高可用技术的实践分享

中国系统架构师大会 · 北京

2016-10-27

马耀泉

提纲

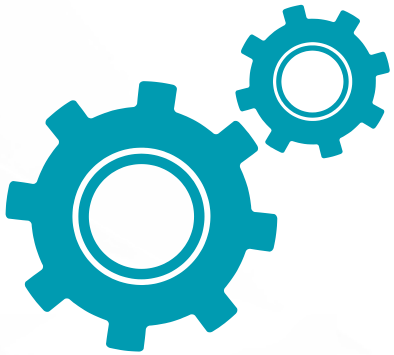
1. 云平台的高可用需求
2. 基础组件的高可用实践
3. 平台监控系统
4. 升级系统
5. 未来工作展望

1 云平台的高可用需求

1 云平台的高可用需求

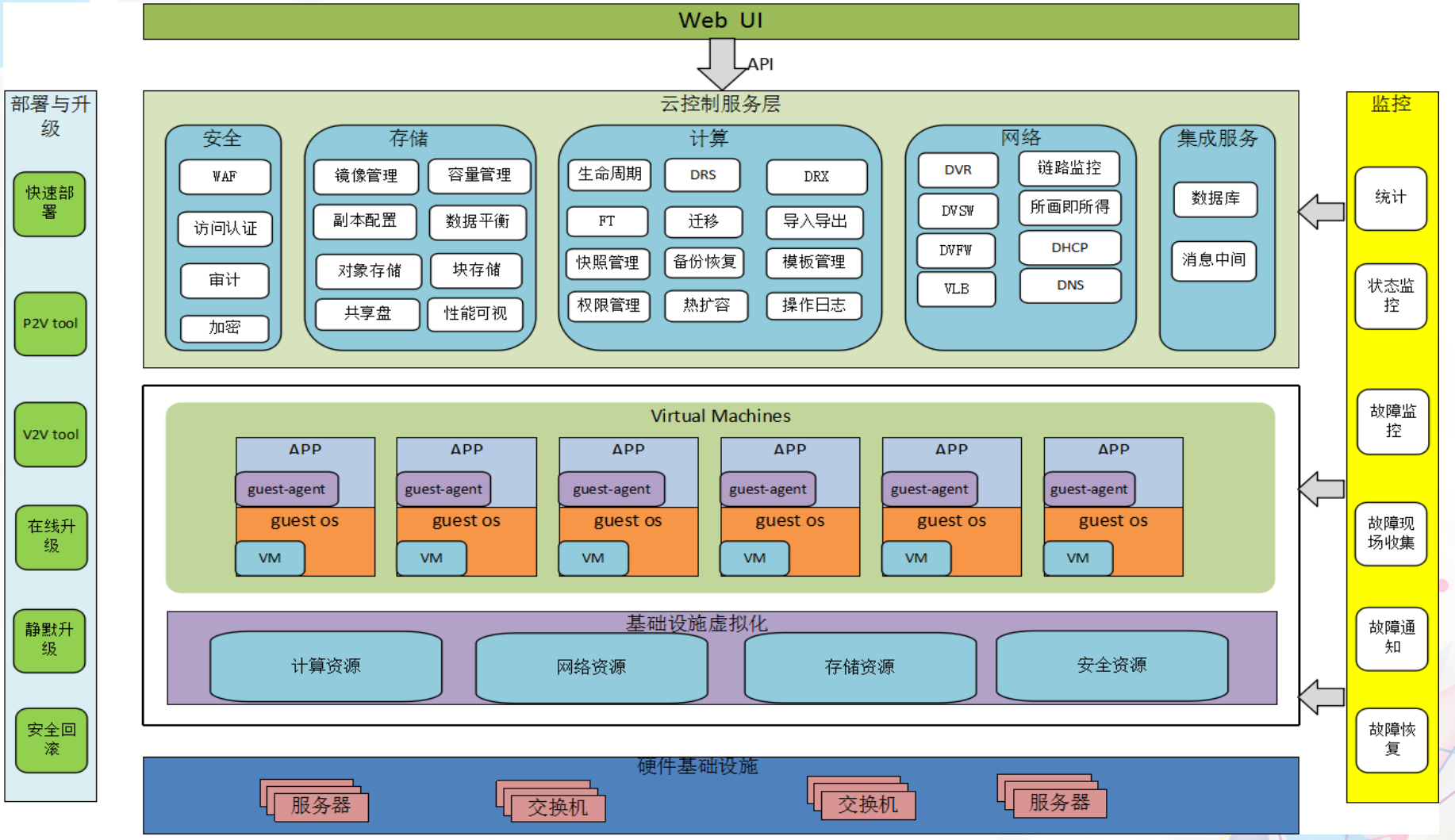
- 如何保证基础平台的稳定性
- 如何监控业务和快速恢复
- 在升级过程中如何保证业务的连续性

2 基础组件的高可用设计

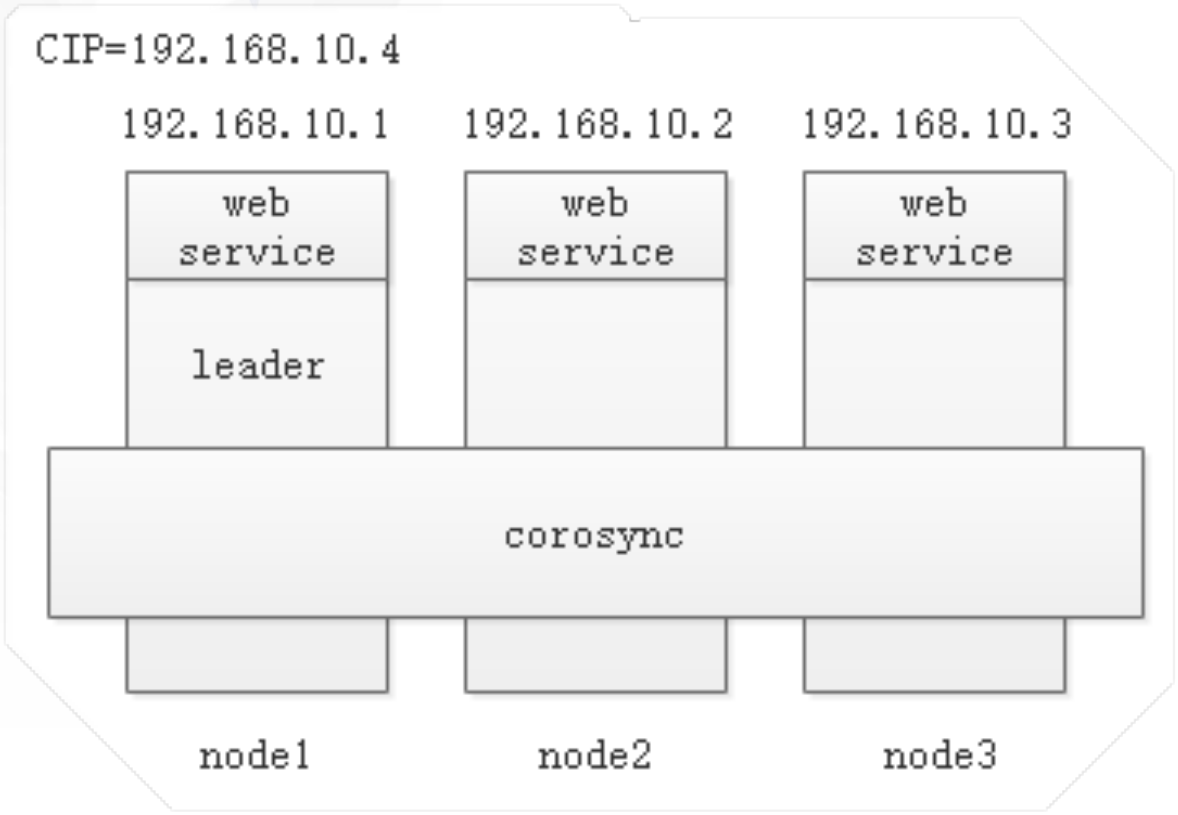


2 基础组件的高可用设计

整体架构



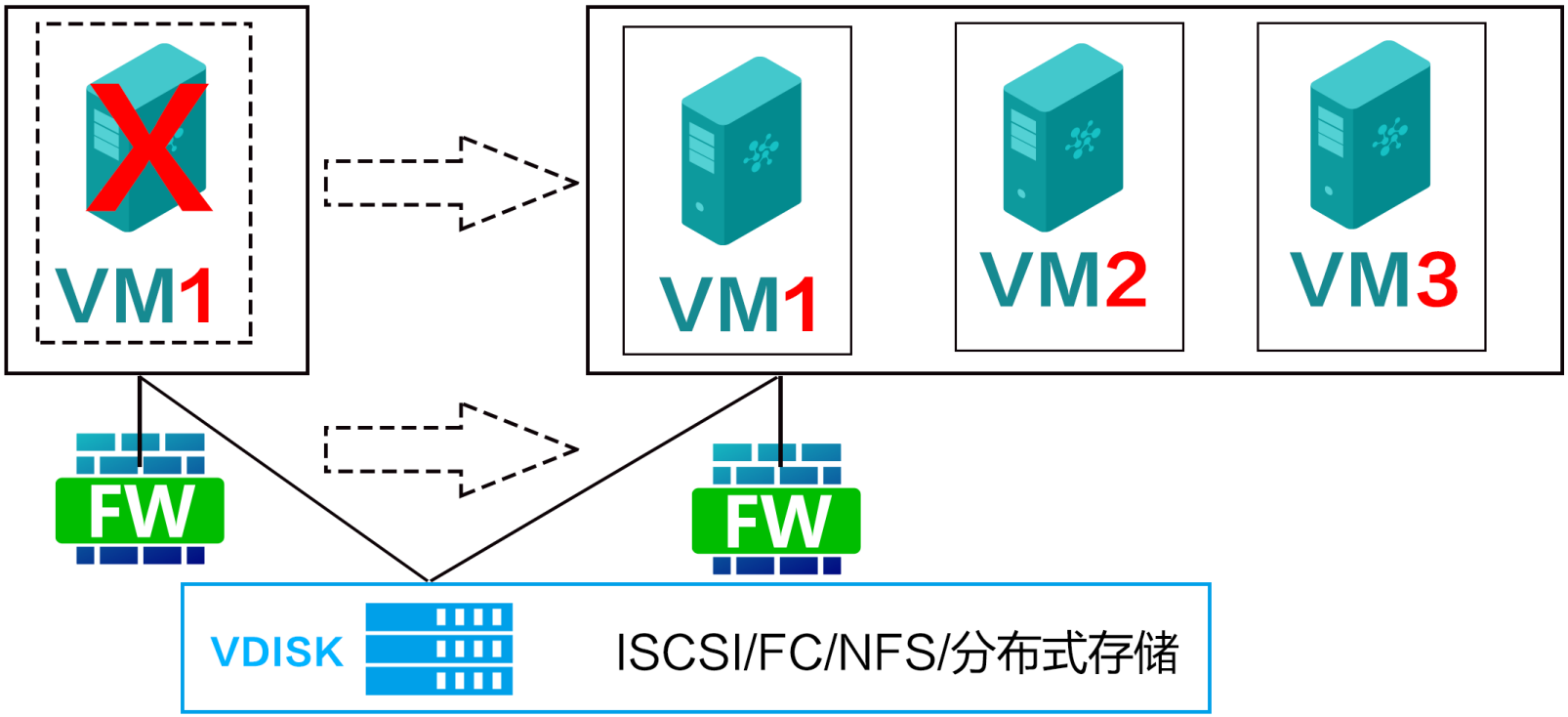
控制服务层之集群高可用



- **去中心化设计**
- 集群基础配置使用集群文件系统存放；
- 使用corosync维护成员关系；
- 集群leader故障，自动推选；
- 配置集群IP，跟随leader；

② 基础组件的高可用设计

虚拟机高可用（1）



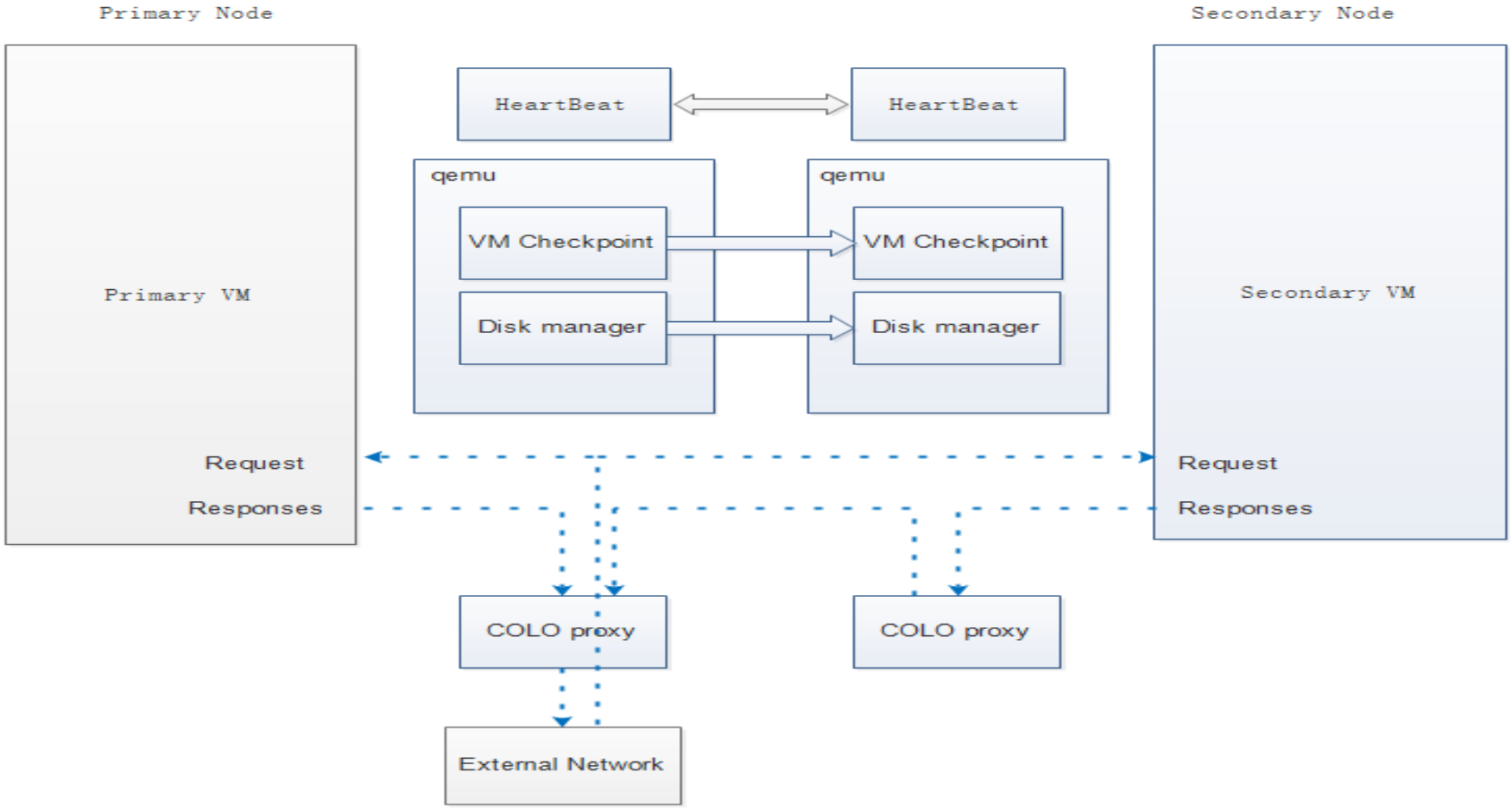
Failover（故障切换）：

- 业务恢复时间（RTO）：系统 + APP + 探测时间（可配置）；
- 需要共享存储（外置存储或者 分布式存储）；
- **网络可达，网络的配置和防火墙策略跟随；**

虚拟机高可用（2）

FT(Fault Tolerance)技术：

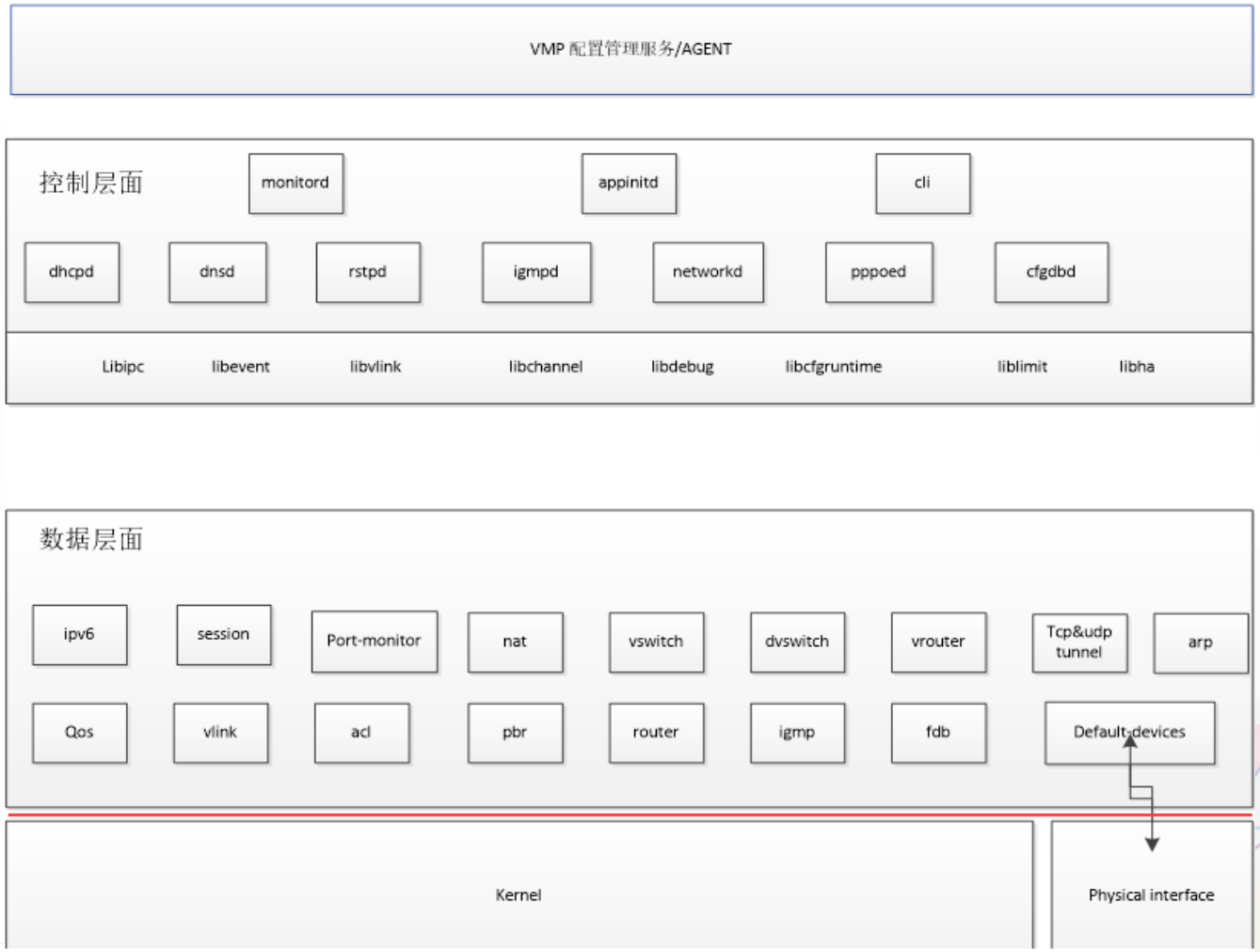
- 基于coarse-grained lock-steping
- 需要万兆网络进行状态同步
- 需禁用虚拟化高级特性（如热迁移）



虚拟网络高可用（1）

应用层协议栈/转发面

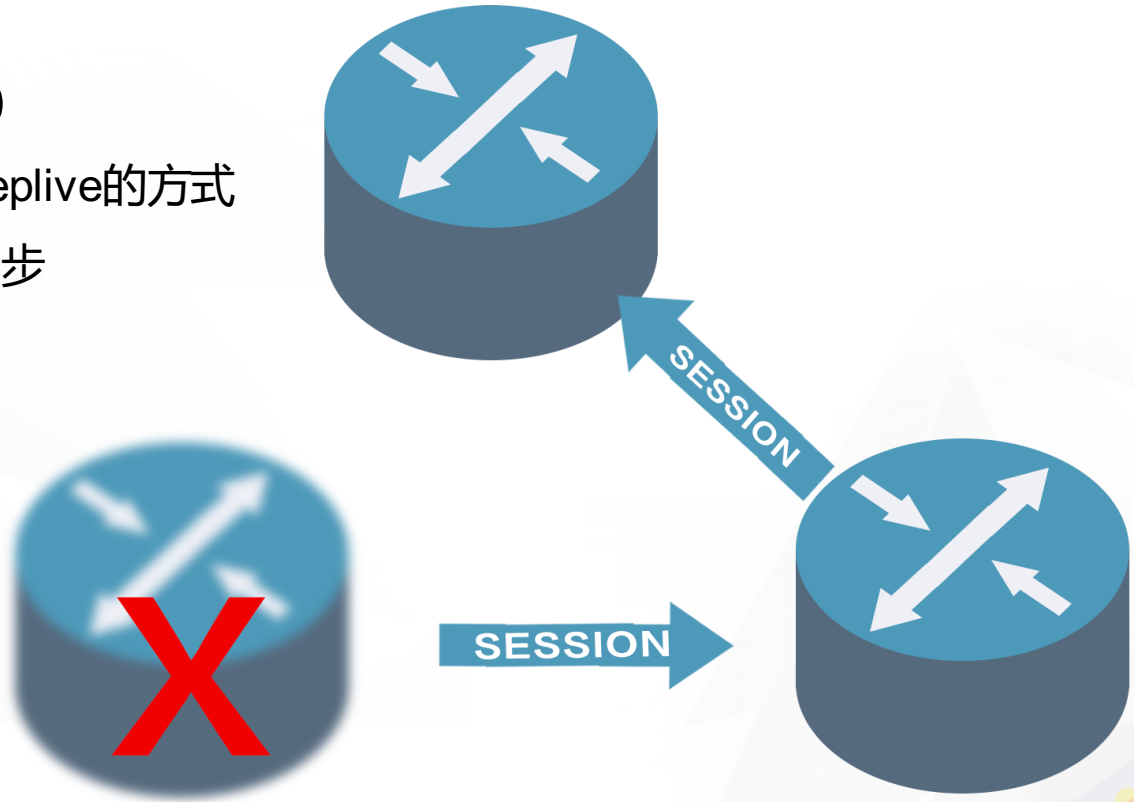
- 网络故障不会导致主机宕机
- 利用DPDK实现高性能报文处理
- 应用层支持主备切换确保业务连续性



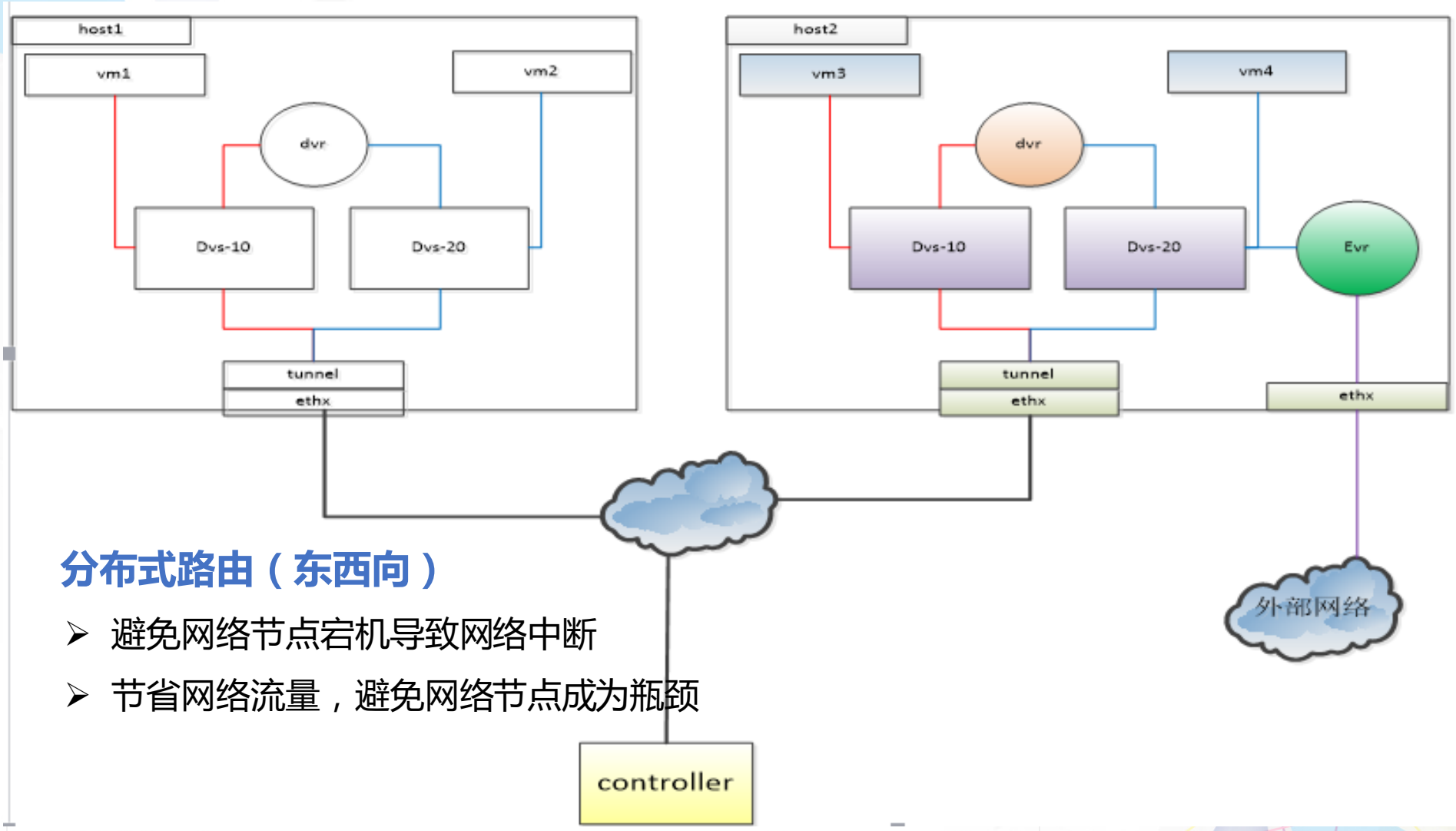
虚拟网络高可用（2）

边界集中路由（南北向）

- 虚拟路由器之间使用keeplive的方式
- 虚拟路由器SESSION同步
- 备选节点重选择



虚拟网络高可用（3）



分布式路由（东西向）

- 避免网络节点宕机导致网络中断
- 节省网络流量，避免网络节点成为瓶颈

虚拟存储高可用（1）

支持数据的多副本

➢防止物理故障导致数据丢失

支持快速修复

➢只修复副本间差异数据

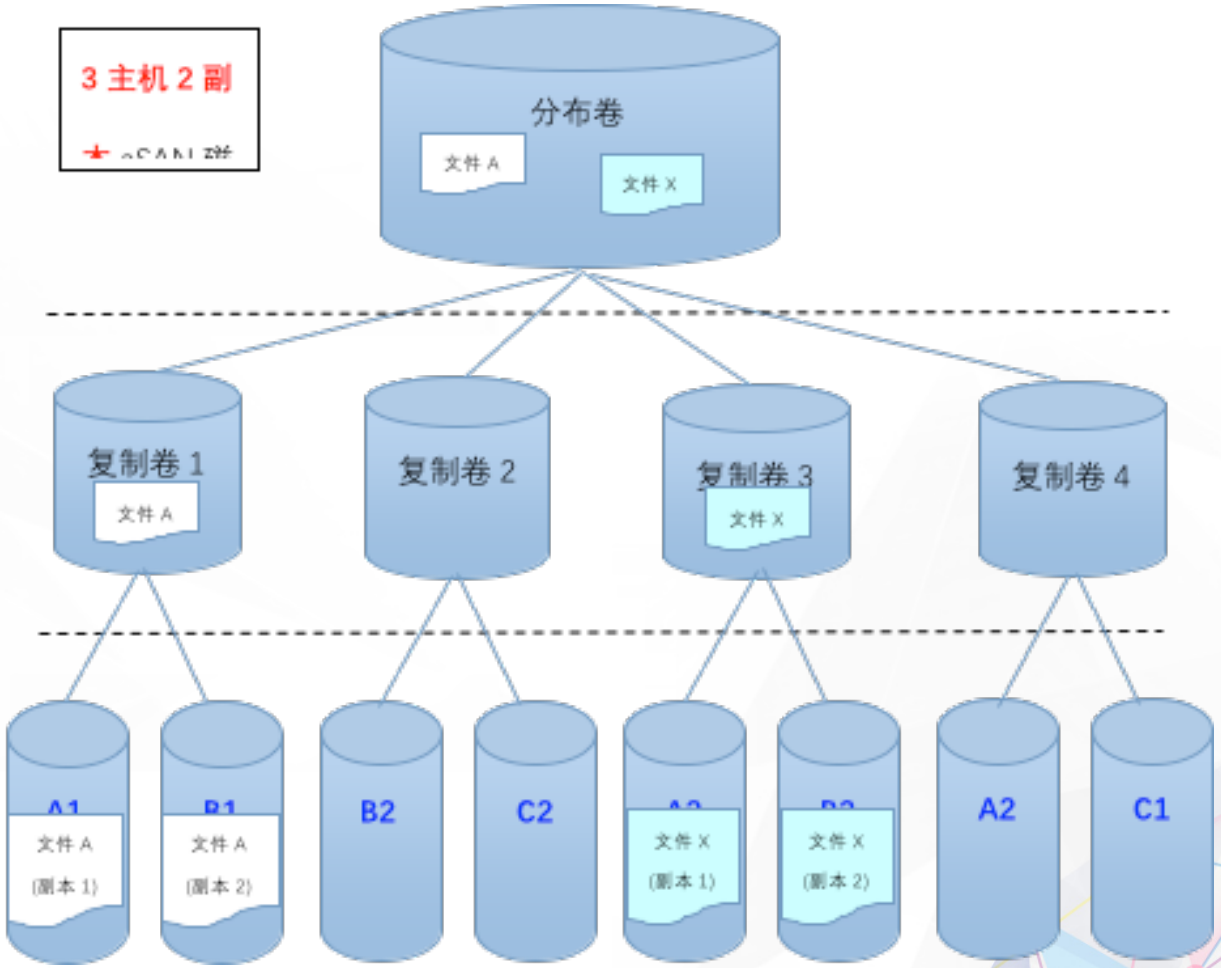
磁盘检测

➢SSD寿命预测

➢坏道告警

无元数据中心

➢避免存在单点故障



虚拟存储高可用（2）

防止脑裂：仲裁机制

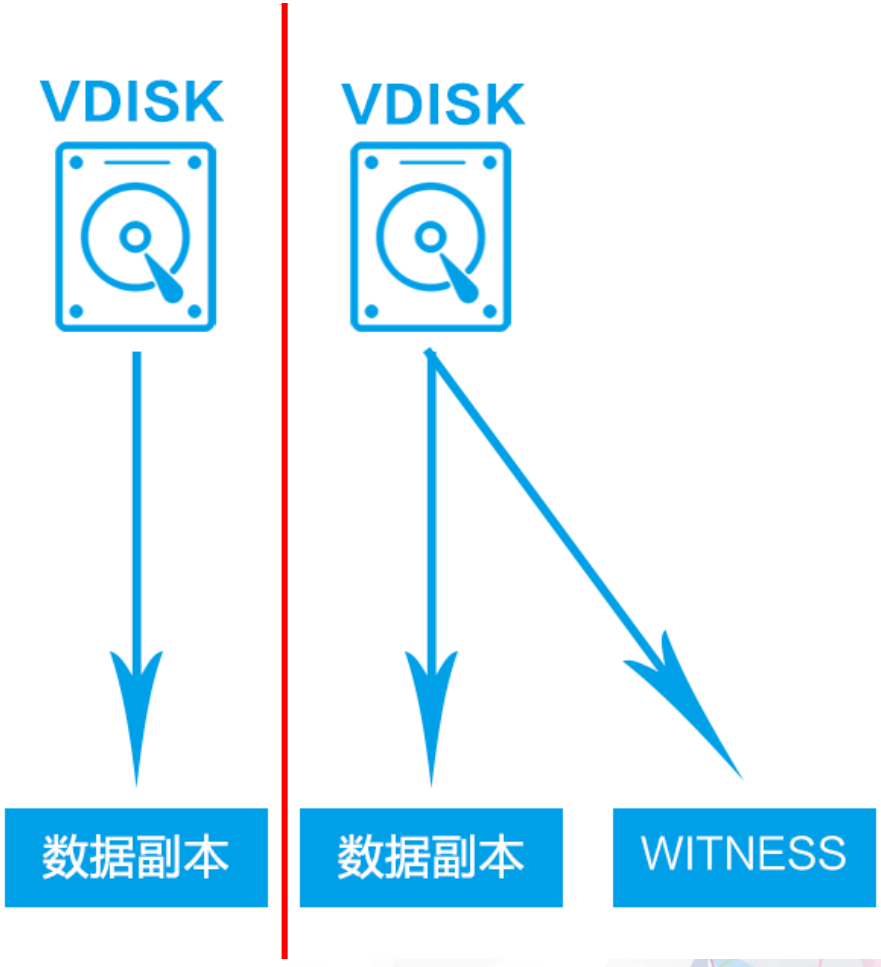
- 增加元数据副本，作为仲裁者
- 只有在超过1/2副本数在线，才允许读写

端到端数据校验

- 解决静默错误的数据损坏
- 解决软件异常导致的数据损坏

数据自愈

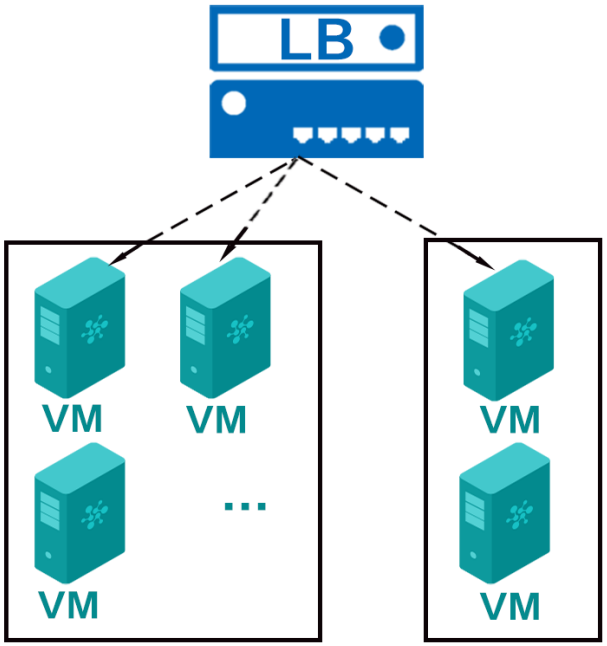
- 第一时间修复可能的数据异常

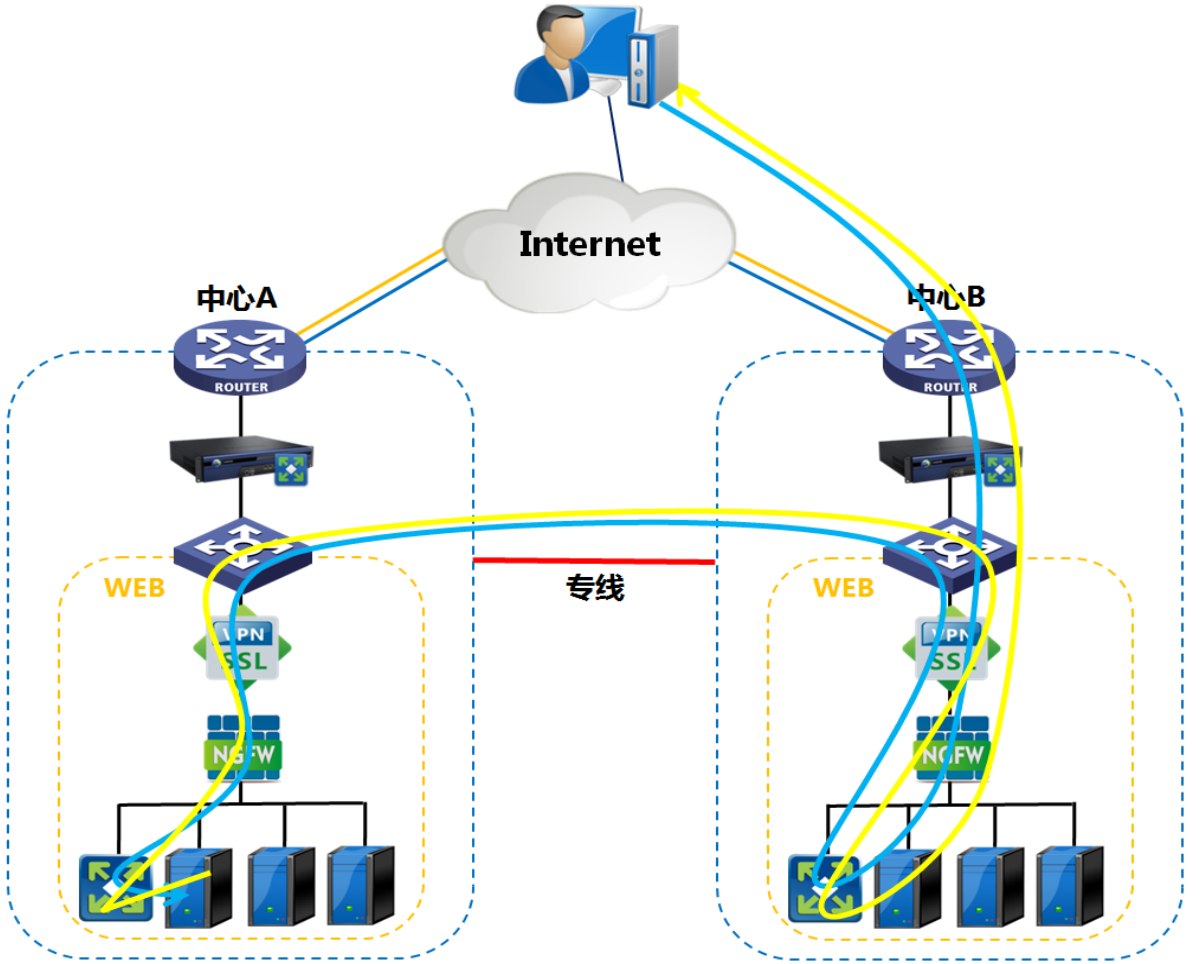


承载业务的高可用

基于负载均衡的双活技术

- 技术成熟度高
- 4-7层的负载均衡
- 可以支持两中心双活

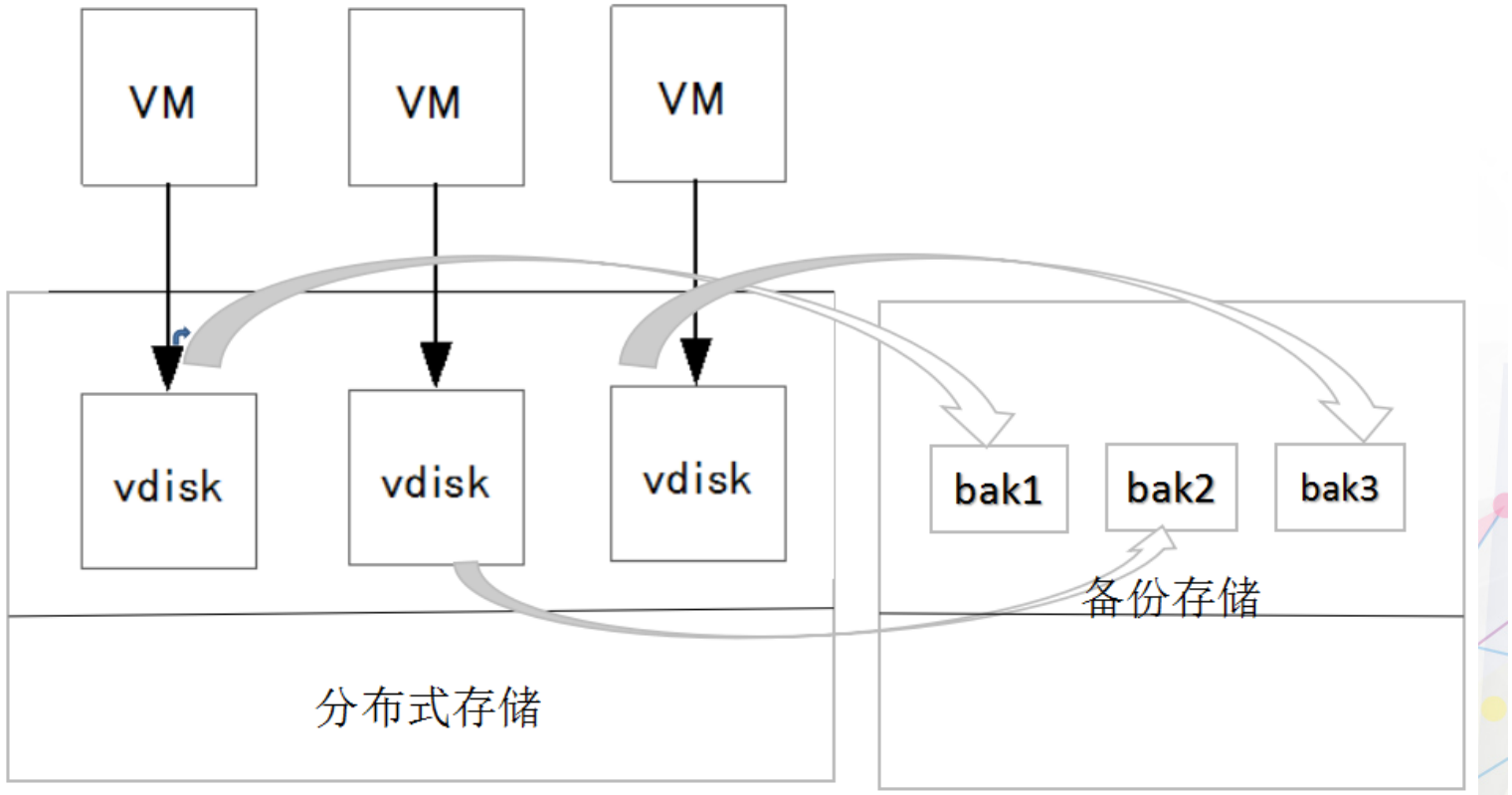




虚拟机备份（1）

基于虚拟机的备份

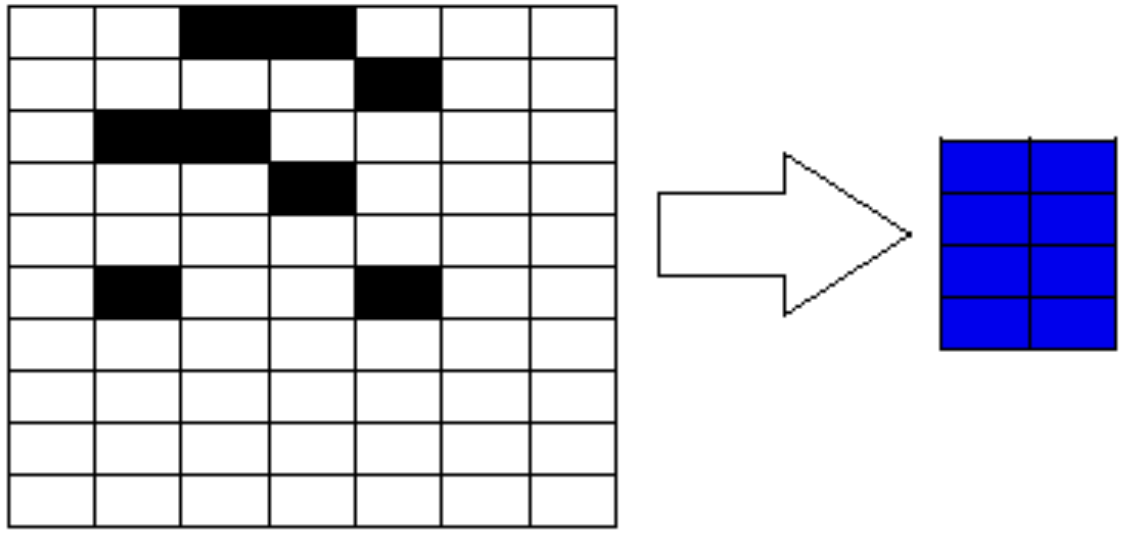
- 每次备份仅仅只是增量数据；
- 支持缓存文件过滤；
- 支持定时备份；



虚拟机备份（2）

虚拟机备份的两种方式

- 通过快照方式，记录两次快照的差异，实现文件增量备份
- 通过位图方式，记录数据变化，实现增量备份
- 快照方式会造成性能的持续降低；位图方式只在备份过程中短暂性能损失



3 监控系统

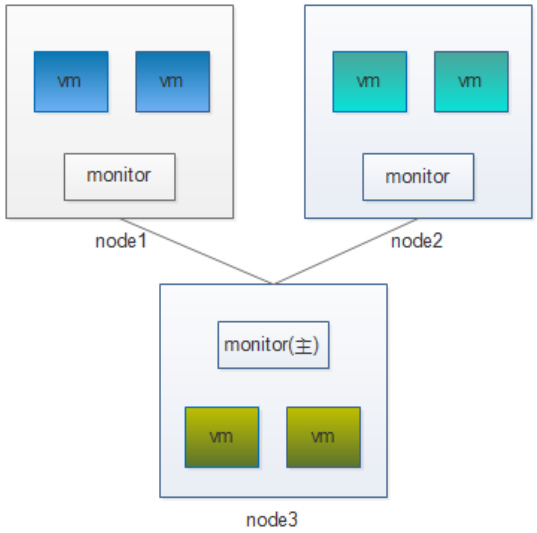
3 监控系统

- 物理主机监控
- 虚拟机监控
- 虚拟网络监控
- 虚拟存储监控

物理主机监控

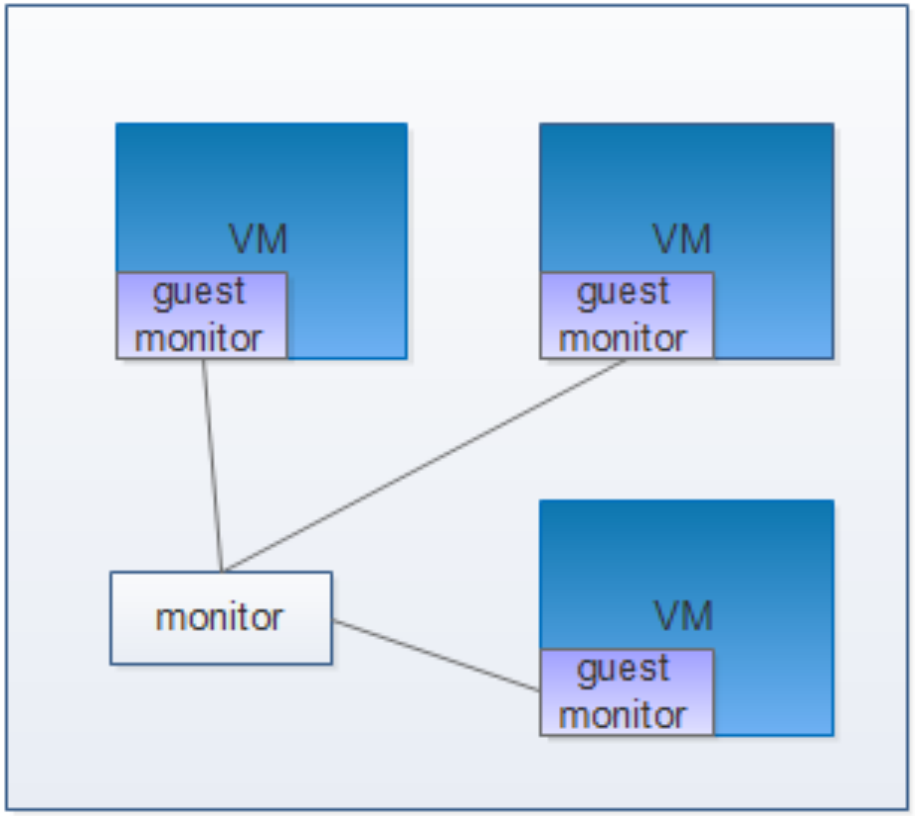
主机级别的监控

- 主机的keepalive，保证主机宕机及时发现；
- 对主机的CPU，内存，网络，磁盘监控；
- 对外置存储和分布式存储状态监控。



node name	CPU	memory	manager network	sotrage network	business network	FC stoarge	server san
node1	98%	13%	×	√	√	√	√
node2	80%	84%	√	√	×	√	√
node3	50%	64%	√	√	√	√	√

虚拟机监控



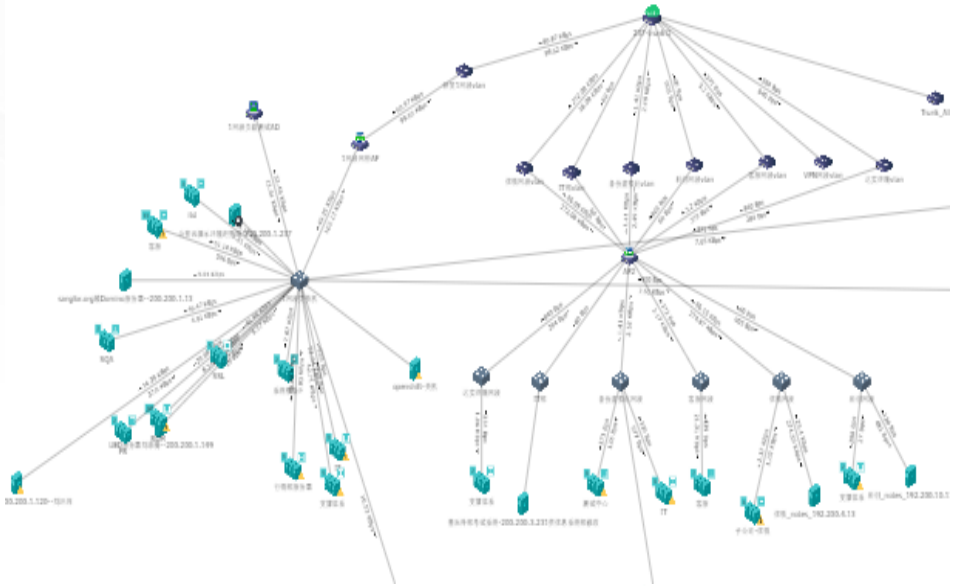
虚拟机级别的监控

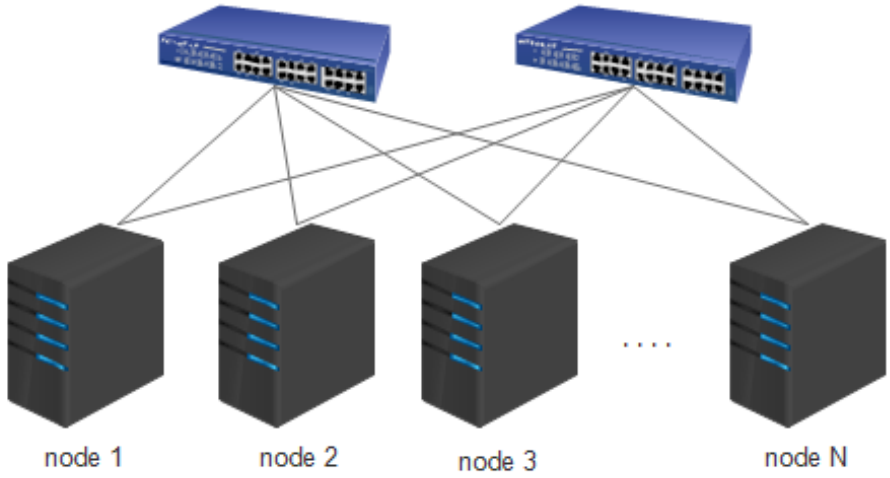
- 虚拟机内部CPU，内存和磁盘占用；
- 实时IO，网络流量；
- hypervisor层异常。

虚拟网络监控（1）

网络监控的问题：

- 虚拟网络拓扑和物理拓扑相差大；
- 网络流量在虚拟平台内部流转，流量不可见；
- 网络排障无法使用传统手段；

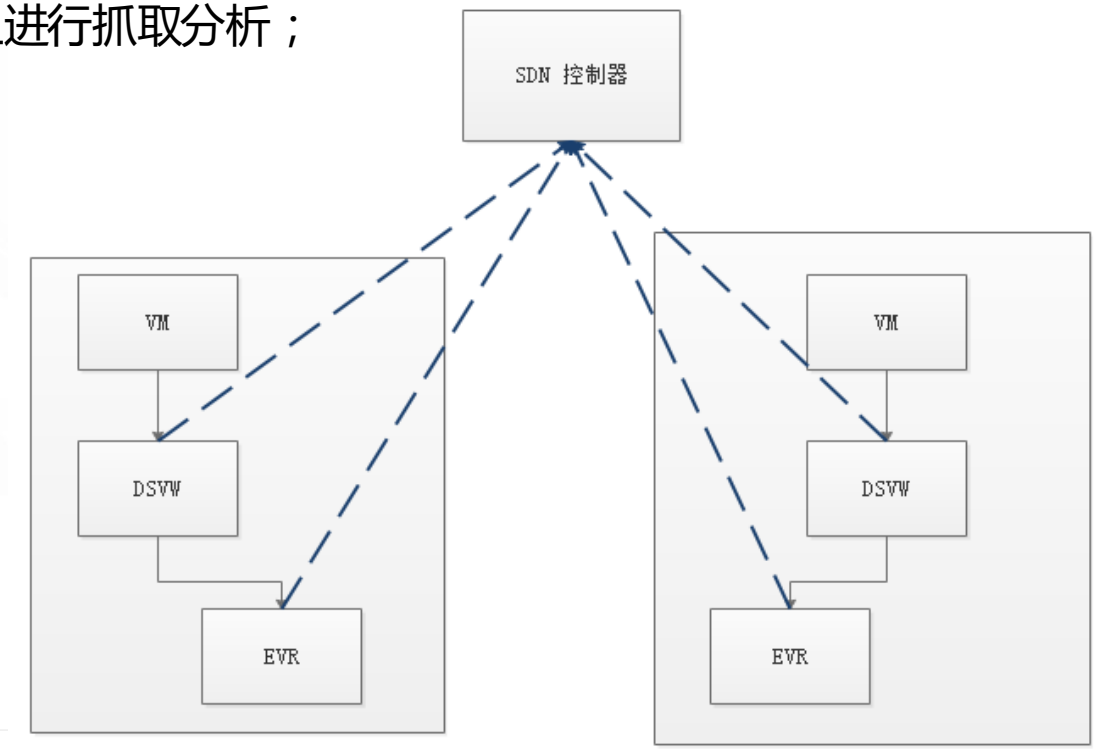




虚拟网络监控（2）

网络监控功能：

- 可以在任意网络节点或者虚拟机发送数据包模拟业务；
- 数据包经过的所有网络设备和物理设备都返回显示；
- 数据包由于ACL或者路由不可达等错误而丢弃，明确上报；
- 可以在任意网络节点上进行抓取分析；
- 数据包途径路径展现；
- 异常点精确定位；
- 配置错误反馈；



虚拟存储监控

- 存储吞吐能力，展现集群整体的吞吐；
- 磁盘健康状态，检测是否磁盘离线；
- SSD寿命预测，计划替换SSD；
- 缓存命中率，方便排查性能问题；
- 存储网络的链路检测与切换；
- 慢盘检测，及时发现加入硬件性能问题；

4 升级系统的改进

升级系统的设计

模块化设计，使得每个模块可以独立升级

升级过程允许新旧两个模块同时工作、平滑替换

虚拟机可以在不同版本的hypervisor之间热升级

升级改进 - 热升级

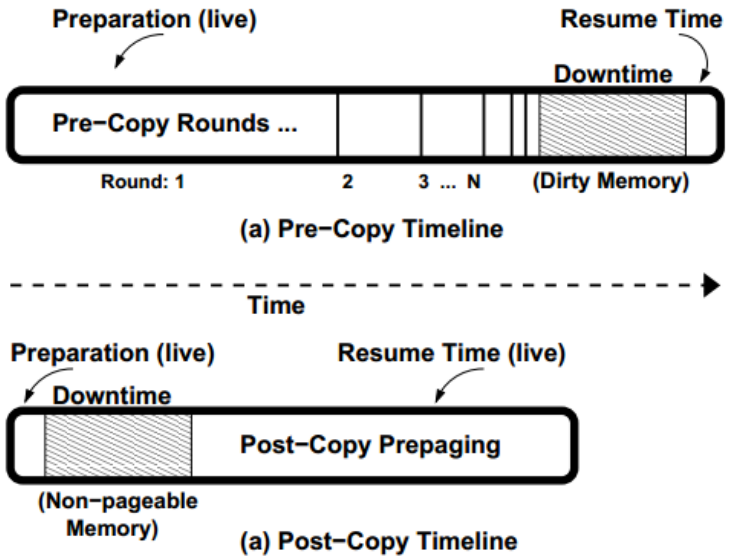
当前热升级最常用的方式为跨主机的热迁移；

优点：

业务基本无感知；

缺点：

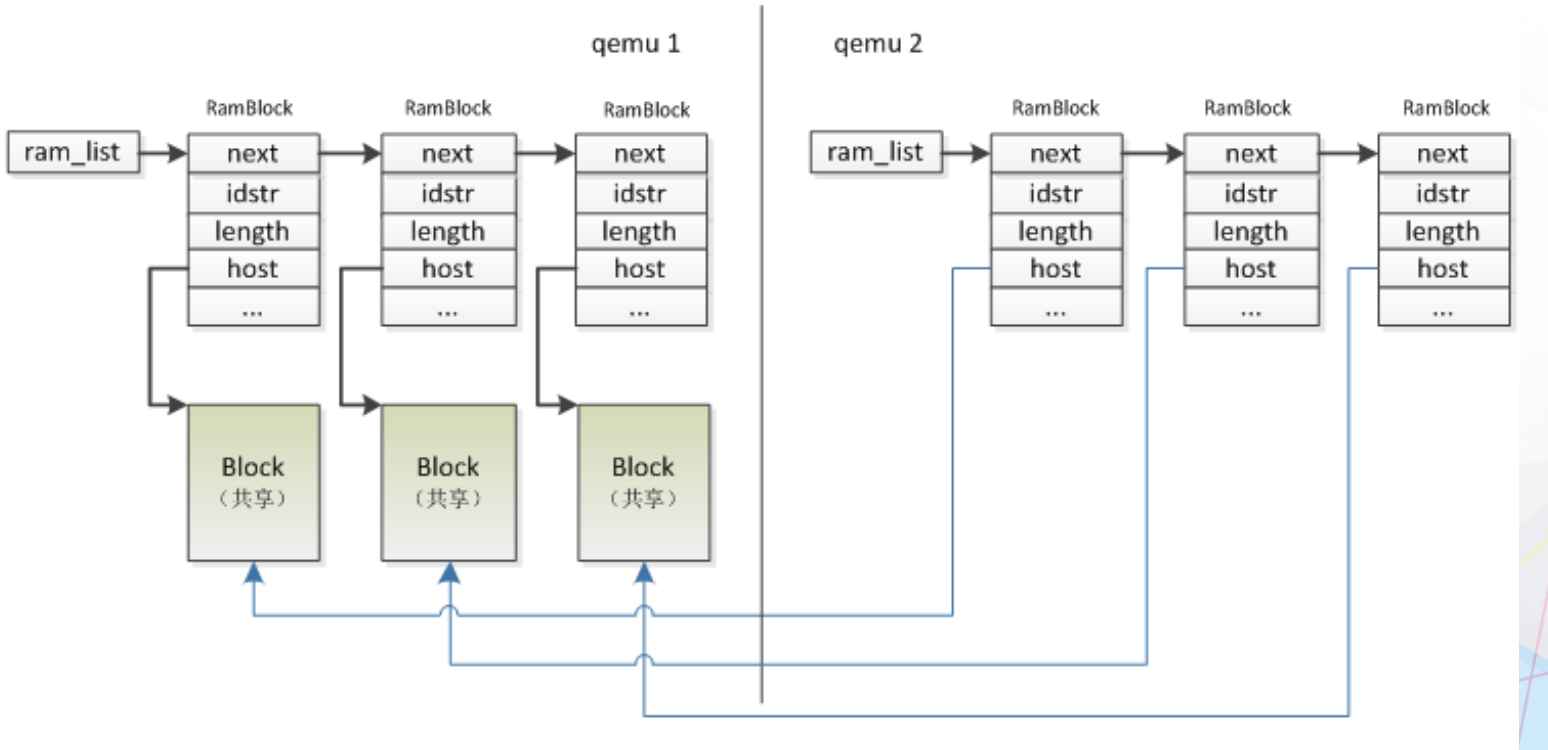
1. 跟虚拟机更新内存速度有关，可能会导致迁移时间过长，甚至迁移失败；
2. 迁移过程依赖网络的稳定性和性能，如果网络不稳定可能会导致迁移失败；



升级改进 - 热升级

热升级改进：

- 本地迁移，内存共享无需拷贝；
- 迁移时间短，跟虚拟机业务没有关系；
- 提高并发迁移速度，缩减升级时间；



5 未来工作展望

5 未来工作展望

1. 虚拟机业务恢复时间（RTO/RPO）持续优化，争取异地恢复做到秒钟级恢复；
2. 提供混合云方案提高性价比的备份方案；
3. 跟第三方管理设备结合，提供更高丰富的业务监控机制；

THANKS

SequeMedia 盛拓传媒 | IT168.com 专注导购16年 | ChinaUnix.net | ITPUB www.itpub.net