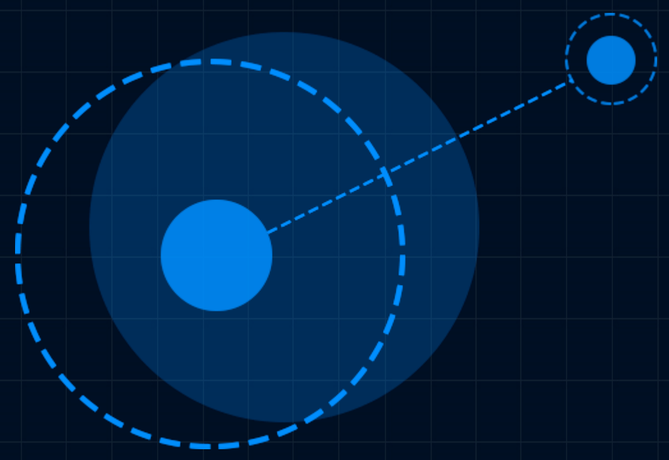
云原生虚拟化方案探索与实践



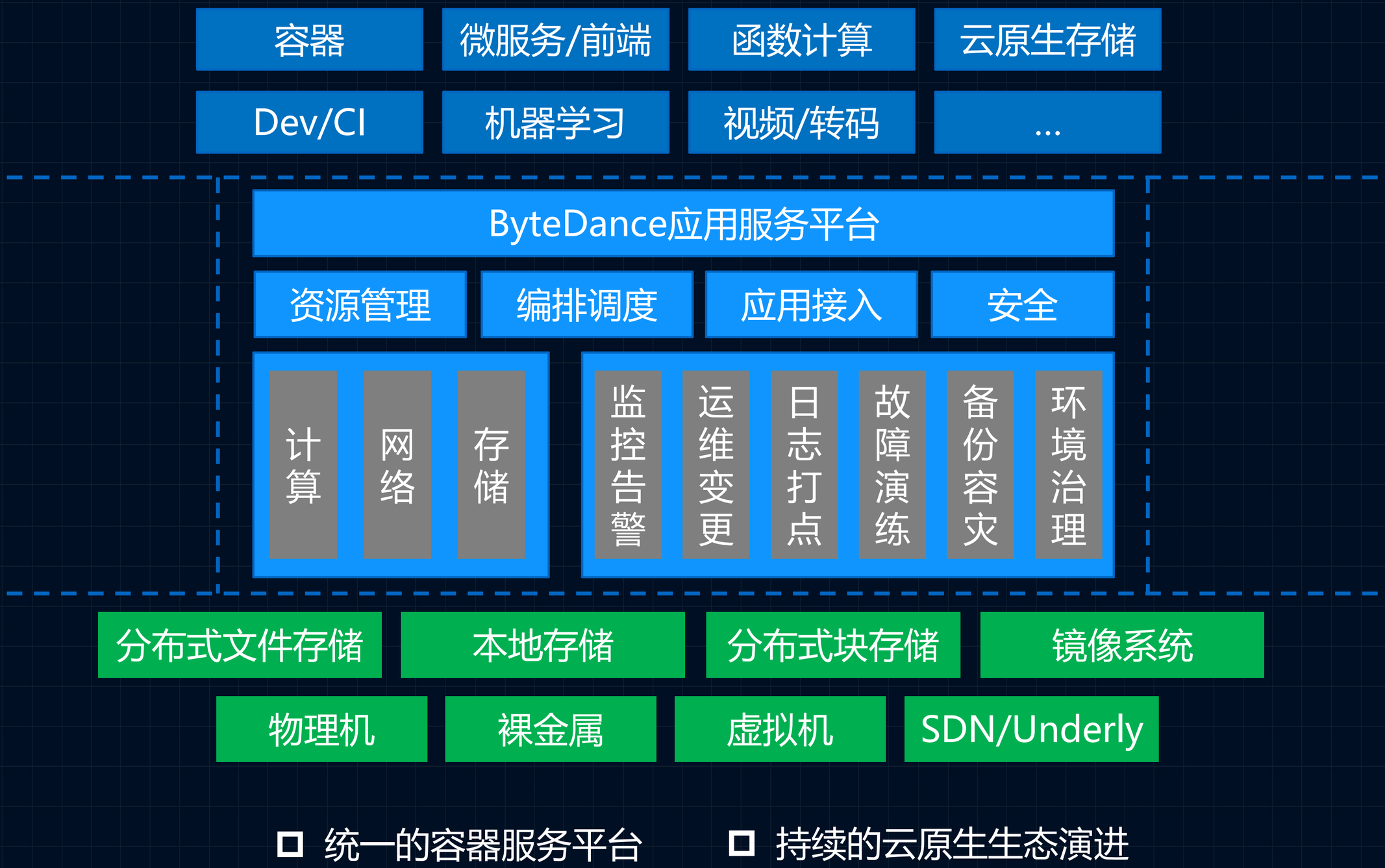
演讲人

字节跳动STE团队 柴稳

目录

背景与思路

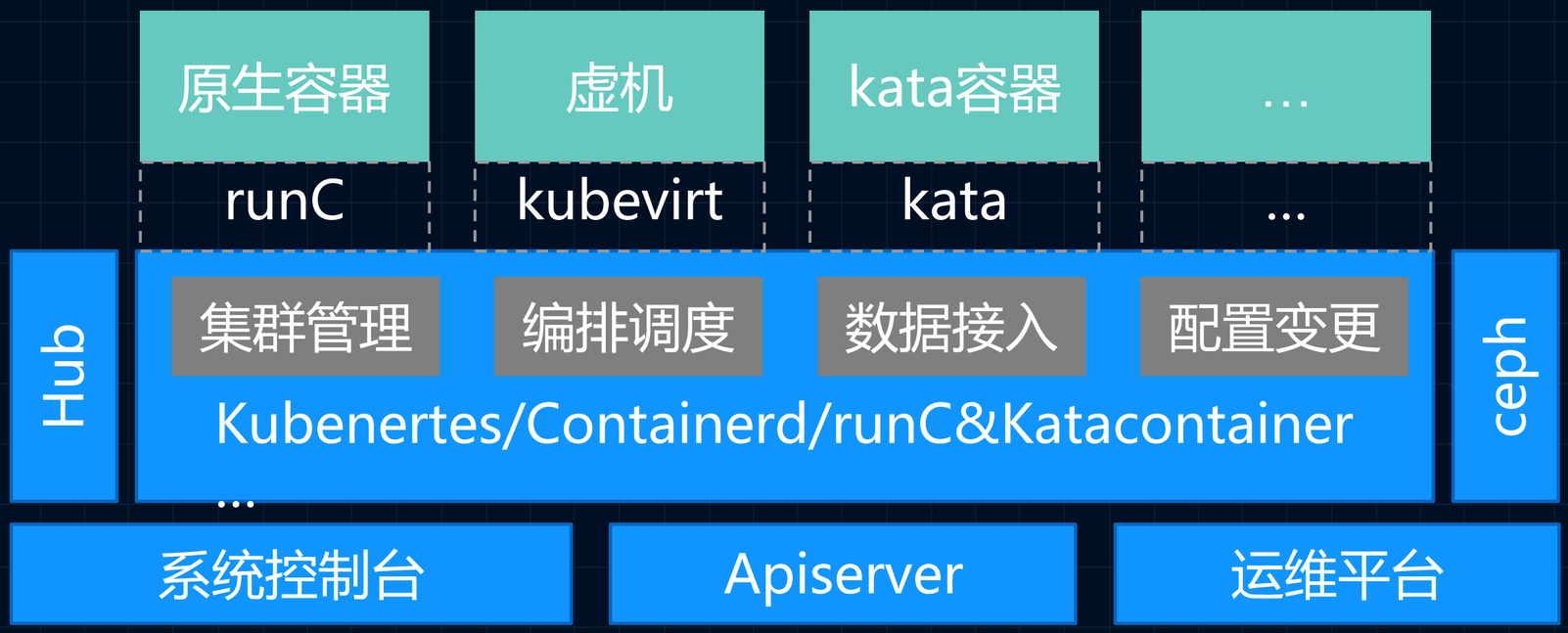
当前业务技术形态



- 统一的容器服务平台
- 持续的云原生生态演进

背景与思路

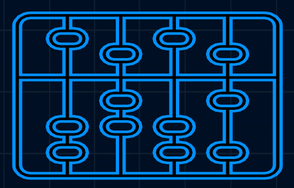
快速验证平台



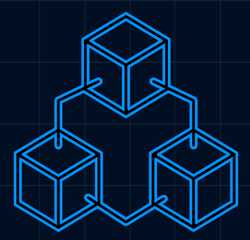
- 功能齐全
- 快速适配
- 灵活变更
- 快速验证
- 半生产环境打磨

背景与思路

探索方向



计算框架



存储/网络接入

集群管理

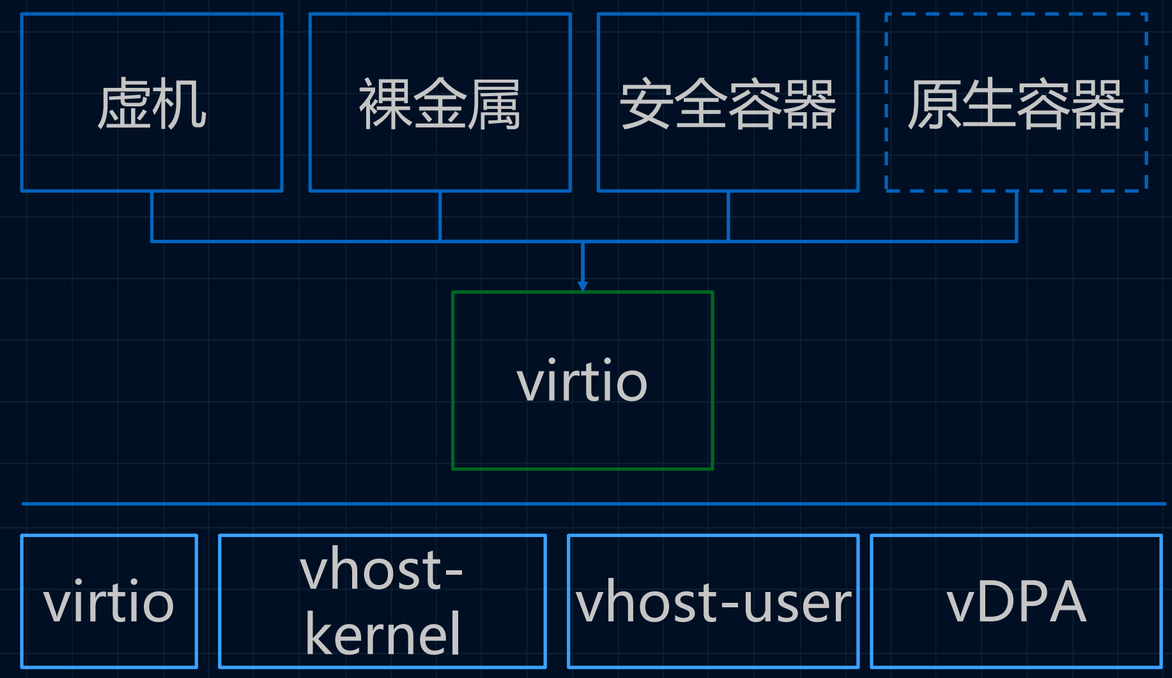
监控运维

目录

存储的探索与实践

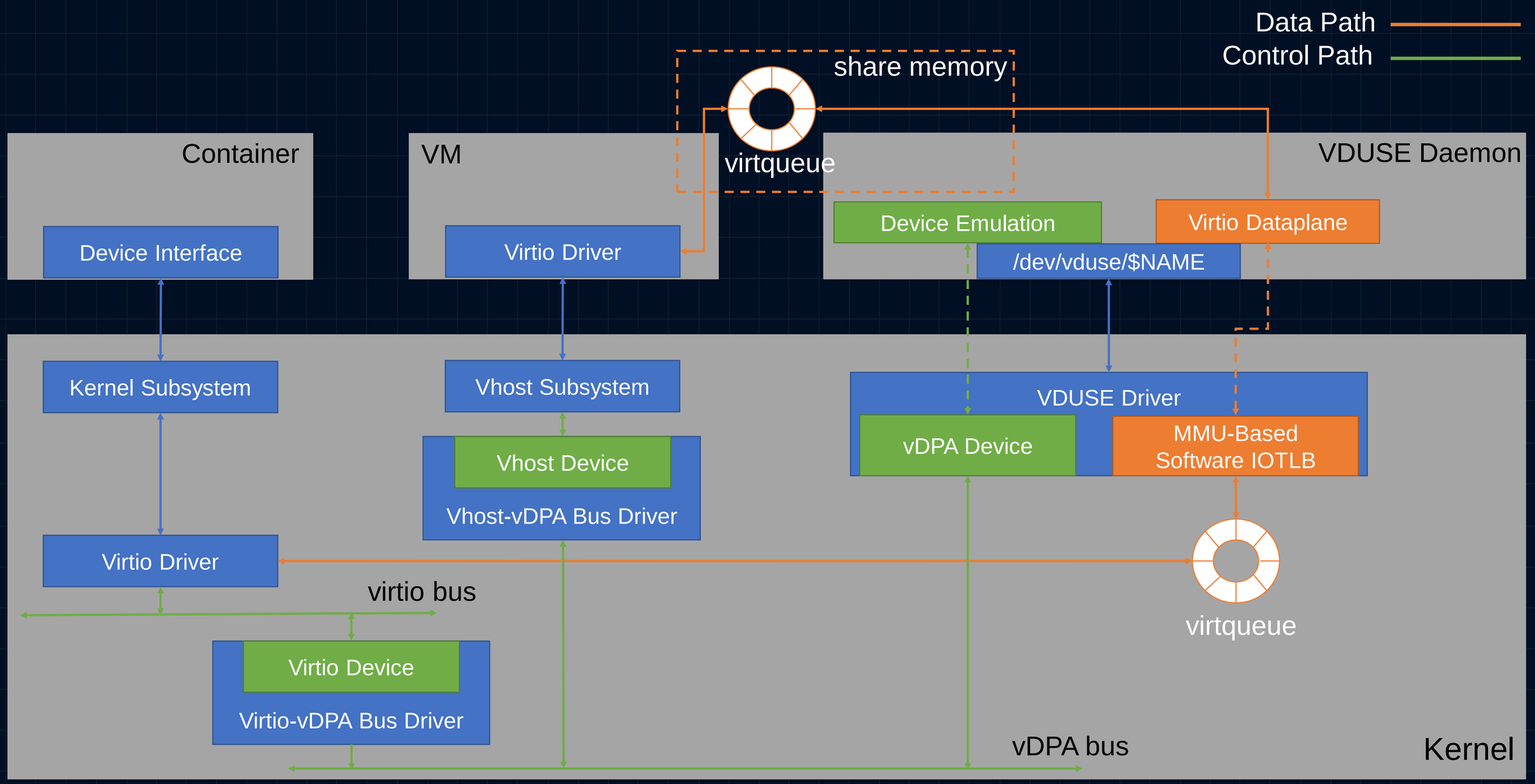
VDUSE的方案介绍

- 现实意义
  - 统一的用户态设备接入方案
  - vDPA硬件设计方案的数据验证
  - 应用视图的一致性
- 实现路径
  - virtio-vdpa/vhost-vdpa 设备框架
  - bounce buffer/shared 内存访问模型
  - kick/irq-injection机制



存储的探索与实践

VDUSE的方案介绍

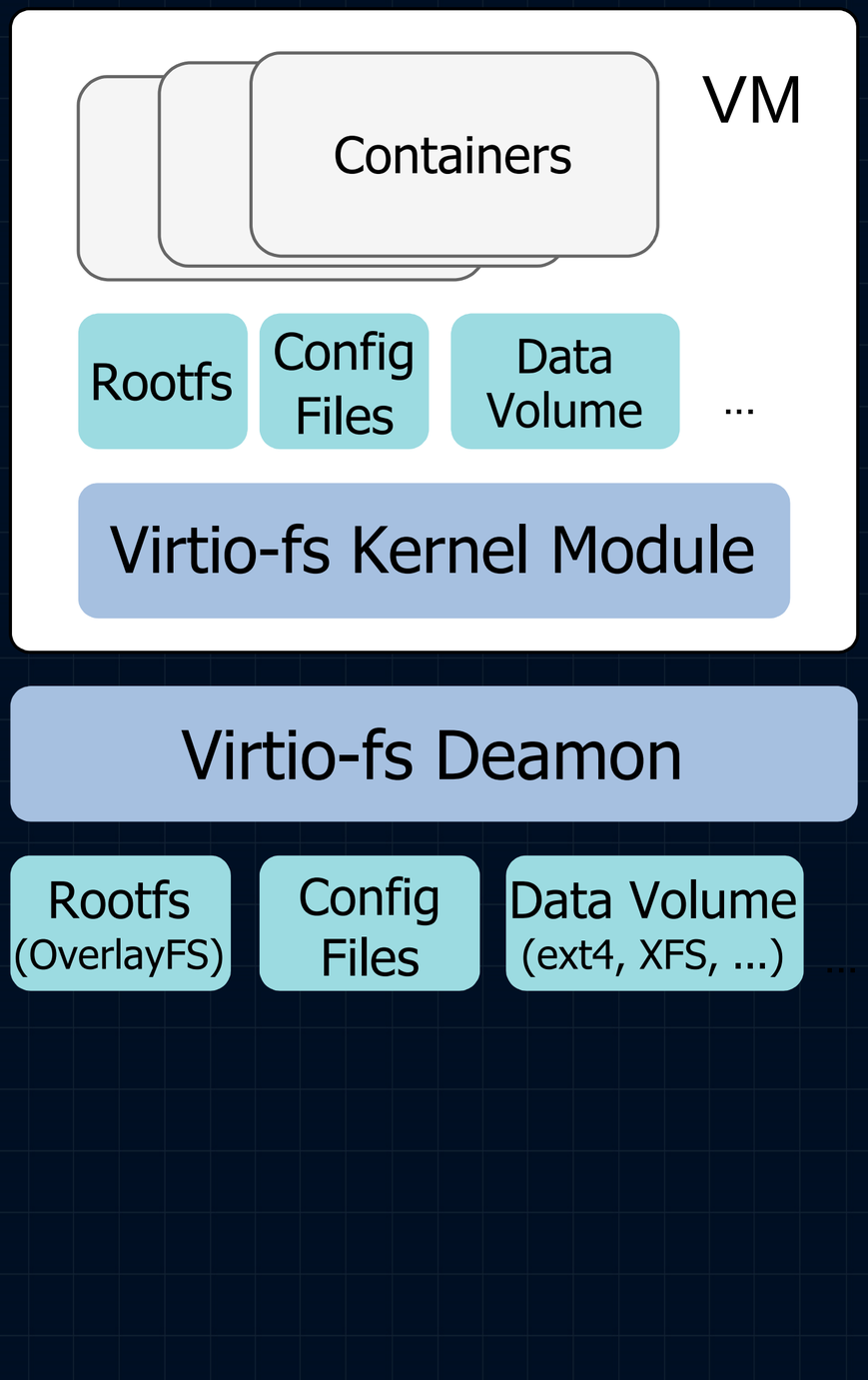


- 5.15合入linux kernel
- 场景化的性能/CPU消耗收益明显
- 内部存储服务逐步上量中
- OSV基础库/DPDK/SPDK等方案逐步支持中

存储的探索与实践

virtiofs的增强

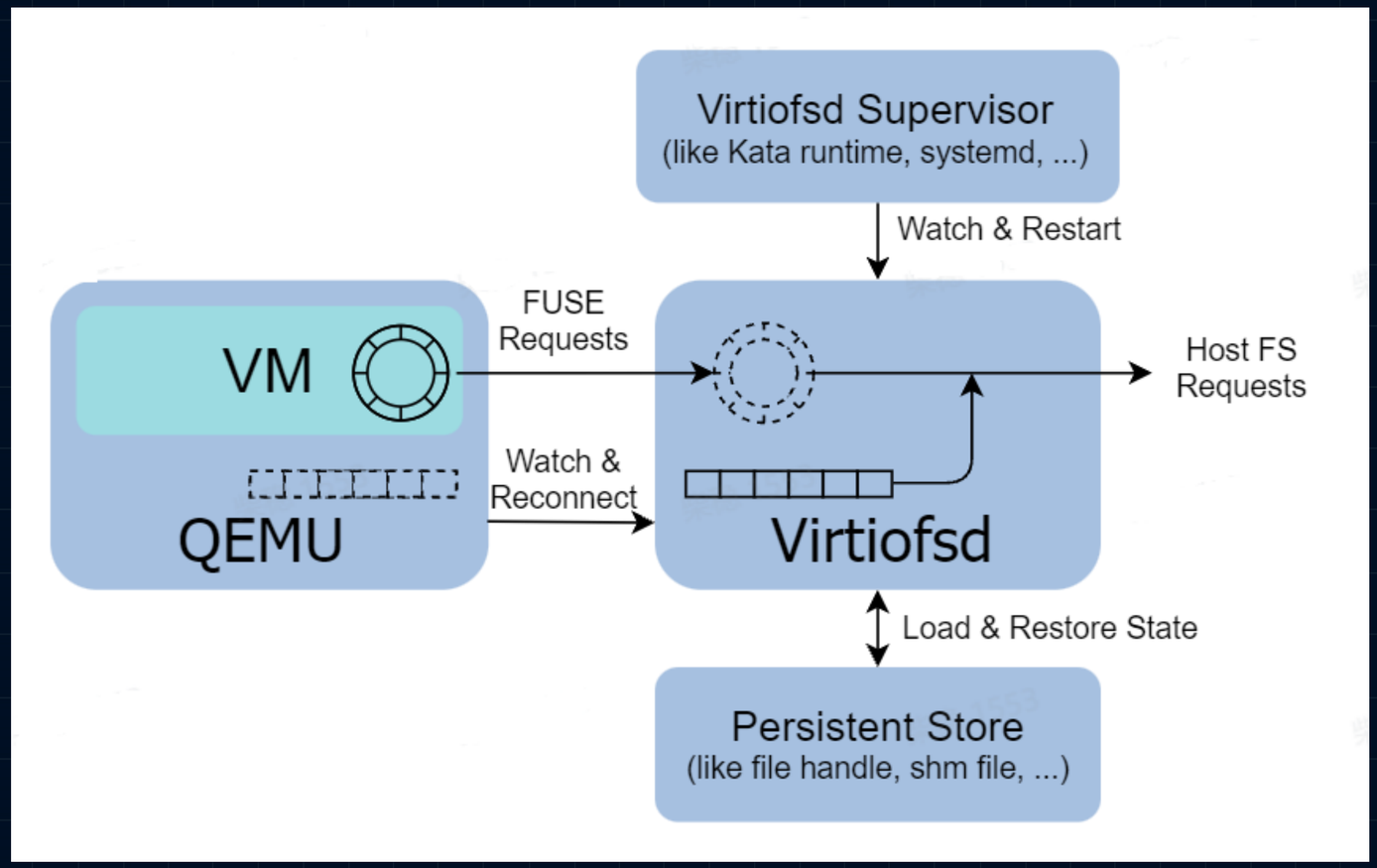
- Crash&Recovery重连
- 热升级与热迁移



存储的探索与实践

viritofs的增强

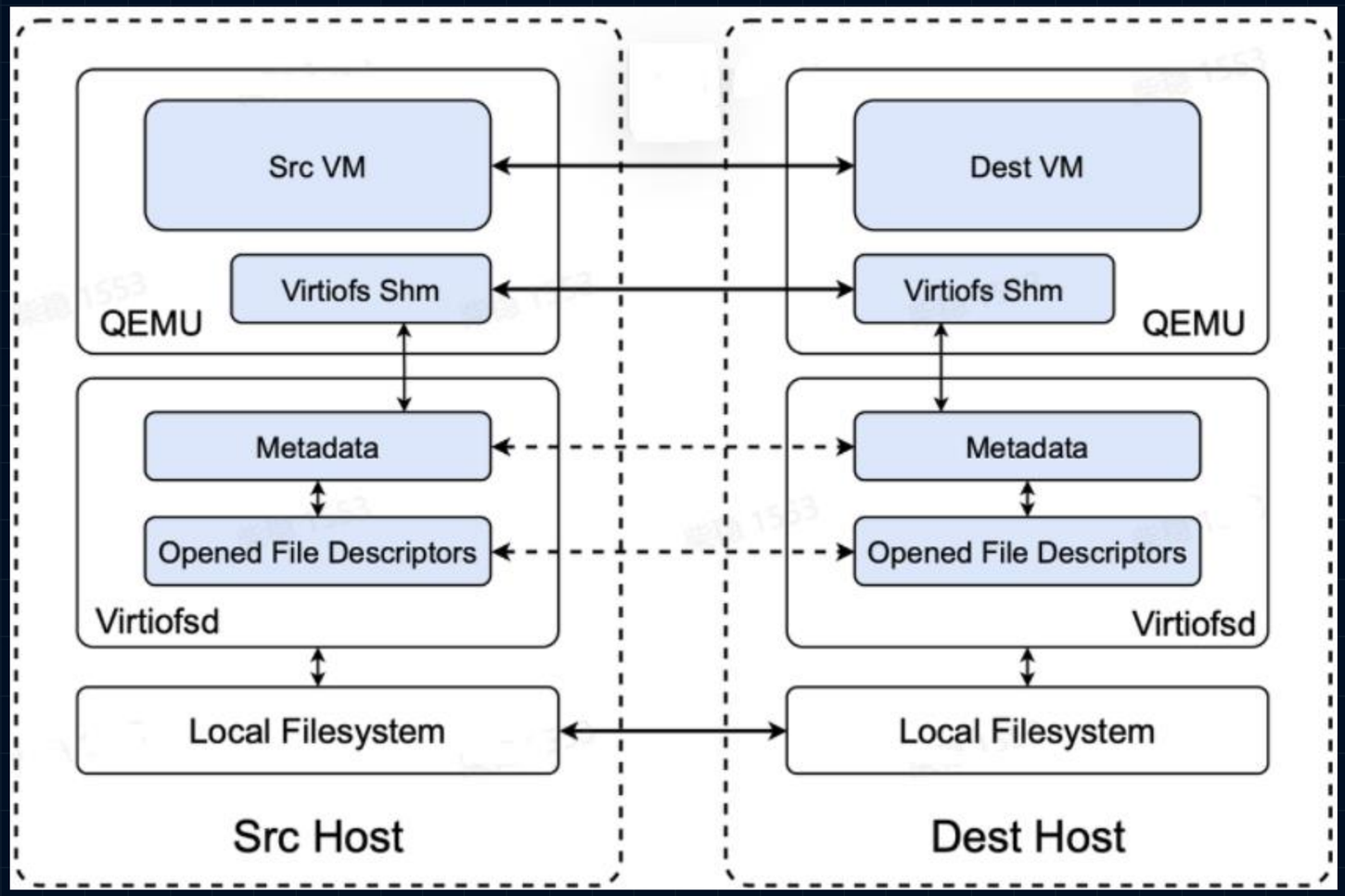
- FUSE请求的repaly
- vhost-user inflight 请求的tracking
- viritofsd 自身状态的save/restore
- 幂等的保证与遗留问题



存储的探索与实践

viritofs的增强

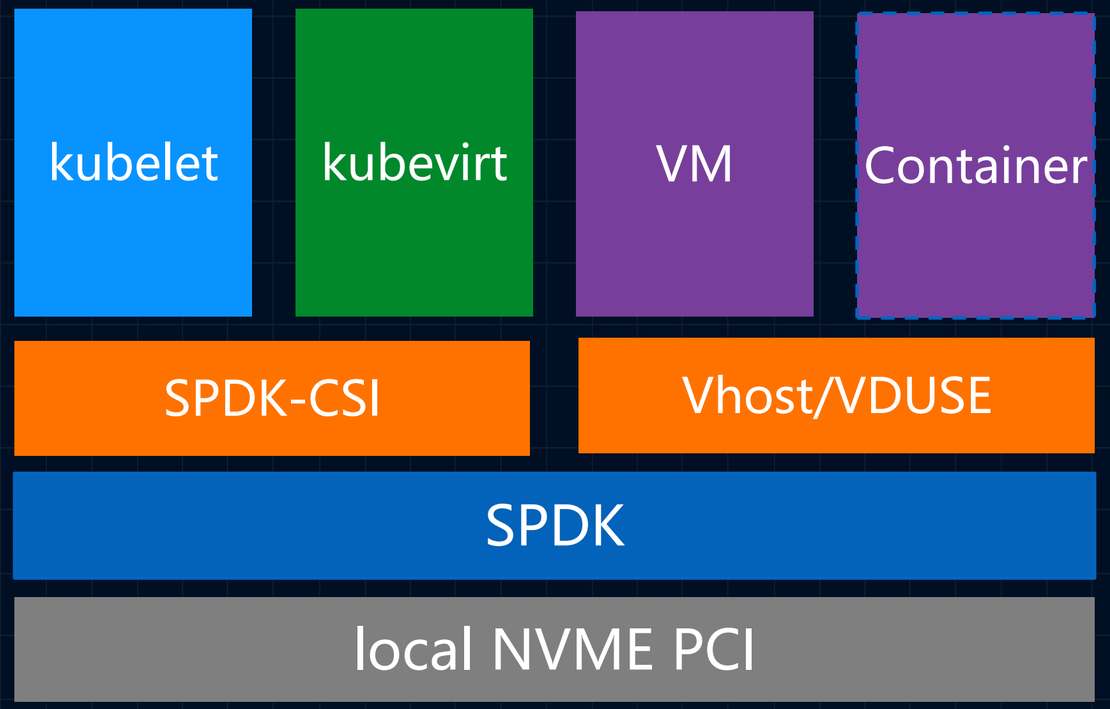
- virtiofs只读挂载
- 元数据热迁移
- vhost-user-fs设备状态的迁移
- 有状态服务virtiofs daemon的状态迁移



存储的探索与实践

kubevirt/SPDK本地盘

- 支持本地PCIE SPDK target
- systemd/daemonset的服务部署方式
- 支持虚机/容器的统一接入方式



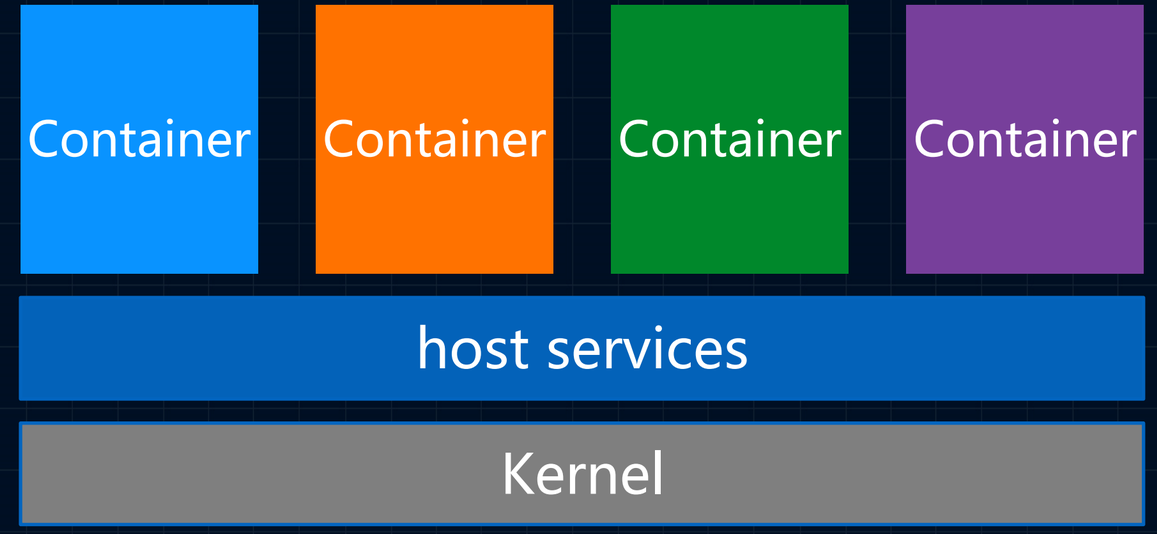
目录

计算的探索与实践

Kata-BM介绍

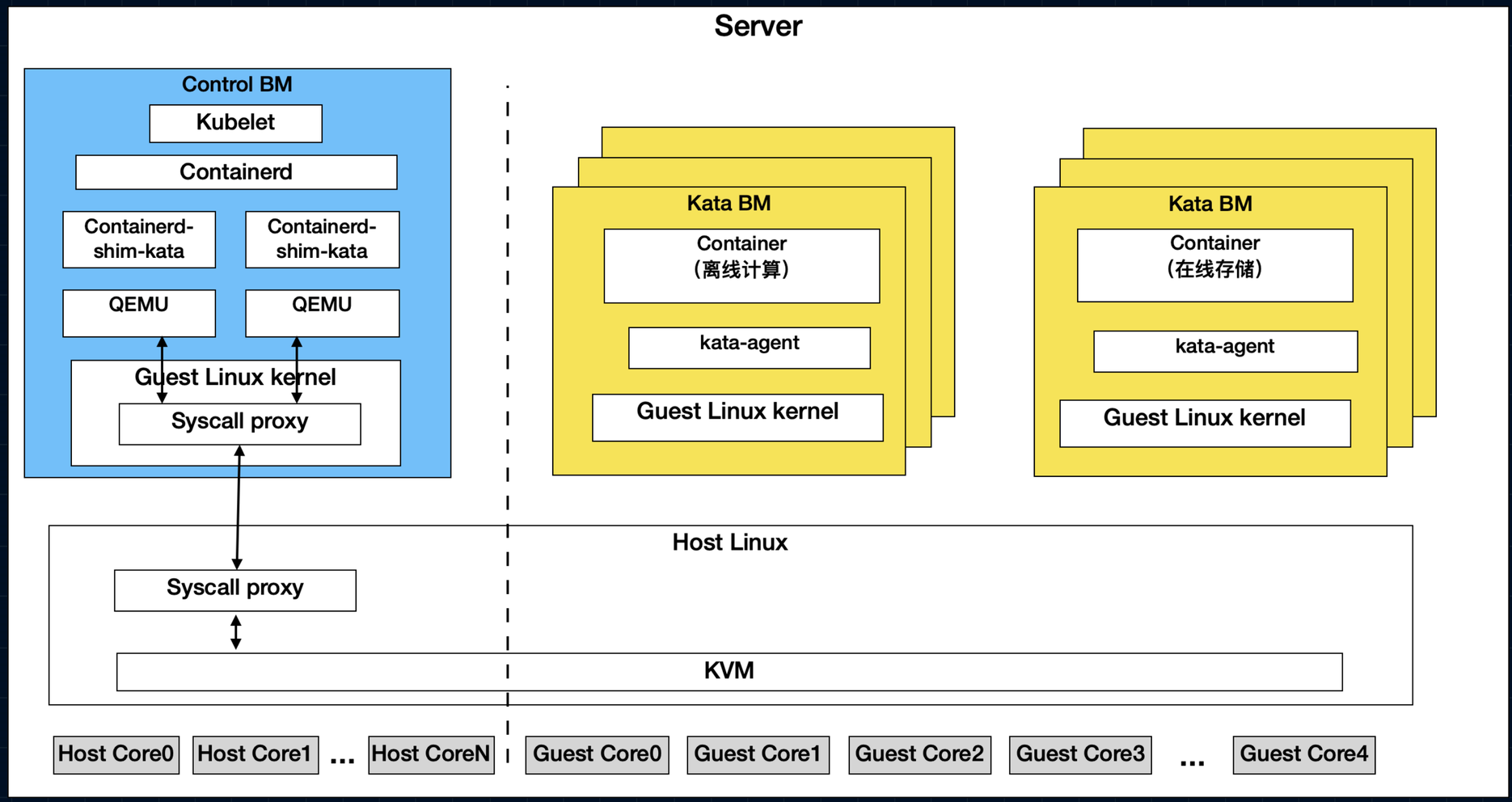
- 非同质化业务容器混布及服务质量保证问题
- 单server越来越高的cpu密度带来的扩展性问题

- 调度/内核事件噪音
- 单一宏内核的扩展性
- server粒度cpu的规模化增长趋势
- 传统虚拟化的损耗



计算的探索与实践

Kata-BM介绍

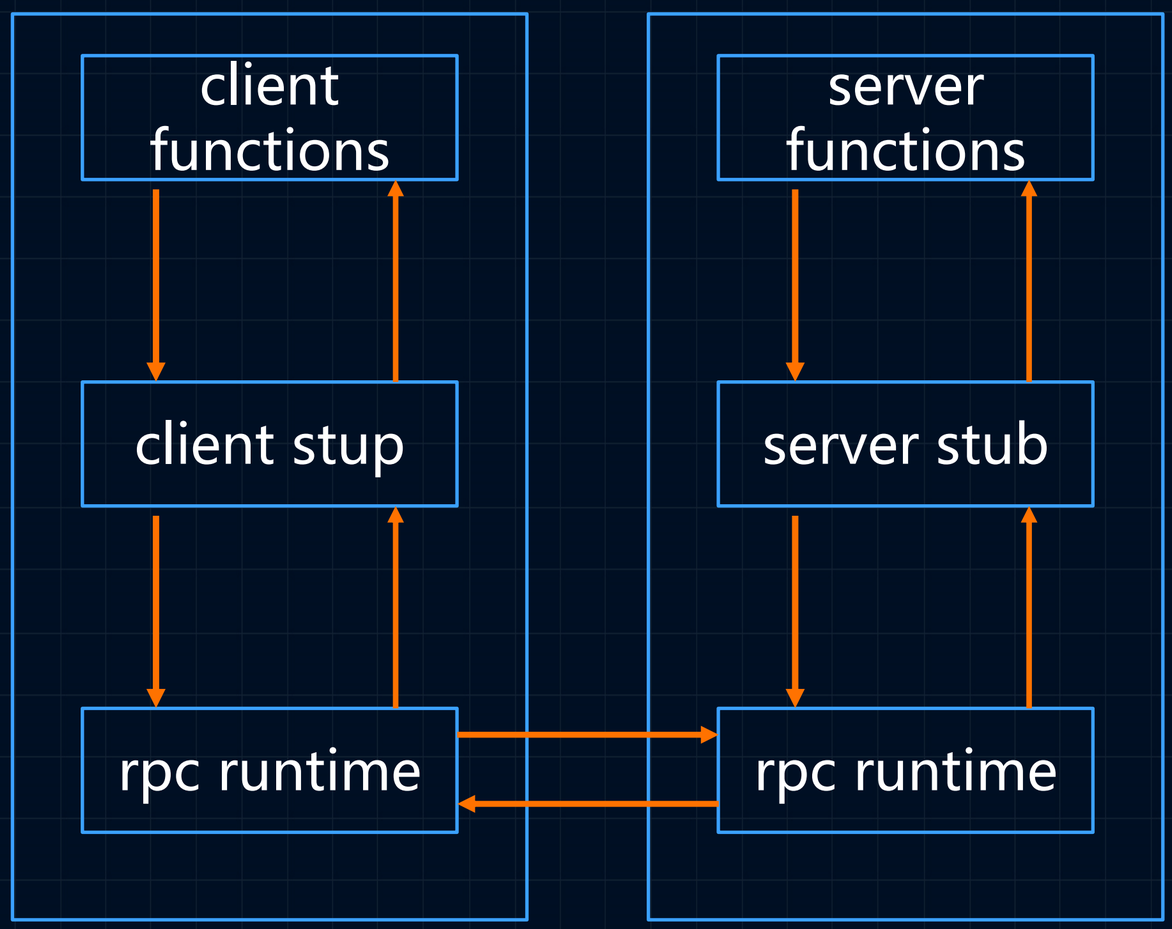


- kata-runtime适配支持
- 使用预分配的物理cpu及内存页面
- 延迟可预期收益明显
- 去VMEXIT/去EPT消除虚拟化overhead
- 管控/业务分离，系统调用代理业务BM的control plane

计算的探索与实践

IPC/RPC的性能优化

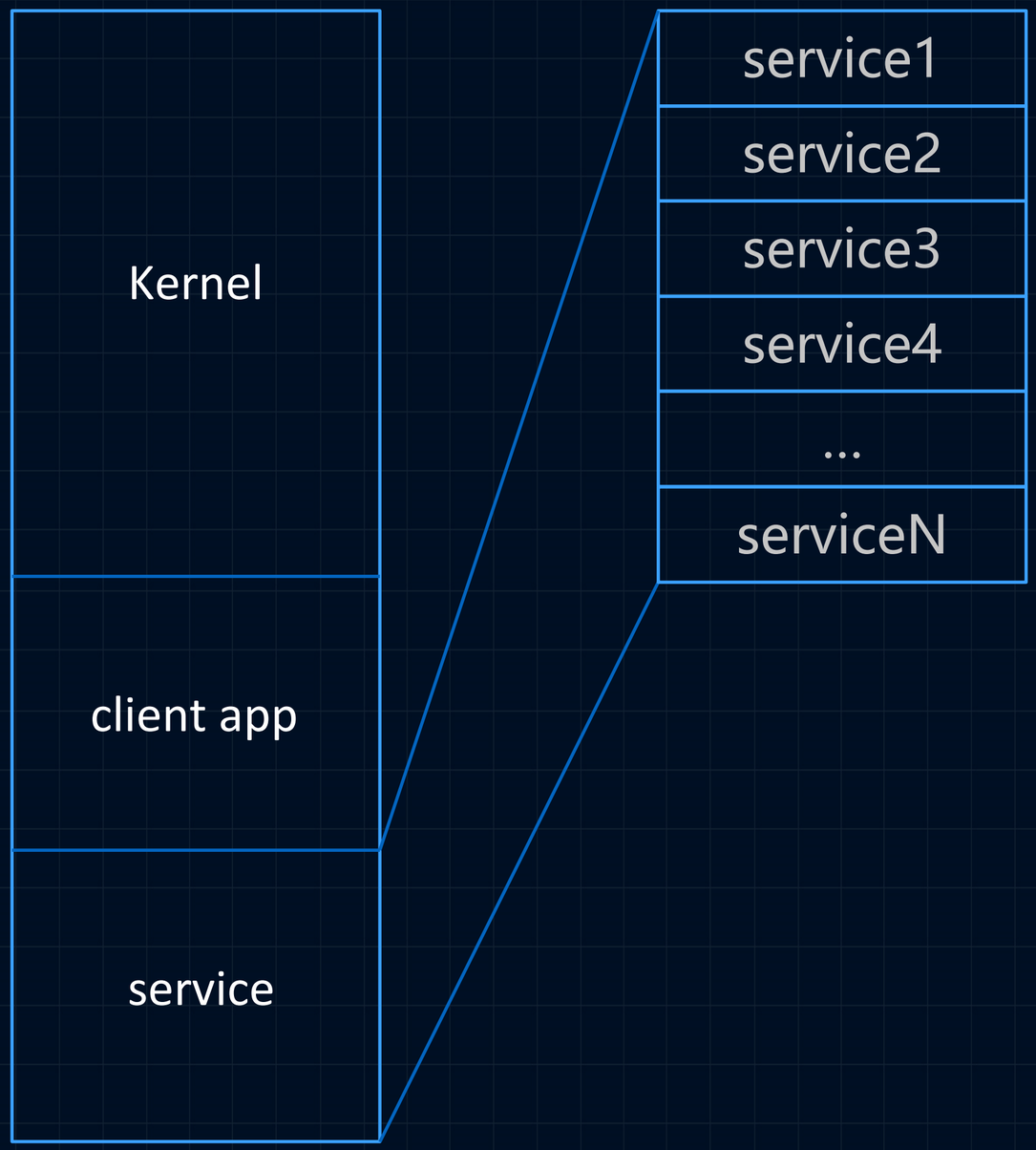
- 微服务场景的核心transaction
- 数次syscall
- 参数复制及序列化相关过程
- 数次的地址空间切换



计算的探索与实践

IPC/RPC的性能优化

- 虚地址空间预分配
- client与service协商划分userspace虚地址空间
- 快路经转化为本地函数调用
- 陷入内核的调整处理
- language/binding的线程调度模型差异处理



目录

未来的一些想法

一些思考

- 应用微架构(如微服务/微前端等)的趋势会越来越明显
  - 宏内核的调度/通信/资源隔离的支持力度
  - 内核对于workload sensitive机制
  - 类微内核的优势
- 开放处理器架构设计的机会(如risc-v等)
  - 针对云原生技术特点的定制优化
  - 容器化/虚拟化底层硬件特性的增强设计
- CPU内存等chip资源分布式逻辑池化的可行性
  - in-chip/out-chip的时间边界临近，cpu内存等传统节点资源的逻辑化
  - 虚拟化的形态演进

ThankS

字节跳动 STE 团队技术交流

该二维码 7 天内 (10 月 27 日前) 有效，重新进入将更新