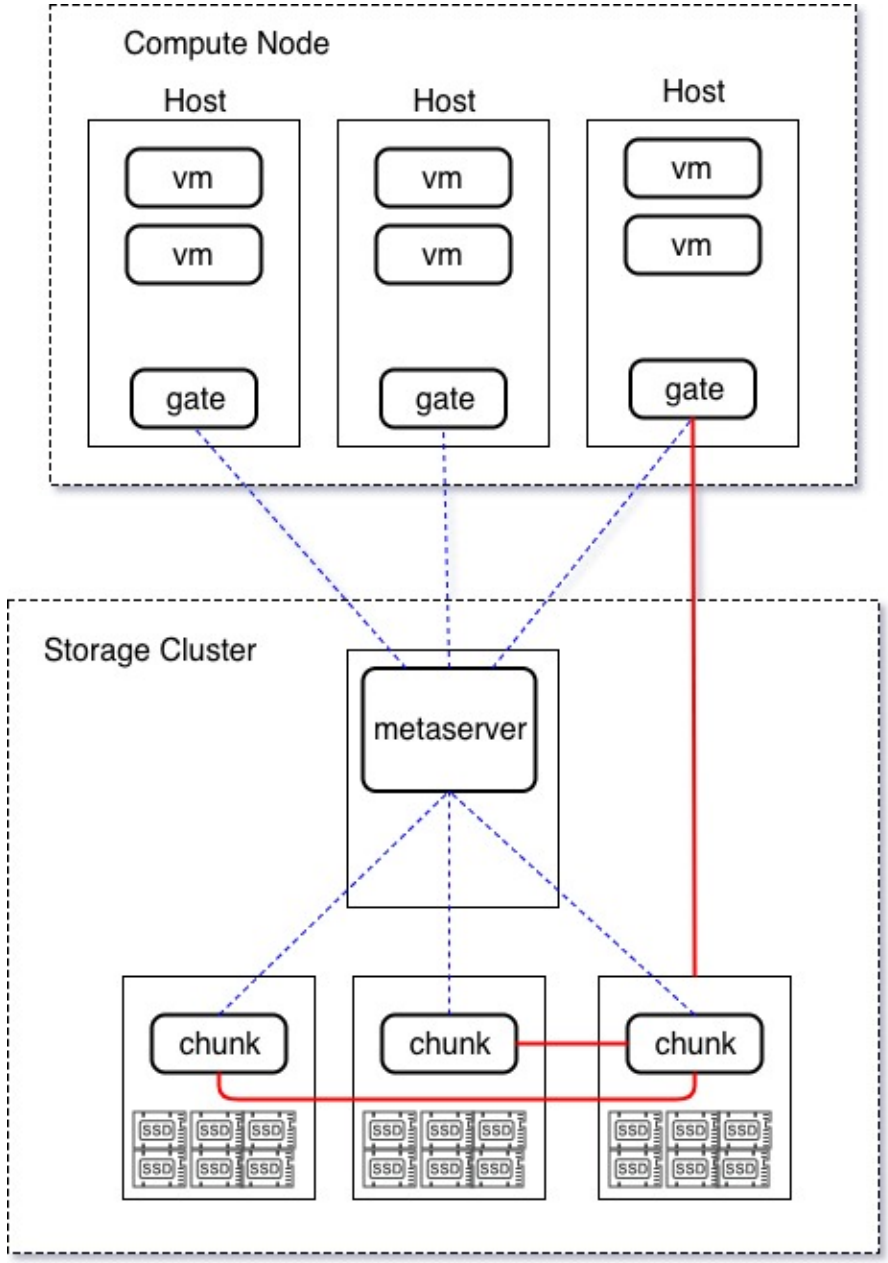
- UDisk架构介绍和性能目标
- 结合SPDK Vhost优化IO接入层性能
- RDMA网络加速
- 使用SPDK NVMe Driver加速持久化层

UDisk产品概述

UDisk(UCloud Disk)向虚拟云主机提供块存储服务，采用分布式架构，通过多副本冗余保证数据持久性。当前的SSD云盘虽然可靠性高，能够宕机快速迁移，但性能上相对本地SSD盘没有优势。我们希望作出一款同时高可靠，高性能的产品，让客户使用到快速可靠的存储。伴随硬件升级，包括25Gb网络、NVMe SSD，软件也必须要有能力充分发挥出硬件性能。

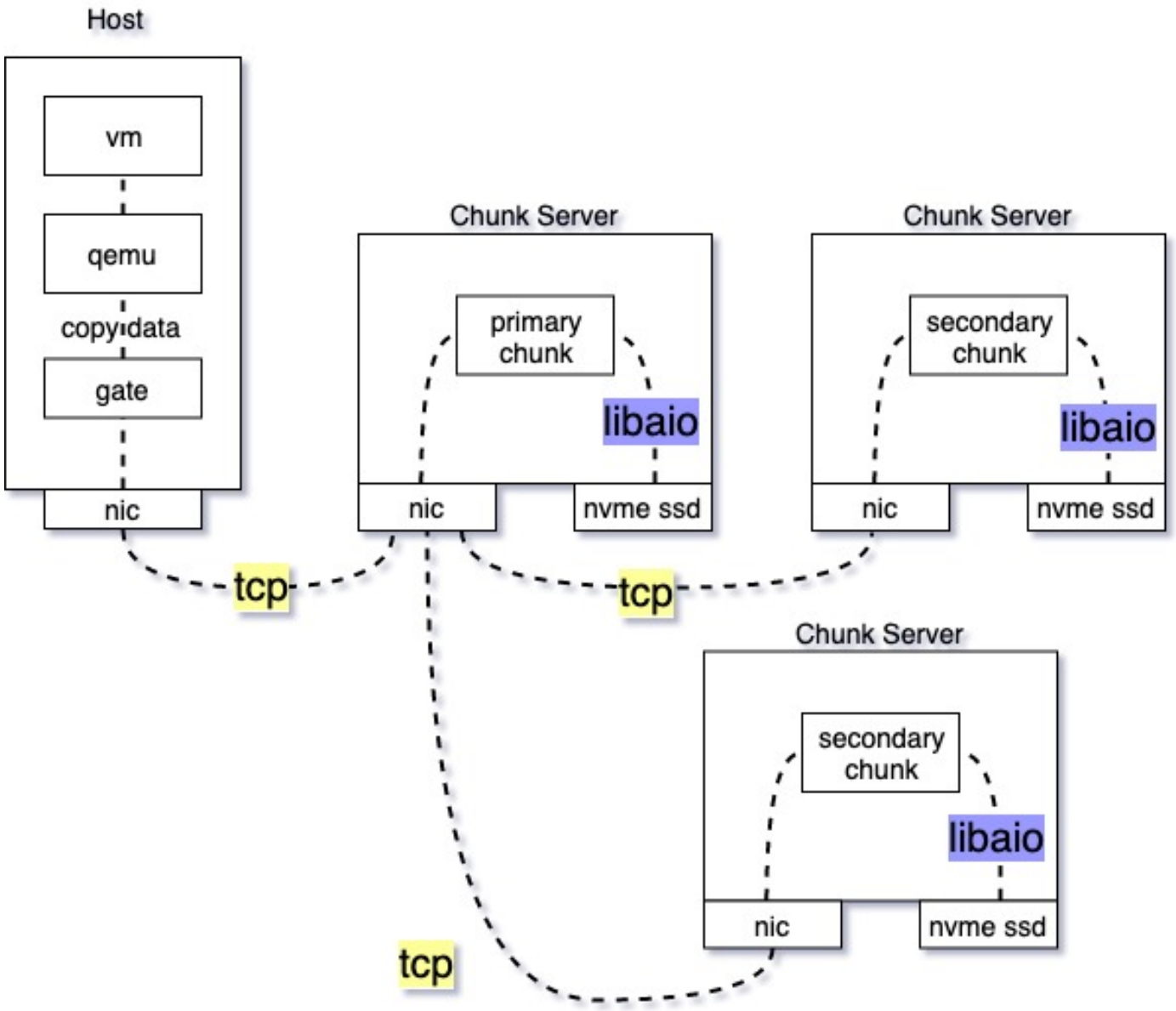
UDisk架构

- Gate：宿主机上的UDisk Client，将虚机IO转发至后端存储节点完成持久化
- Chunk：提供块存储IO服务
- Metaserver：提供路由表的元数据服务



UDisk IO路径

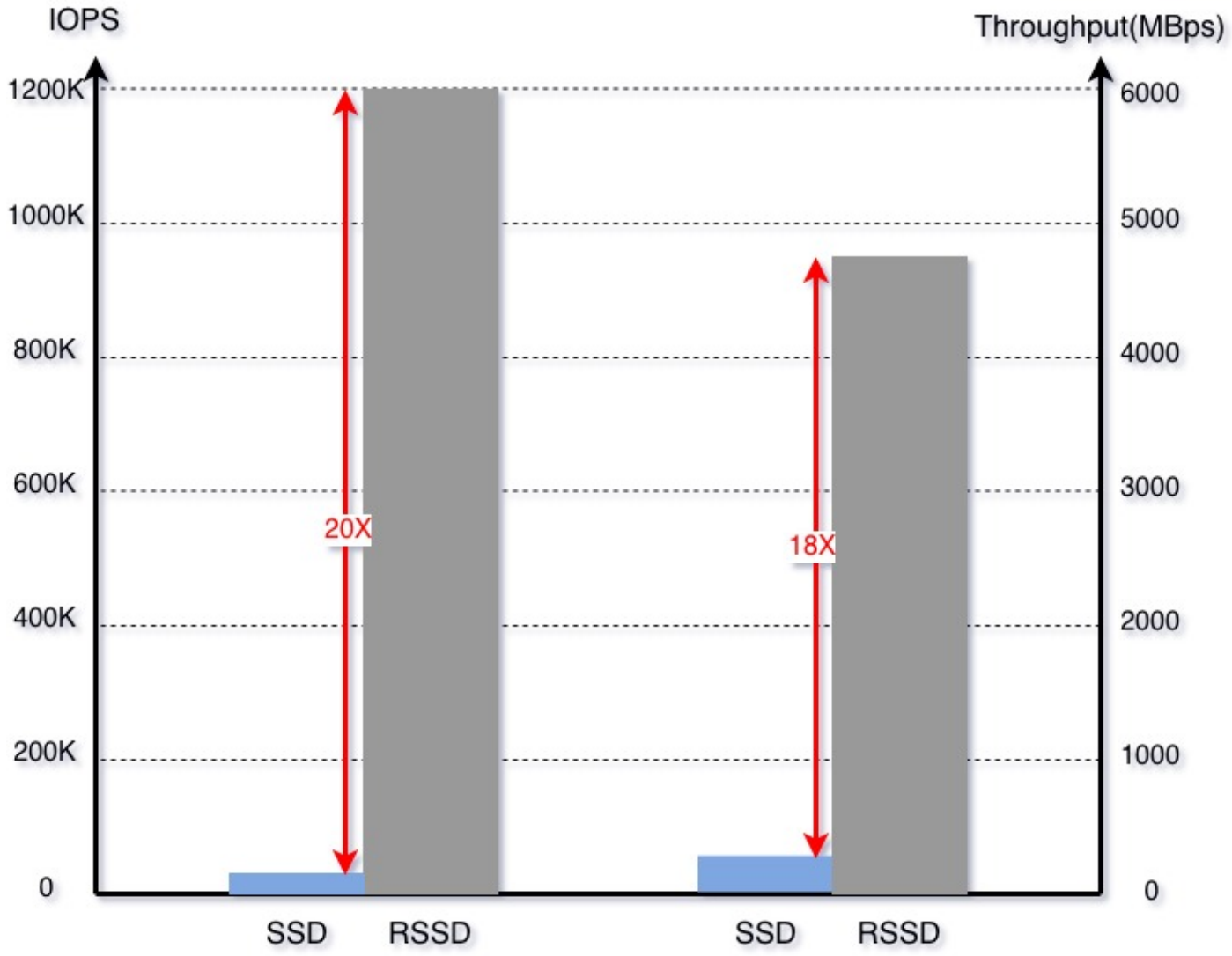
- 虚机IO经过QEMU转发
- QEMU到Gate进程转发IO
  存在拷贝
- TCP网络
- Libaio方式读写NVMe SSD



性能目标

优化前云盘提供
最大24000 iops
最大吞吐带宽达到260MBps
时延500us

优化后云盘提供
最大120万IOPS
最大吞吐带宽达到4.8GBps
时延100us

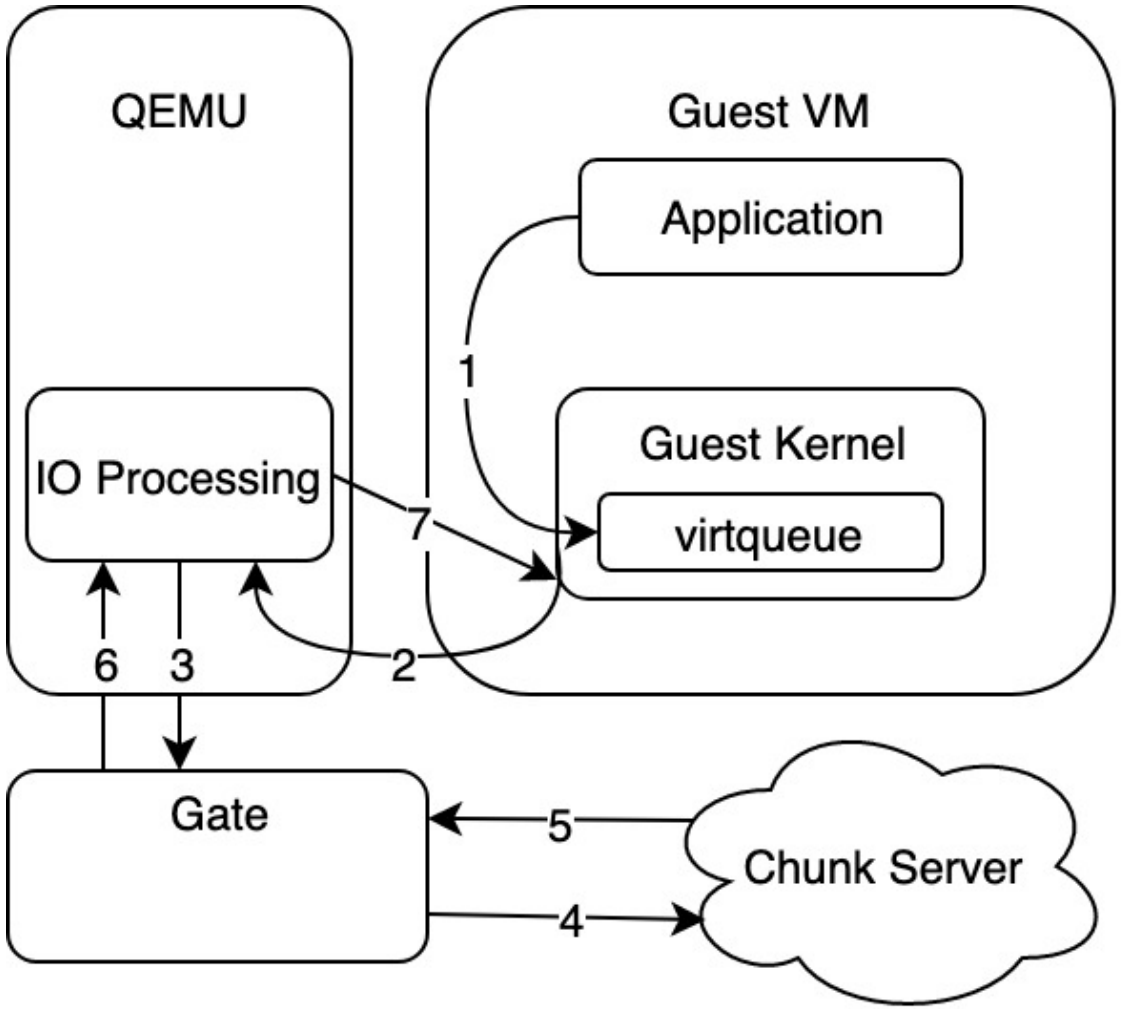


IO接入层瓶颈

- IO经过QEMU的UDisk驱动层， QEMU里的UDisk驱动层通到Gate监听的unix domain socket的转发IO请求，存在额外的拷贝开销

- 零拷贝
- IO路径bypass QEMU

目前采用的QEMU Virtio方案



1：提交IO到virtqueue

2：通过ioeventfd机制通知QEMU（中断）

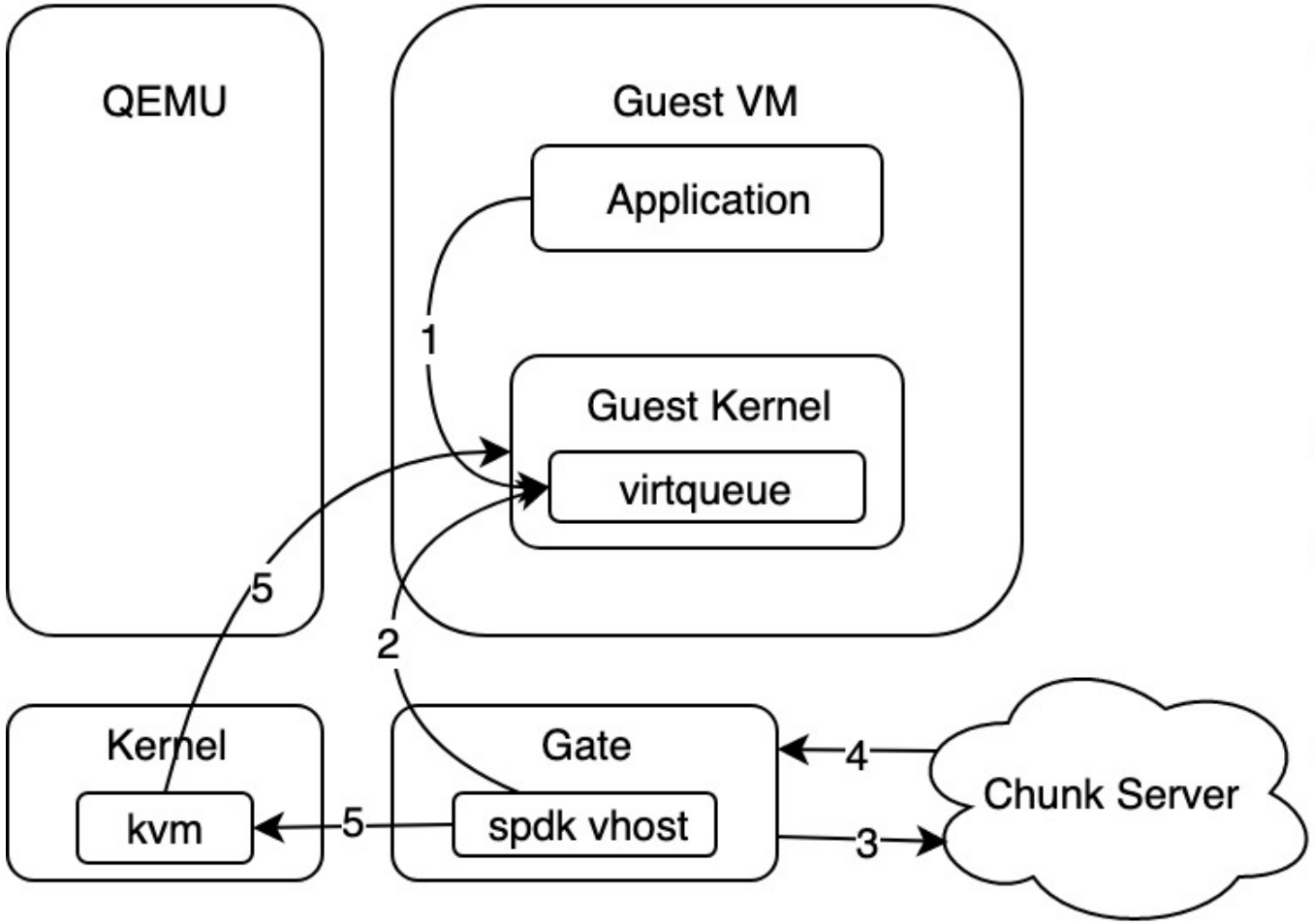
3：QEMU将IO通过unix domain socket转发给Gate（拷贝）

4-5：后端存储集群处理来自Gate的IO请求

6：通过socket返回IO完成响应

7：通过irqfd通知Guest IO完成

新设计的SPDK Vhost方案



vhost和Guest共享内存（包括vring和其中desc的buffer，vhost负责将vm_pa翻译到host_va）

1：提交IO到virtqueue

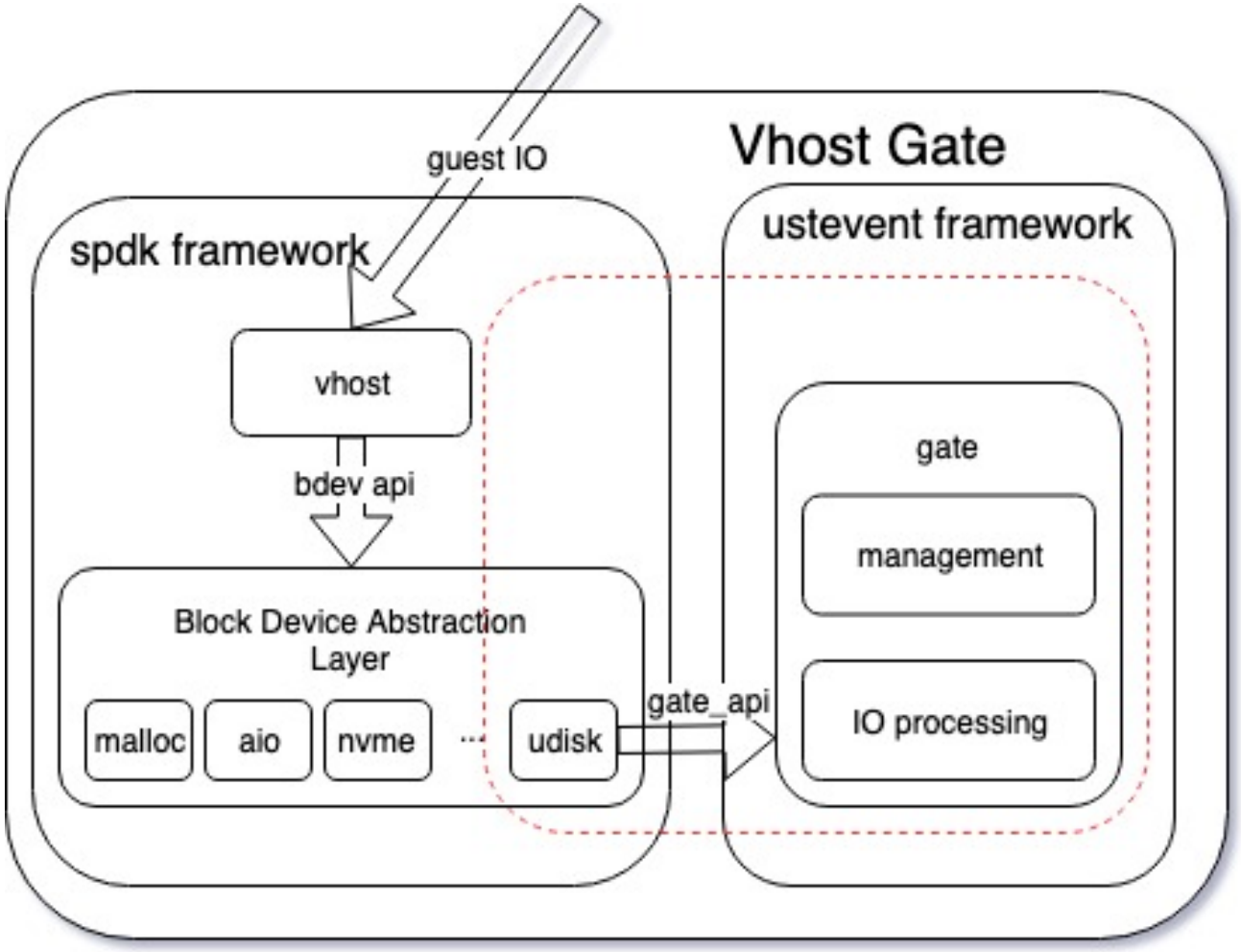
2：轮询virtqueue，处理新到来的IO

3-4：后端存储集群处理来自Gate的IO请求

5：通过irqfd通知Guest IO完成

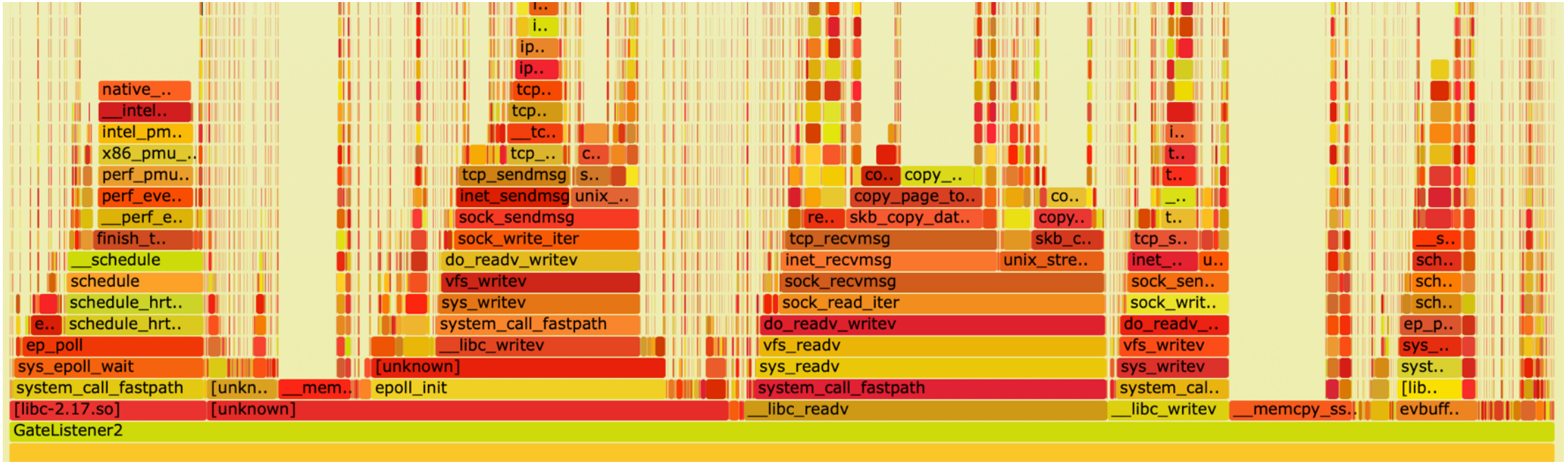
集成SPDK Vhost

- 数据通过共享内存方式传递
- 采用轮询方式，guest向host提交IO没有额外中断开销
- SPDK良好的模块抽象层次，编写自定义的bdev就可以完成Gate和Vhost的对接
- Vhost模块二次开发，支持热升级
- 运维工具的开发，通过SPDK RPC管理Vhost Controller和Bdev，控制每个云盘的QoS



TCP瓶颈

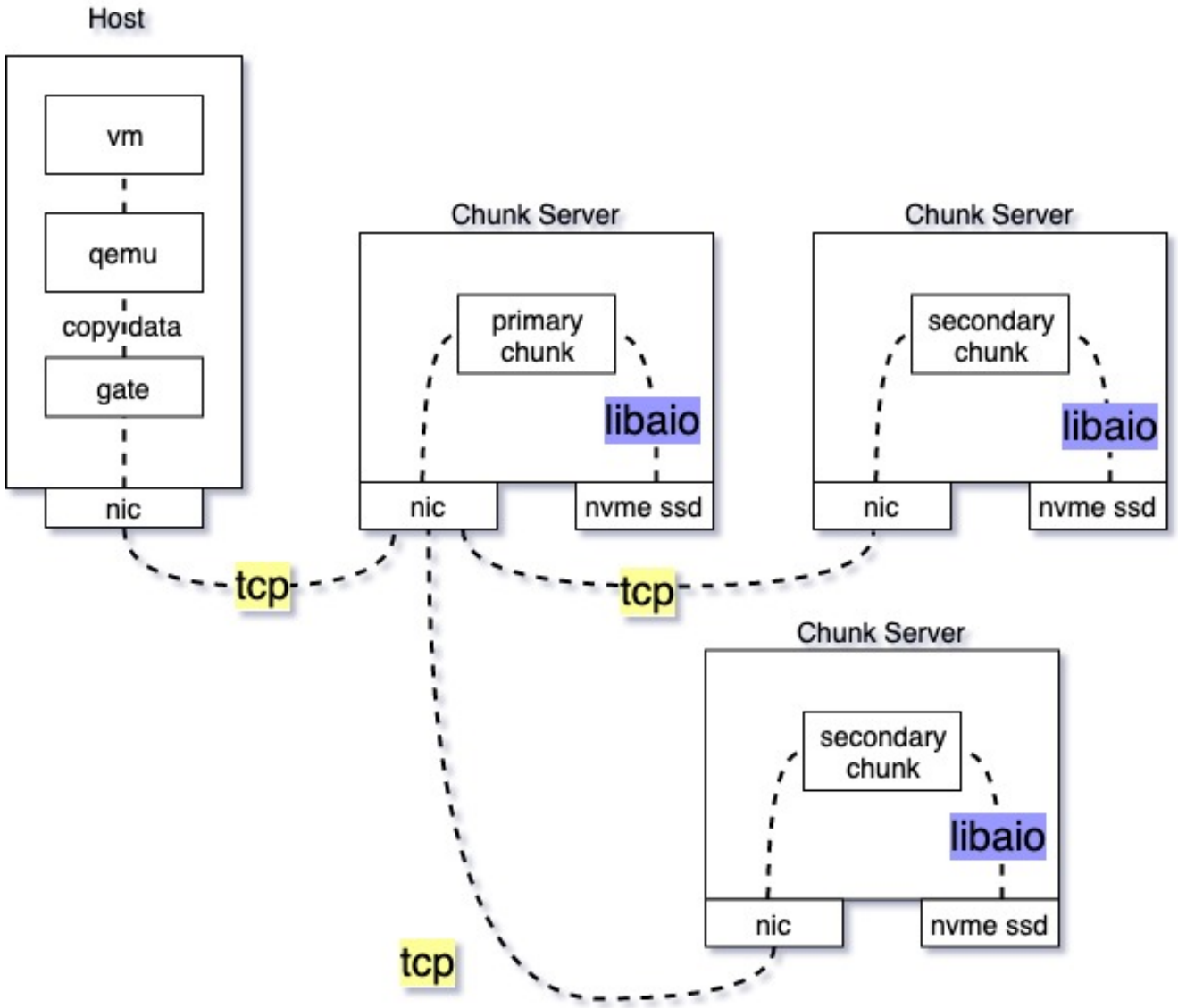
Host侧跑大压力IO测试时，通过perf抓取Gate进程函数热点，TCP协议栈上CPU开销巨大。后端存储节点上也是如此。



TCP瓶颈

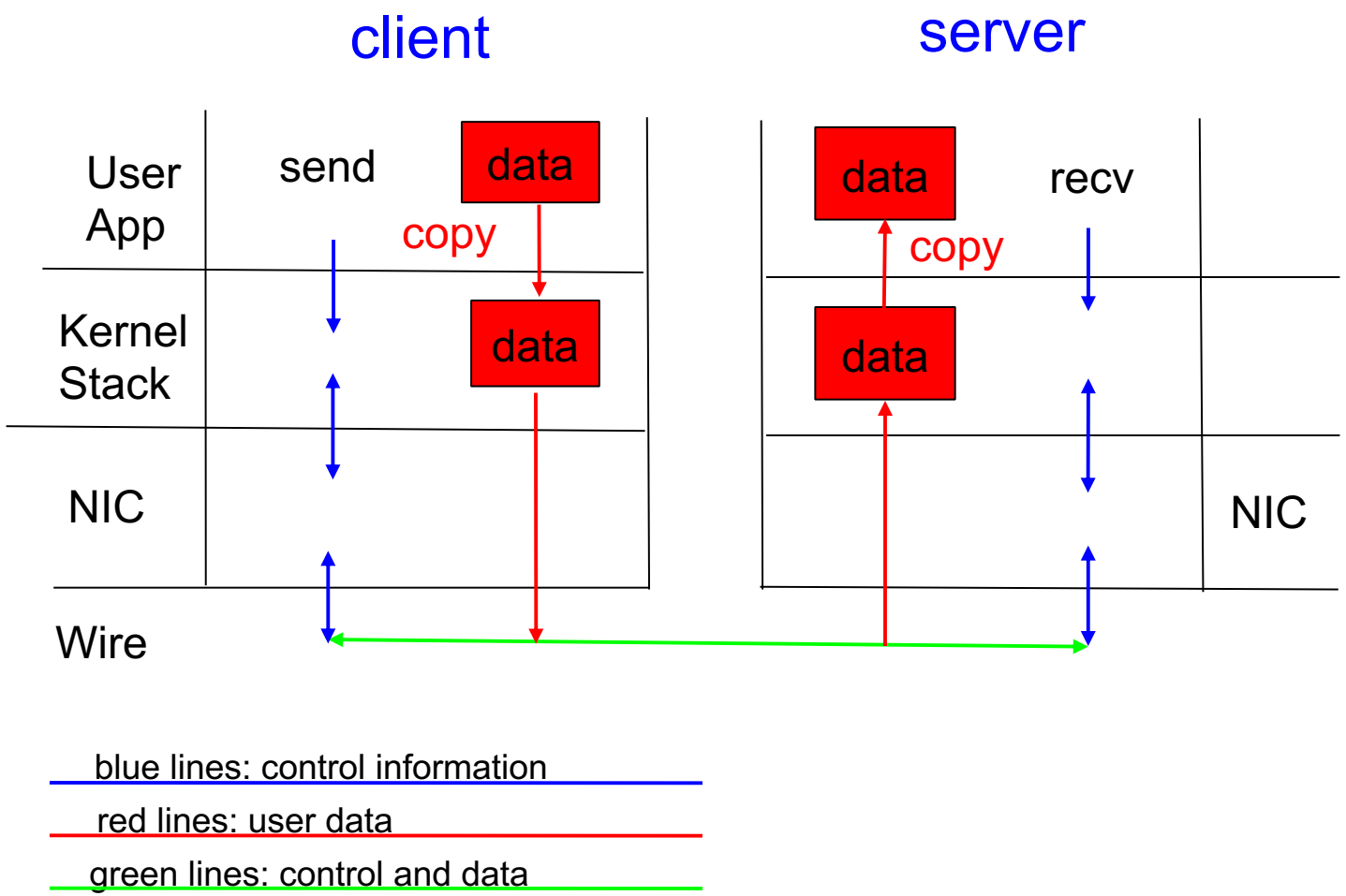
- 高吞吐场景下，70%～80%的CPU都消耗在TCP栈
- Host是计算节点，CPU资源宝贵
- 写IO产生额外两次复制写，时延明显增加

- 高吞吐、低时延、CPU offload
- RDMA



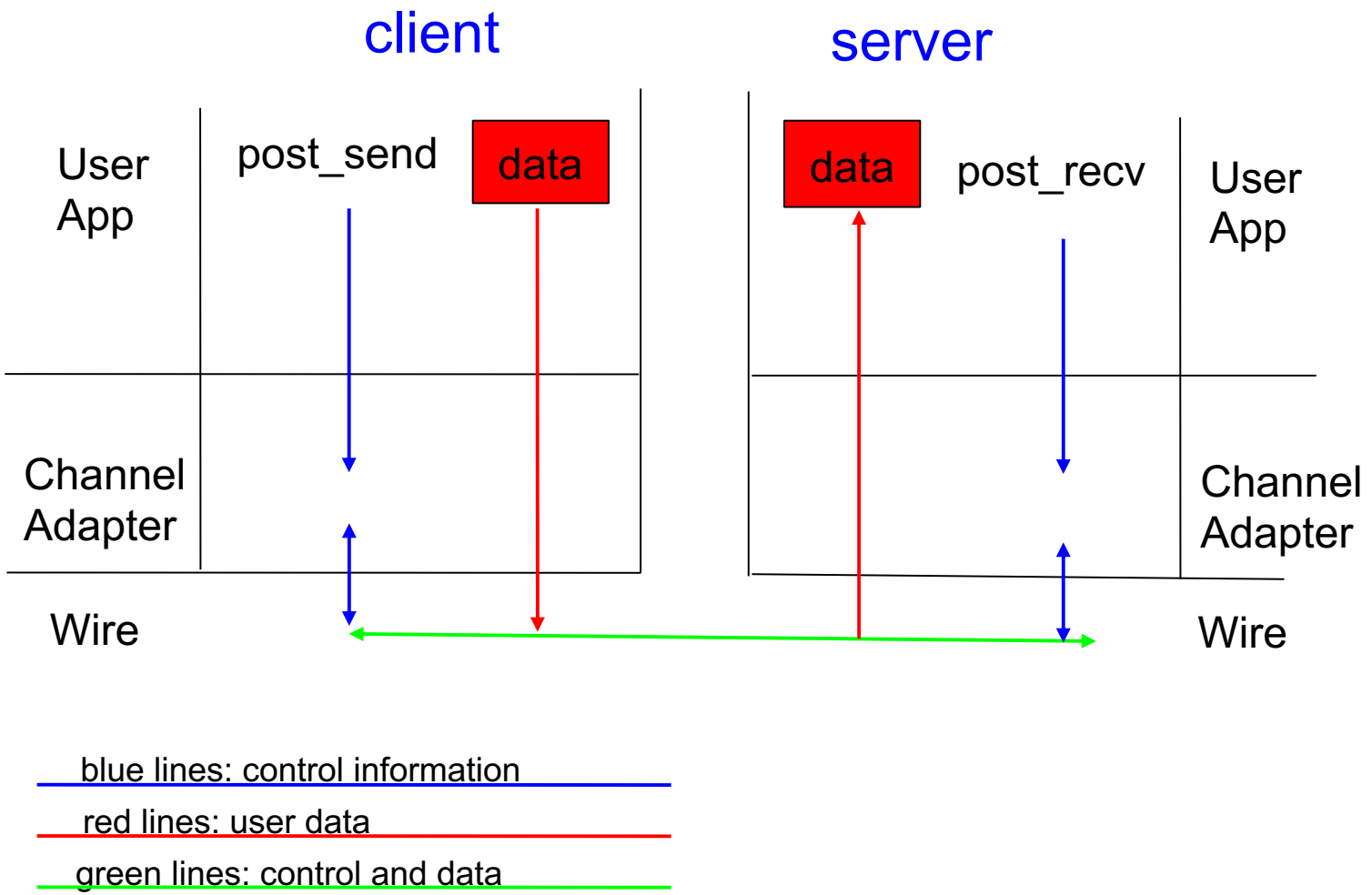
TCP

- 数据从用户态拷贝到内核态
- 收发包流程有上下文切换
- 流式传输
- TCP内核栈
- 高时延，CPU开销大



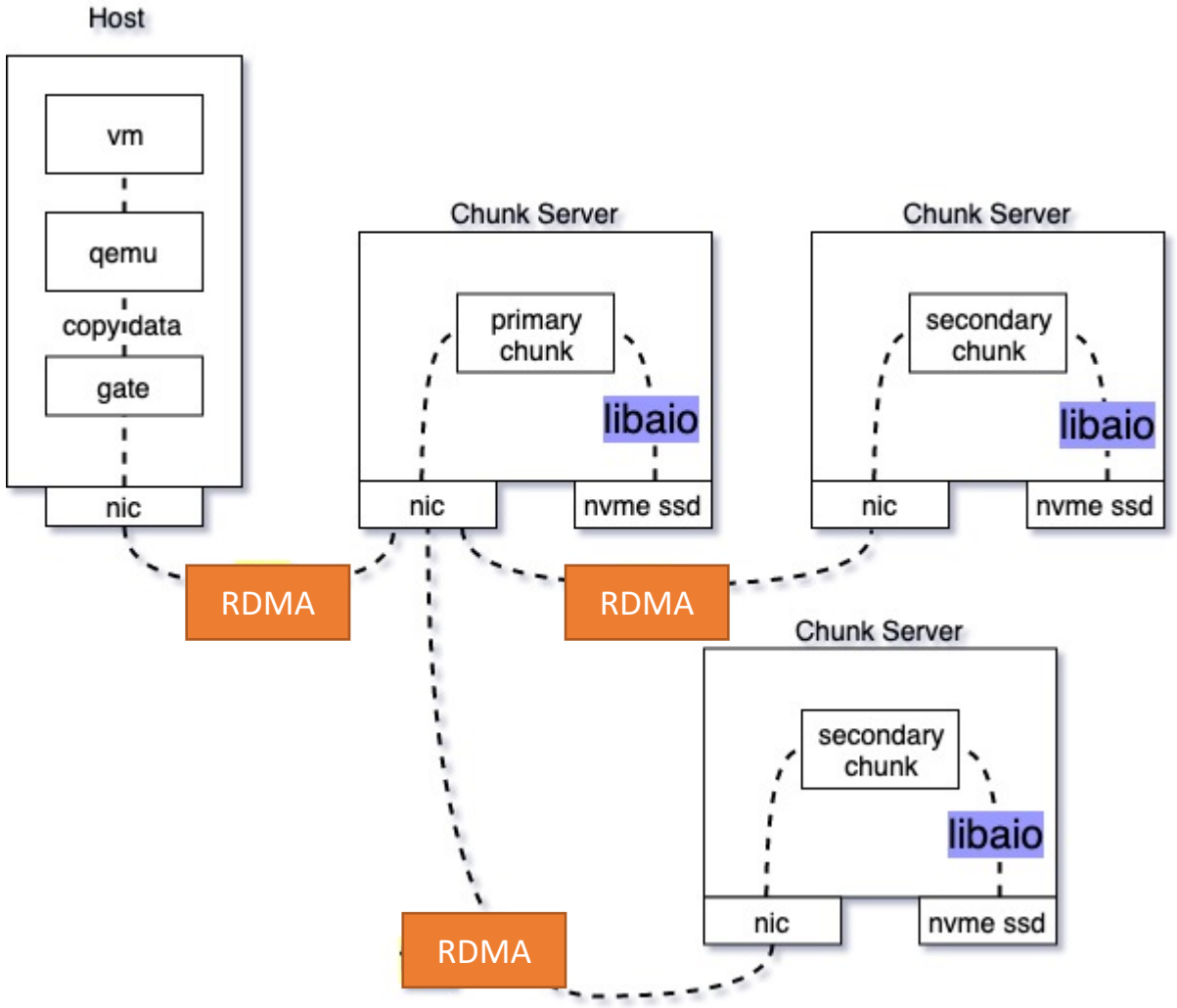
RDMA

- 零拷贝
- 收发包过程没有上下文切换
- 消息式传输
- Kernel bypass
- CPU offload到网卡
- 高吞吐、低时延



RDMA替换TCP

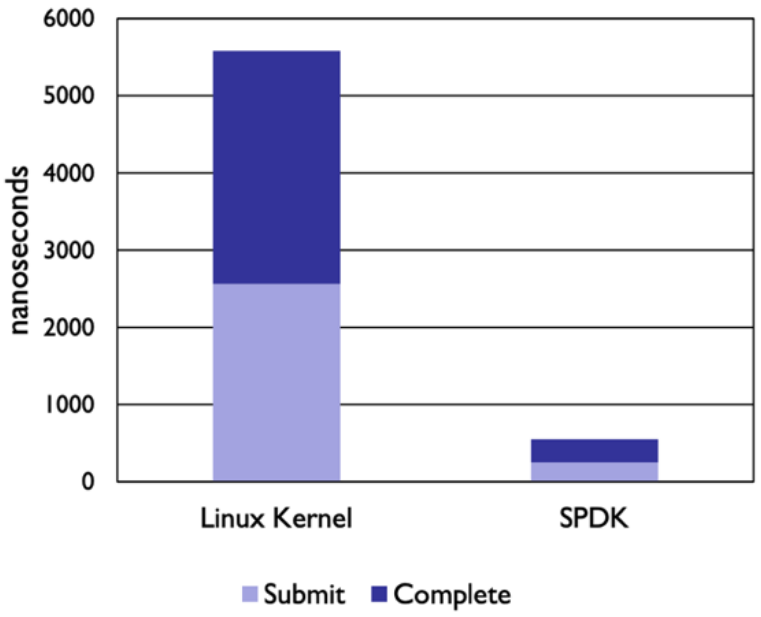
- 计算节点到存储节点RDMA
- 存储节点之间RDMA
- 25Gb双口bonding，提供50Gb吞吐能力（25Gb单口能力无法满足4.8GBps吞吐）
- RoCE v2（基于UDP/IP）



后端持久化层

- 后端存储节点提供更高的IO吞吐能力，存储节点的SSD密度增加
- Libaio with Kernel Driver，ssd性能越来越好，相对地，kernel overhead越来越明显

NVM Express* Driver Software Overhead



Kernel Source of Overhead	SPDK Approach
Interrupts	Asynchronous Polled Mode
Synchronization	Lockless
System Calls	Userspace Hardware Access
DMA Mapping	Hugepages
Generic Block Layer	Specific for Flash Latencies

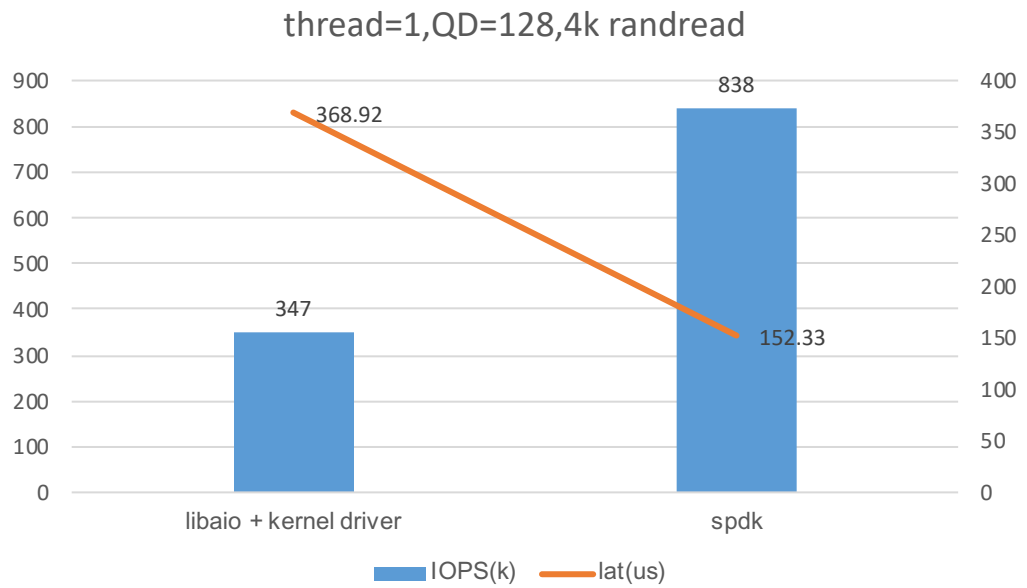
SPDK reduces NVM Express* (NVMe) software overhead up to 10x!

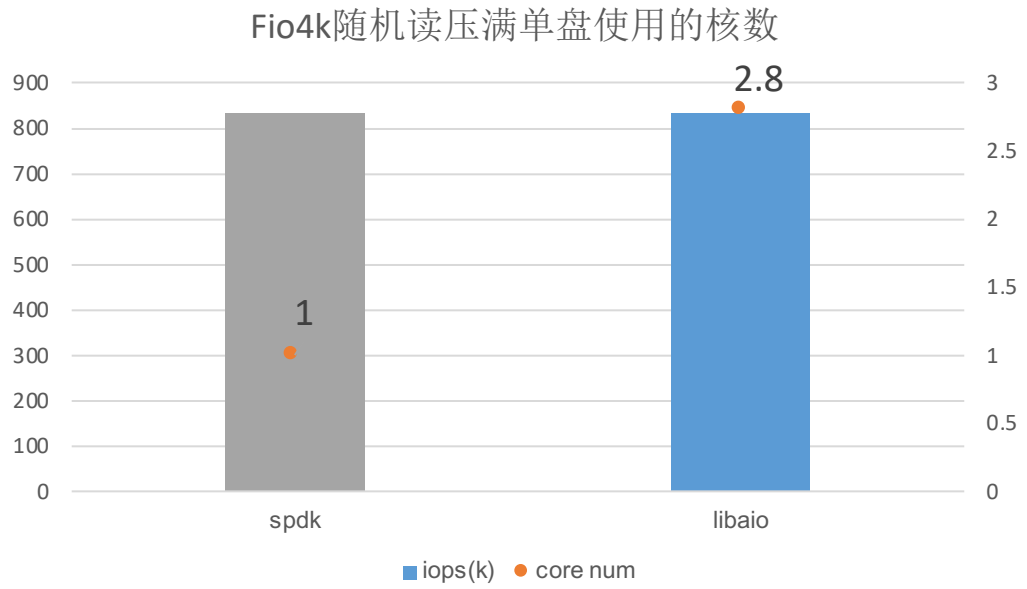
SPDK NVMe Driver实测

- Fio对比测试kernel driver和spdk driver的单核性能

- Libaio with Kernel Driver需要利用3个核才能压到磁盘极限，SPDK NVMe Driver只需要1个核

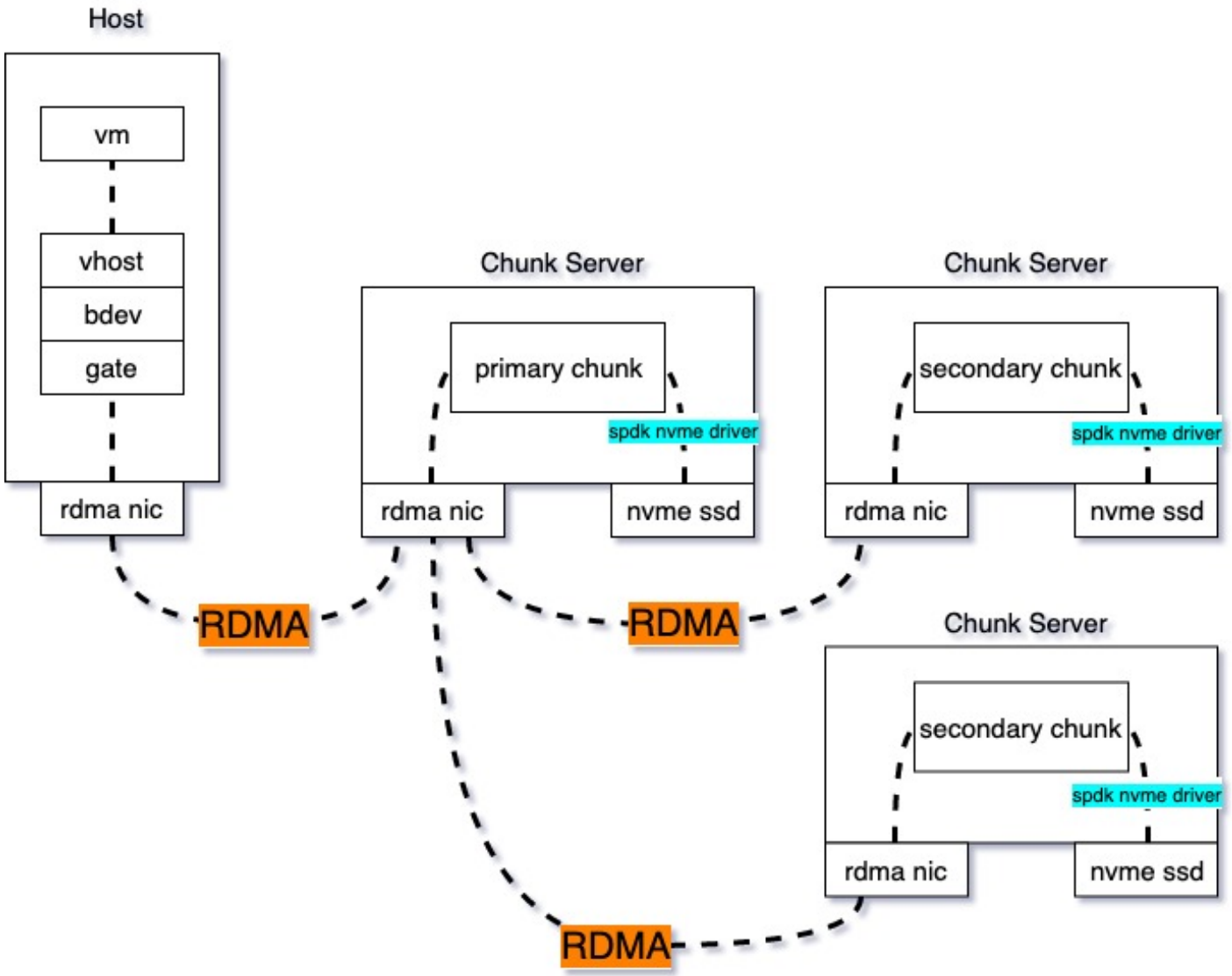
- SPDK NVMe Driver使用轮询模式，同时可以发挥RDMA最佳性能





整体架构优化

- 升级至25Gb网络
- 存储节点ssd盘密度增加
- IO接入层升级，QEMU Virtio 升级至SPDK Vhost
- TCP升级至RDMA
- Libaio with Kernel Driver升级至SPDK NVMe Driver
- IO路径无锁，绑核轮询



总结

- SPDK Vhost加速Host IO Path
- SPDK NVMe Driver加速后端存储IO Path
- RDMA加速网络，解放CPU
- 关键技术分离，升级存量版本，提升用户体验
- 存储开发框架

Q & A

快速定制 • 贴身服务 • U Defined Cloud