JD.COM 京东

容器核心技术及SDN实践

田琪&闫国旗

JD.COM 京东

促进软件开发领域知识与创新的传播

InfoQueue

ArchSummit
全球架构师峰会

[深圳] 2015年7月17日-18日

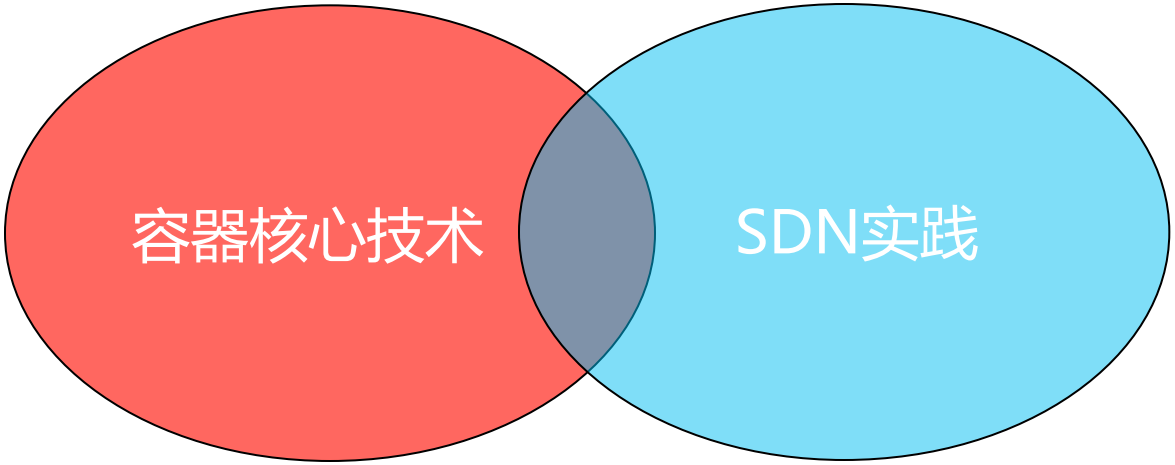
QCon
全球软件开发大会

[上海] 2015年10月15-17日

关注InfoQ官方微信
及时获取QCon演讲视频信息

Agenda

JD.COM 京东



Part 1

- 内核Namespace介绍
- 内核CGroup介绍
- Docker存储驱动选择
- 京东镜像存储系统

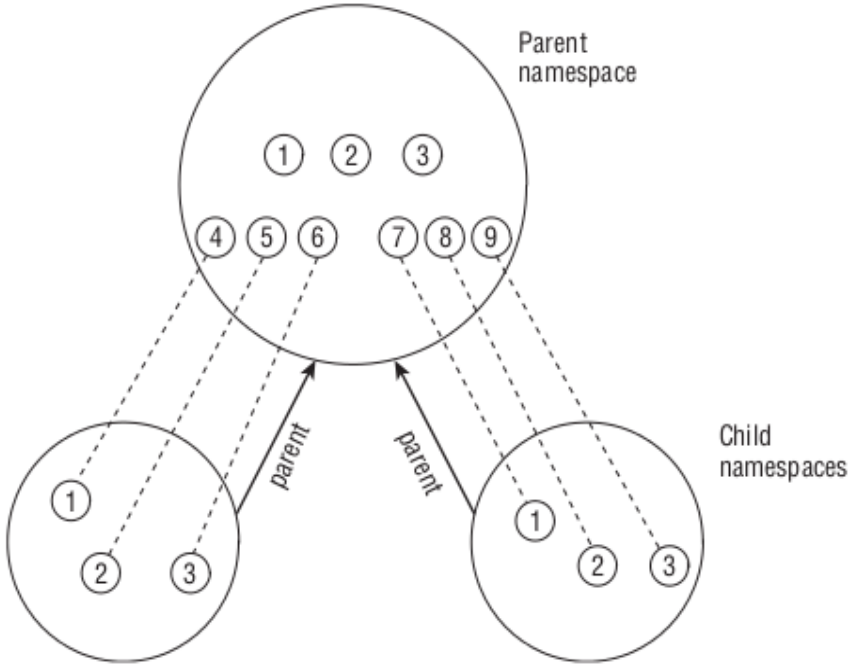
JD.COM 京东

- 提供进程级别的资源隔离
- 为进程提供不同的命名空间视图
- 无hypervisor层，区别于KVM,Xen等虚拟化技术
- 从Kernel 2.4版本引入mnt namespace~3.8引入user namespace仍然持续发展中

Namespace概念及当前支持的种类

JD.COM 京东

- mnt (Mount points)
- pid (Processes)
- net (Network stack)
- ipc (System V IPC)
- uts (Hostname)
- user (UIDS)



- 创建新进程及namespace

```
int clone(int (*fn)(void *), void *child_stack,
          int flags, void *arg, ...
          /* pid_t *ptid, struct user_desc *tls, pid_t *ctid */ );
```

- 加入当前进程到新建namespace中

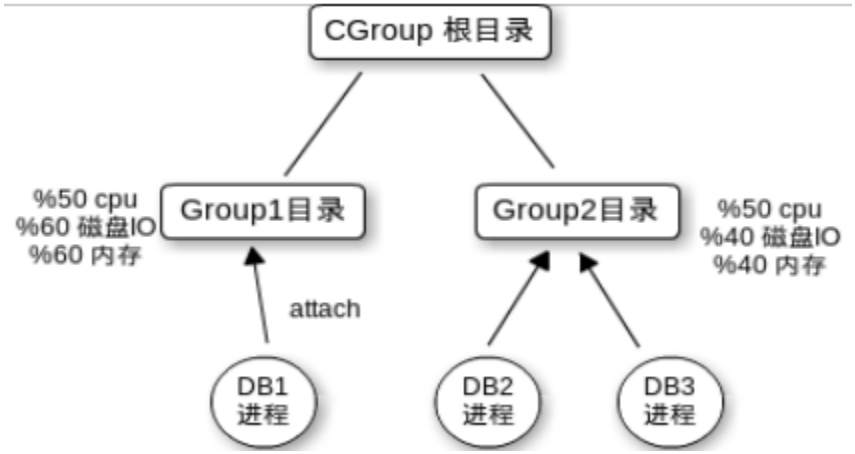
```
int unshare(int flags);
```

- 改变当前进程的namespace

```
int setns(int fd, int nstype);
```

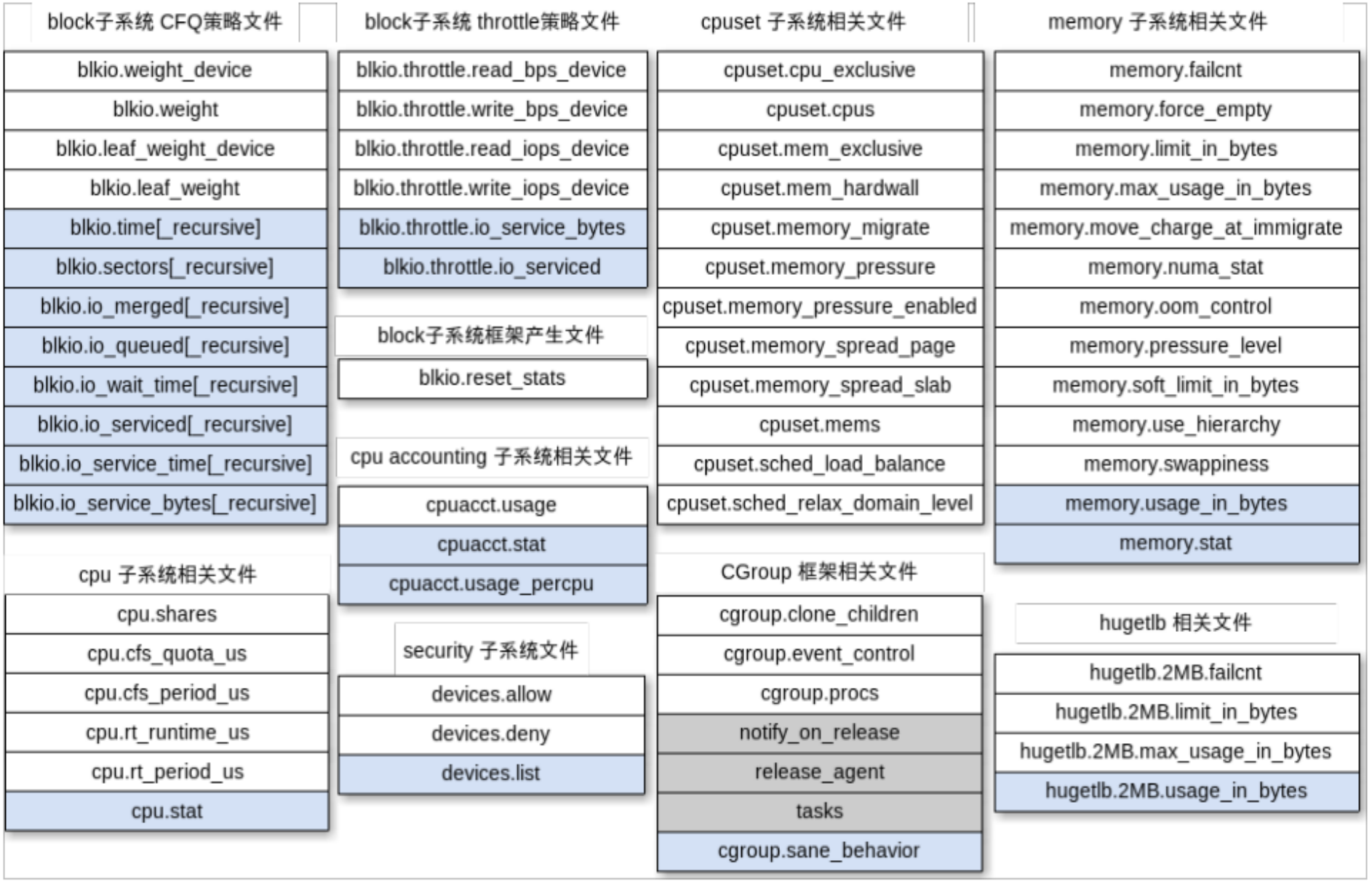
JD.COM 京东

- 提供进程的资源管理功能
- 资源管理主要涉及内存,CPU,IO等
- 不依赖于Namespace，可单独使用
- 管理功能通过VFS接口暴露
- CGroups提供通用框架，各子系统负责实现



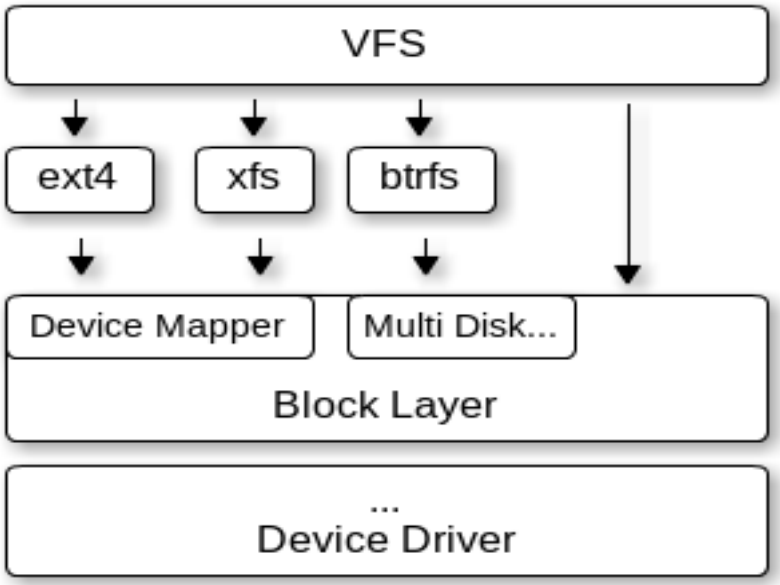
CGroups文件系统

JD.COM 京东



- ns的隔离性不完整
  - 需要更多种类的命名空间
- cgroup IO控制方面问题较多
  - 带宽控制只能CFQ调度器，不适合高速硬件
  - 通用限流策略缺少弹性
  - buffer io无法准确控制

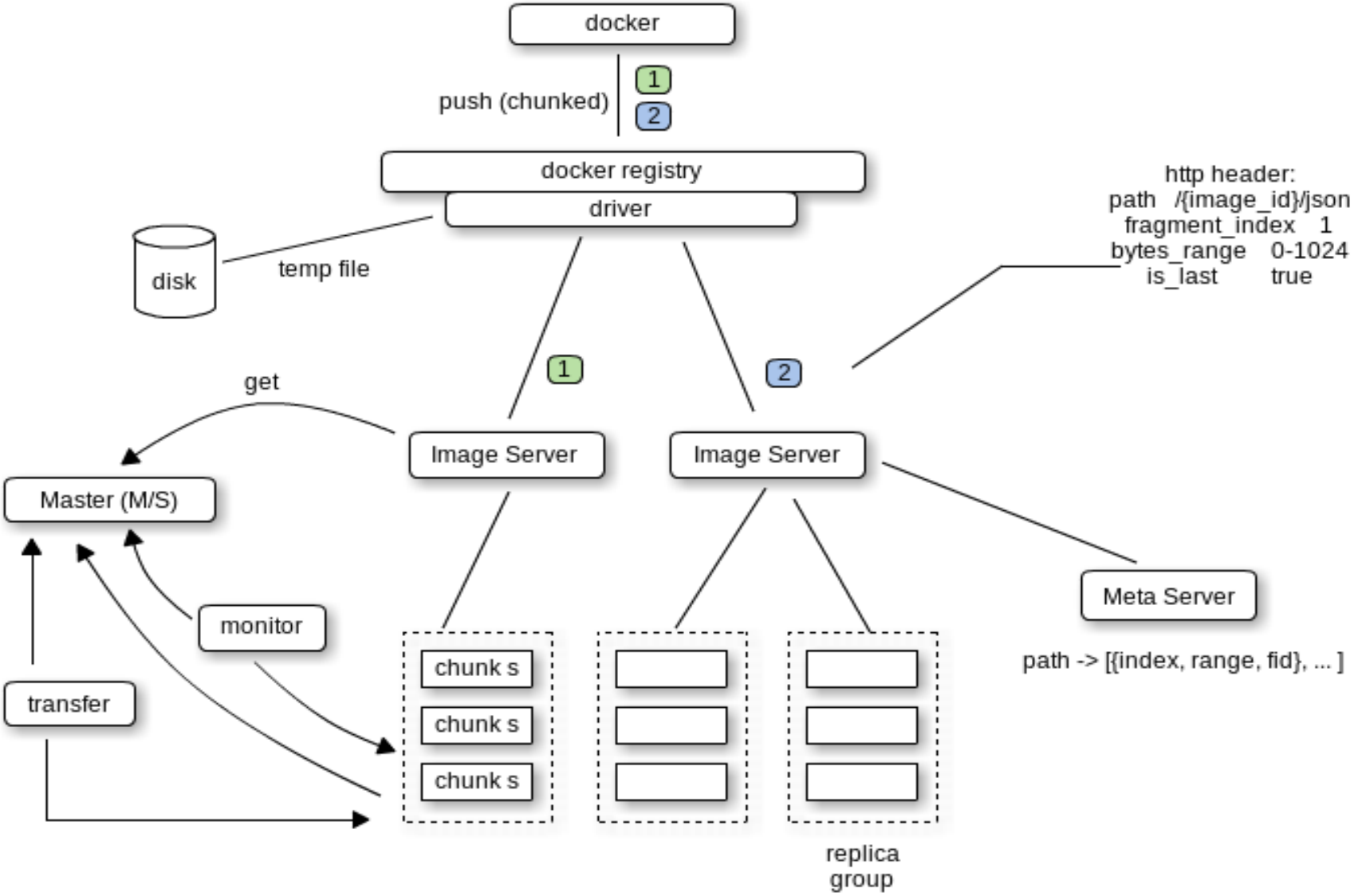
- 需要系统提供CoW
- 文件系统层
  - btrfs
- 叠合文件系统
  - aufs,overlayfs
- 块设备层
  - device mapper



- 内核dentry的游戏
  - merged/
  - work/
  - lower/
  - upper/
- 大文件的copy up会比较慢

JD Image Store

JD.COM 京东



小结

Part 2

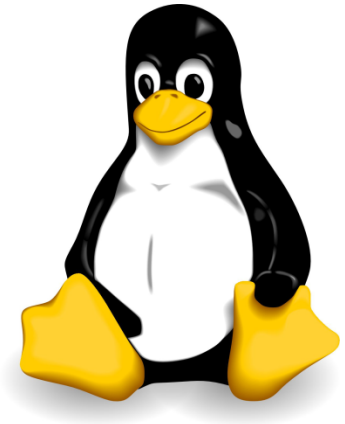
- 网络隔离
- 网络拓朴自定义
- IP资源动态管理与分配
- 网络流量的精细化运营
- 业务和基础网络的融合

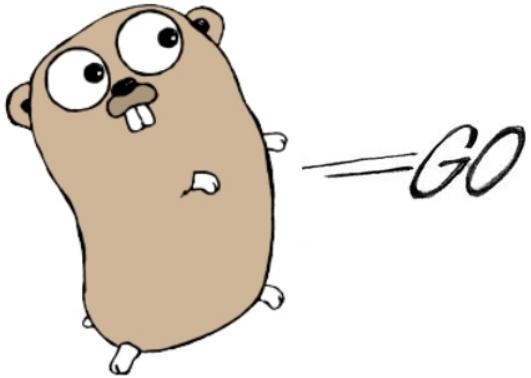
- 项目周期紧迫
- 落地经验不足
- 传统网络架构的平滑过渡
- 技术的实现

- 充分利用已有资源
- 大系统小做，快速迭代
- 避免过度依赖硬件设备
- 方案设计要冗余


OpenFlow


OvS
Open vSwitch



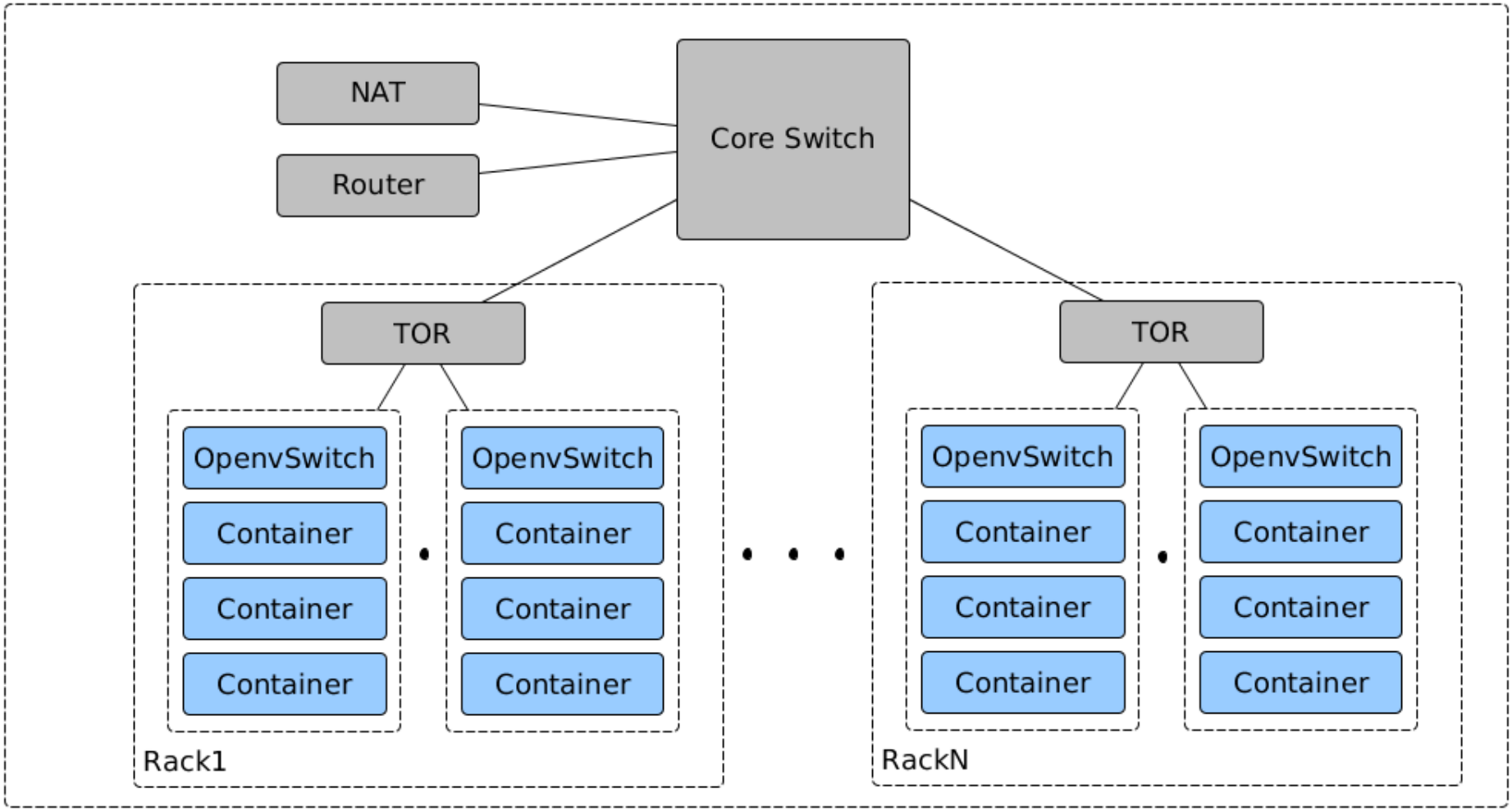

GO

JD.COM 京东

- 控制平面和数据平台分离
  - 集中控制，统一调度
- Overlay Network
  - L2 over L3
- 优化东西 、南北流量路径
  - 计算节点即网络节点
- 优化广播
  - 避免虚拟网络的广播流入承载网

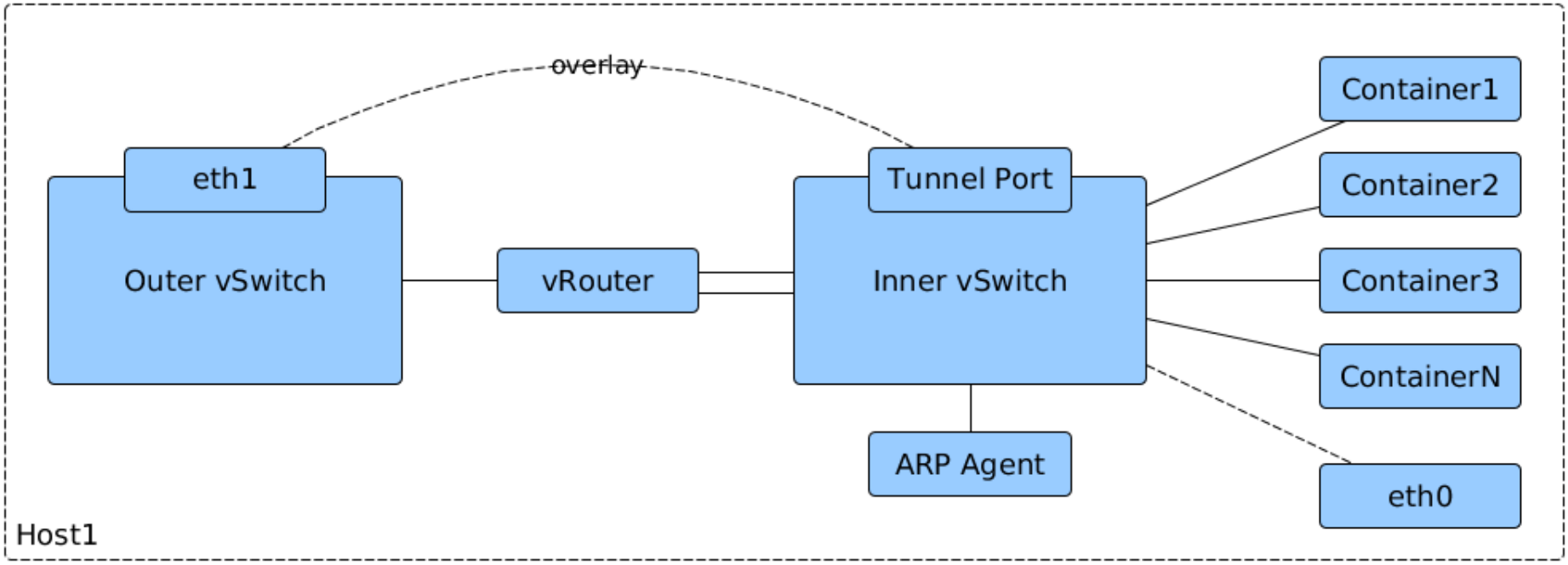
拓扑 - 概览

JD.COM 京东



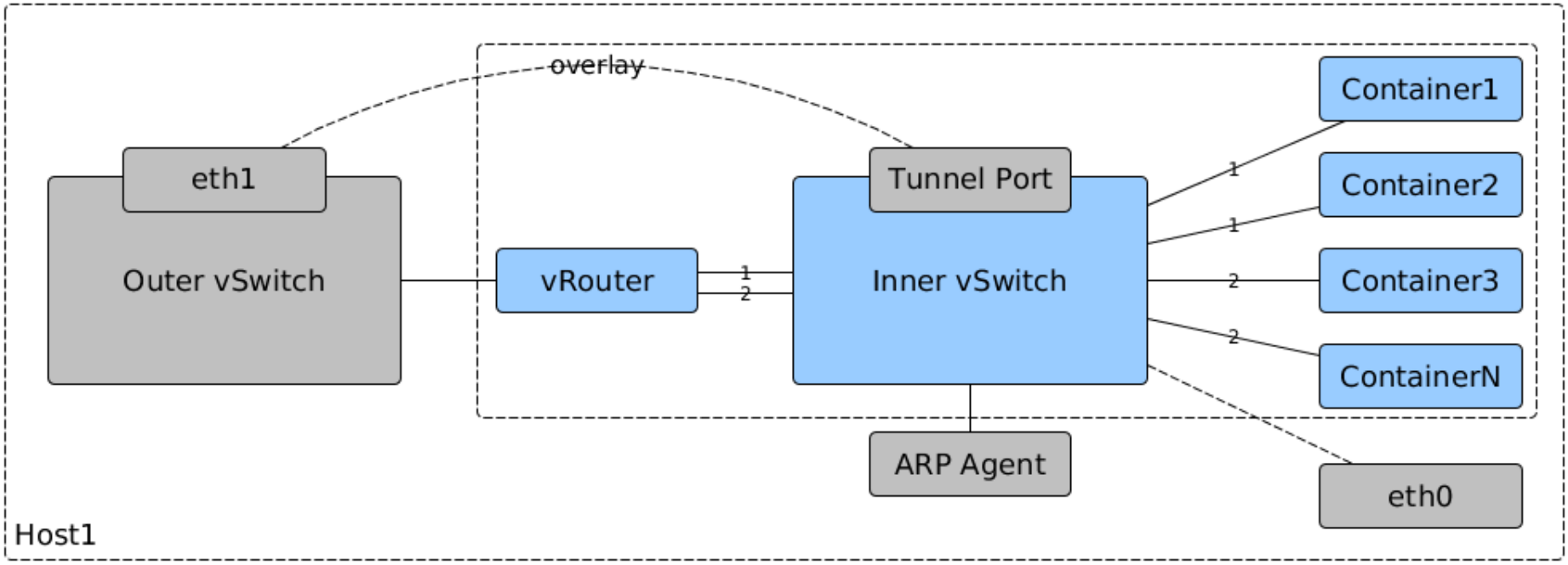
拓扑 - 节点内部

JD.COM 京东



拓扑 - 内部网络

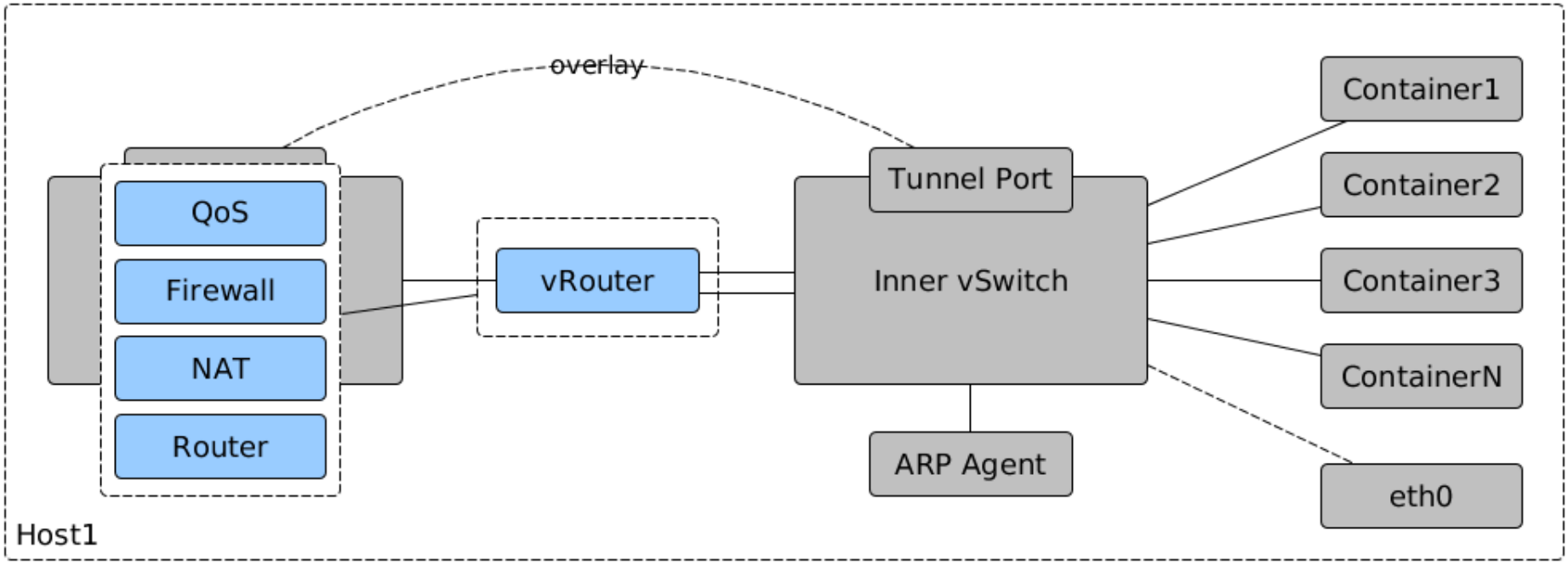
JD.COM 京东



Container1~N: *计算实例*
vRouter: *虚拟路由实例*
Inner vSwitch: *虚拟二层交换机支持Openflow协议*

拓扑 - 虚拟路由

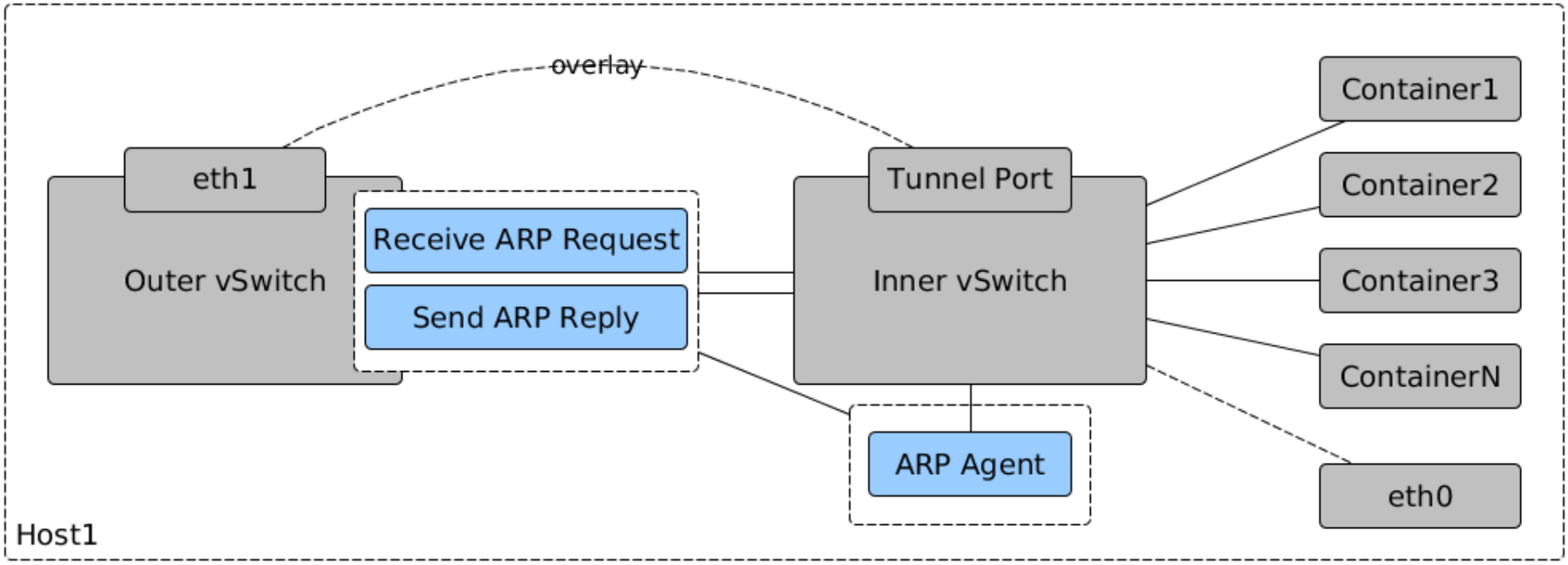
JD.COM 京东



QoS: *带宽、速率限制*
Firewall: *按自定义规则过滤上行及下行流量*
NAT: *源地址、目标地址转换*
Router: *IP包转发*

拓扑 – ARP Agent

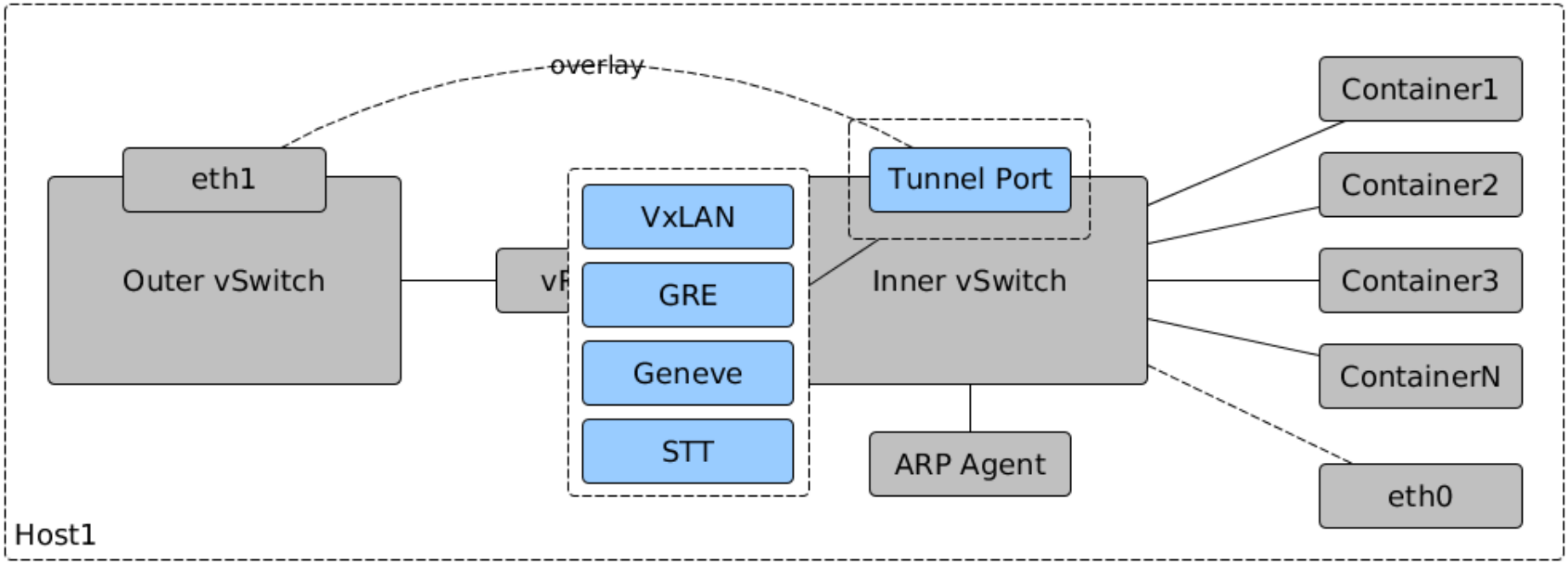
JD.COM 京东



Receive ARP Request: *响应内部虚拟交换机上所有的ARP广播请求*
Send ARP Reply: *根据广播请求的内容，返回正确的IP/MAC映射信息*

拓扑 - Tunnel

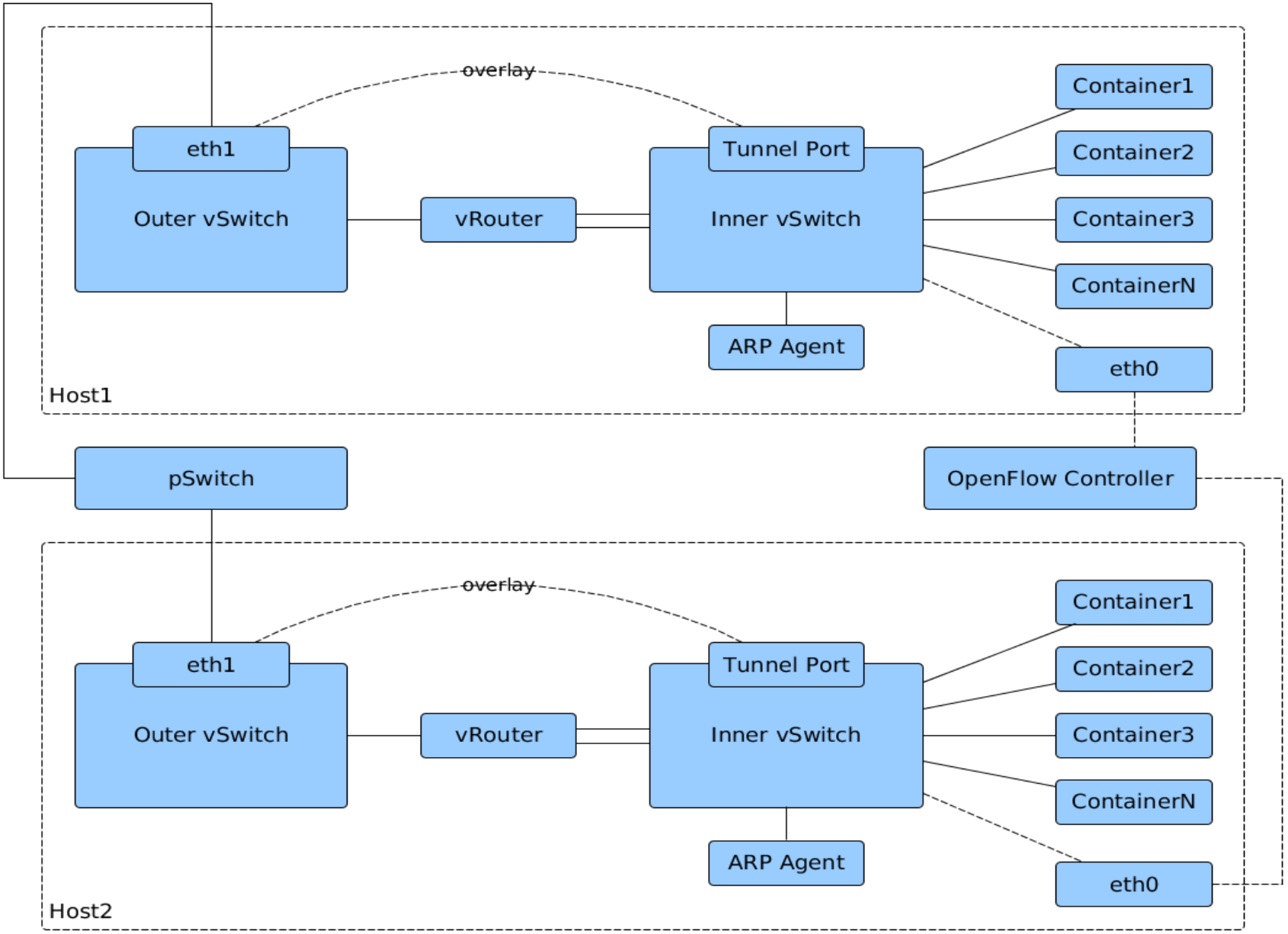
JD.COM 京东



Tunnel Port: *虚拟端口、用于对数据包进行隧道封装*
VxLAN: *Virtual Extensible LAN(supported)*
GRE: *Generic Routing Encapsulation(supported)*
Geneve: *Generic Network Virtualization Encapsulation(3.18)*
STT: *Stateless Transport Tunneling(patch)*

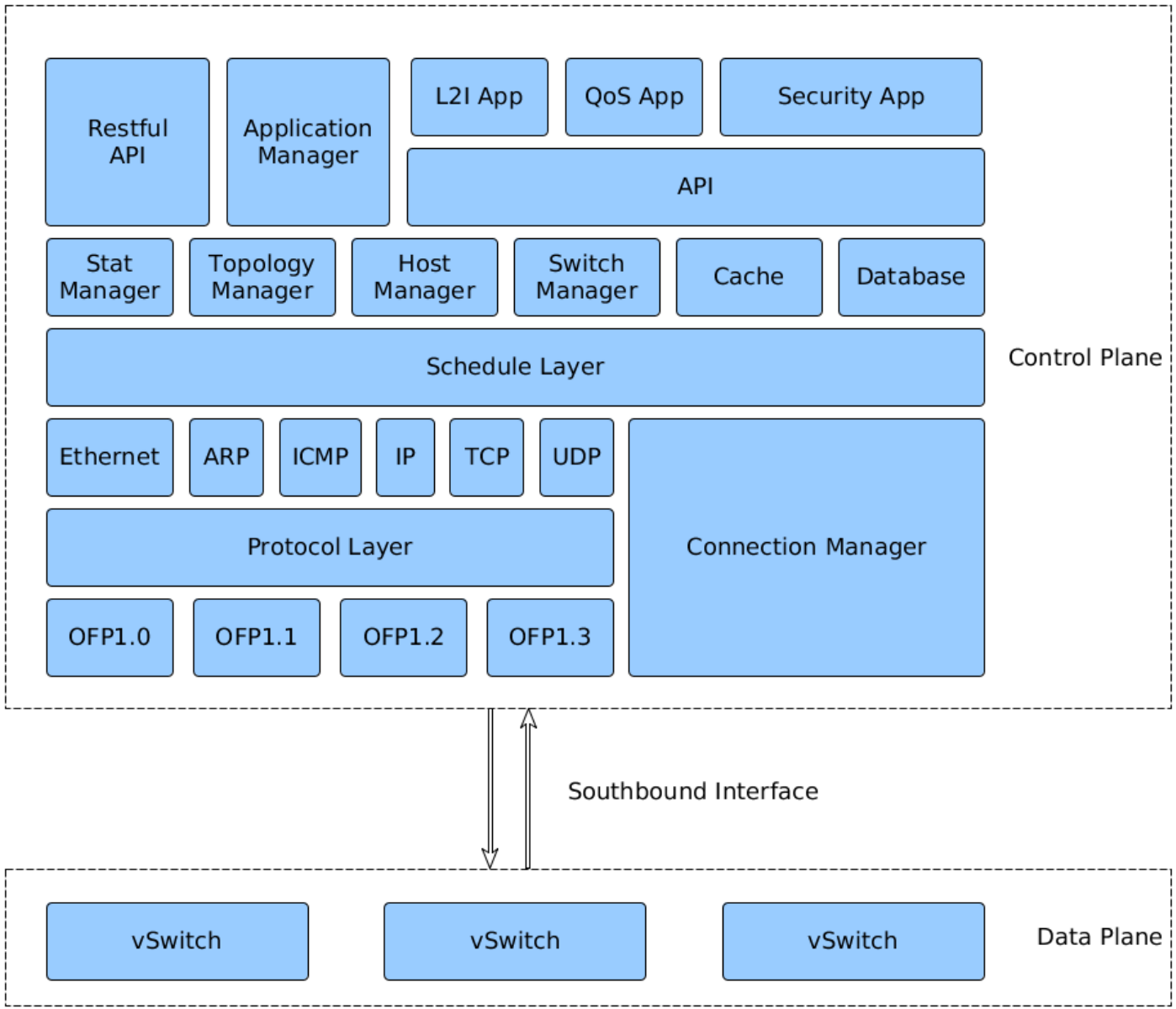
拓扑 – Multiple Host

JD.COM 京东



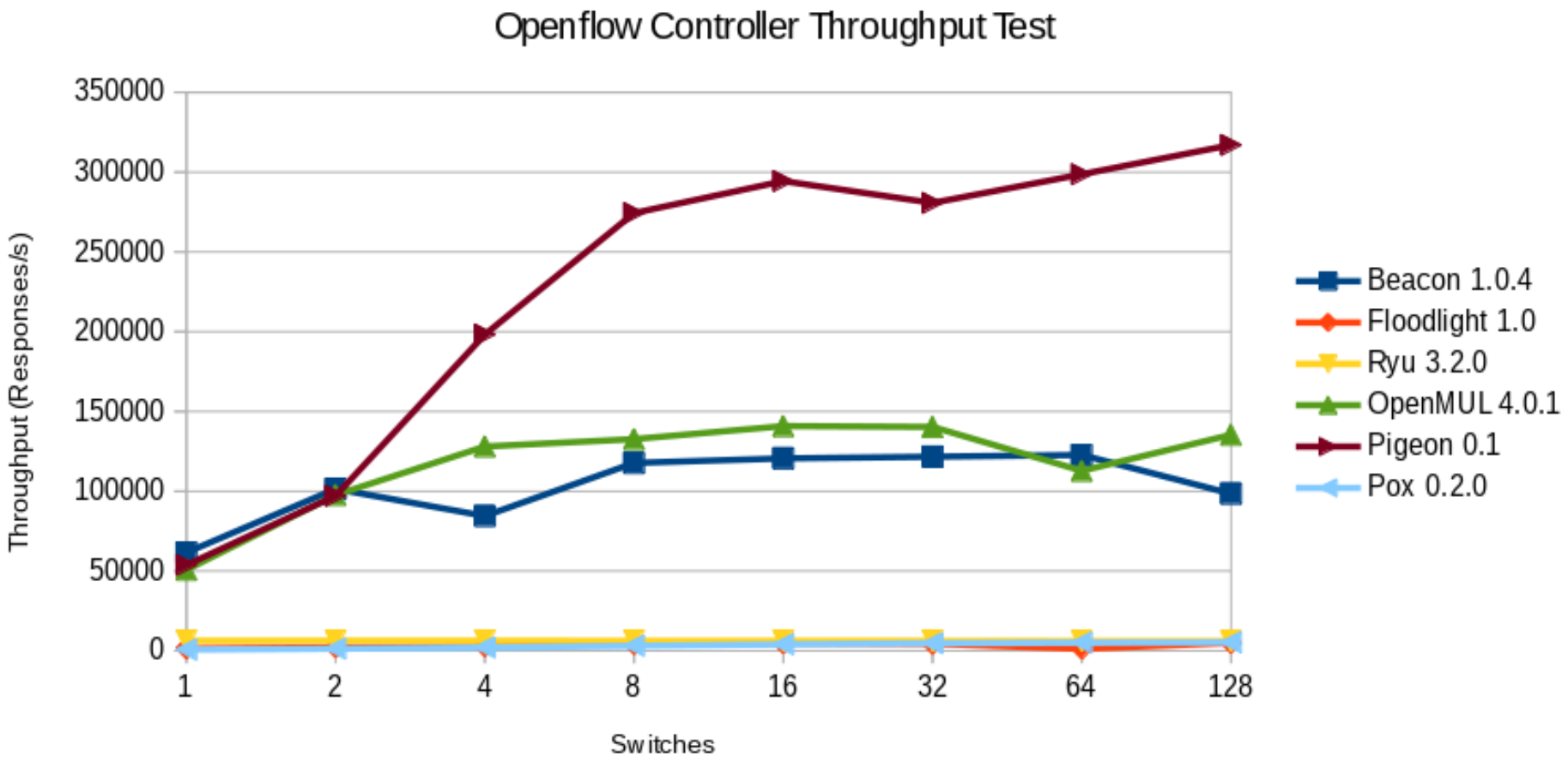
控制器 - 架构

JD.COM 京东



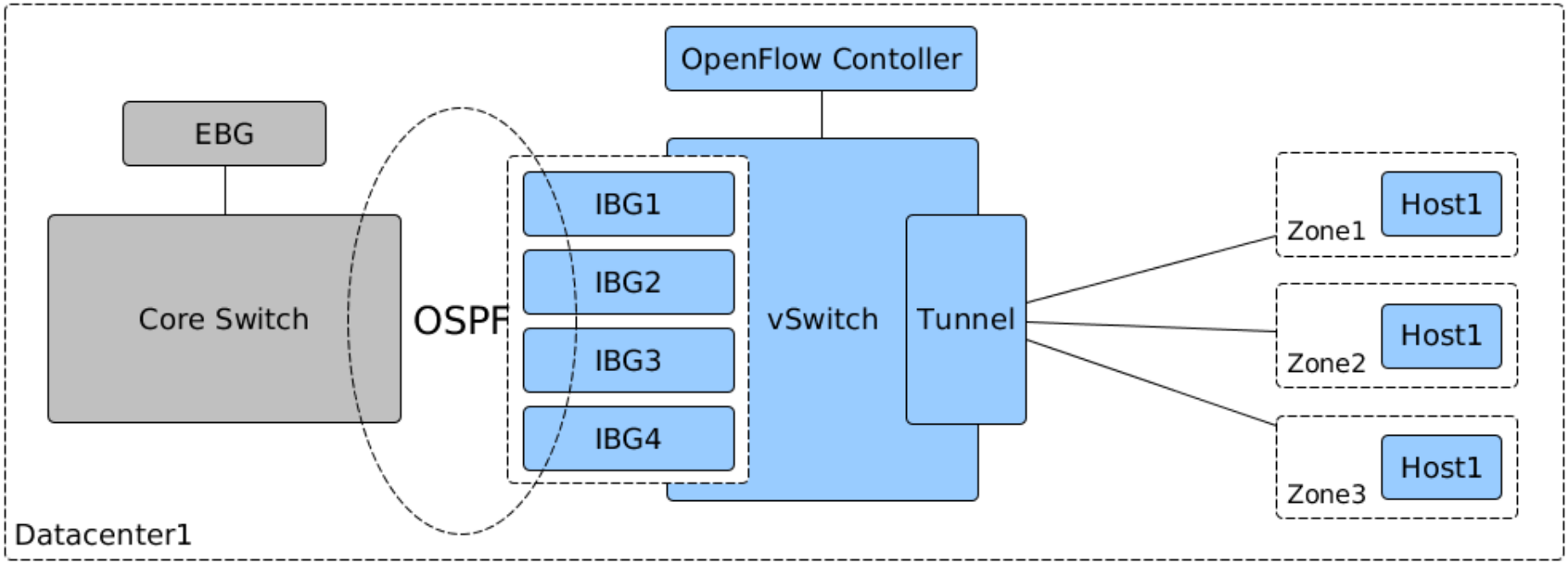
控制器 - 性能测试

JD.COM 京东



现在进行时– IBG

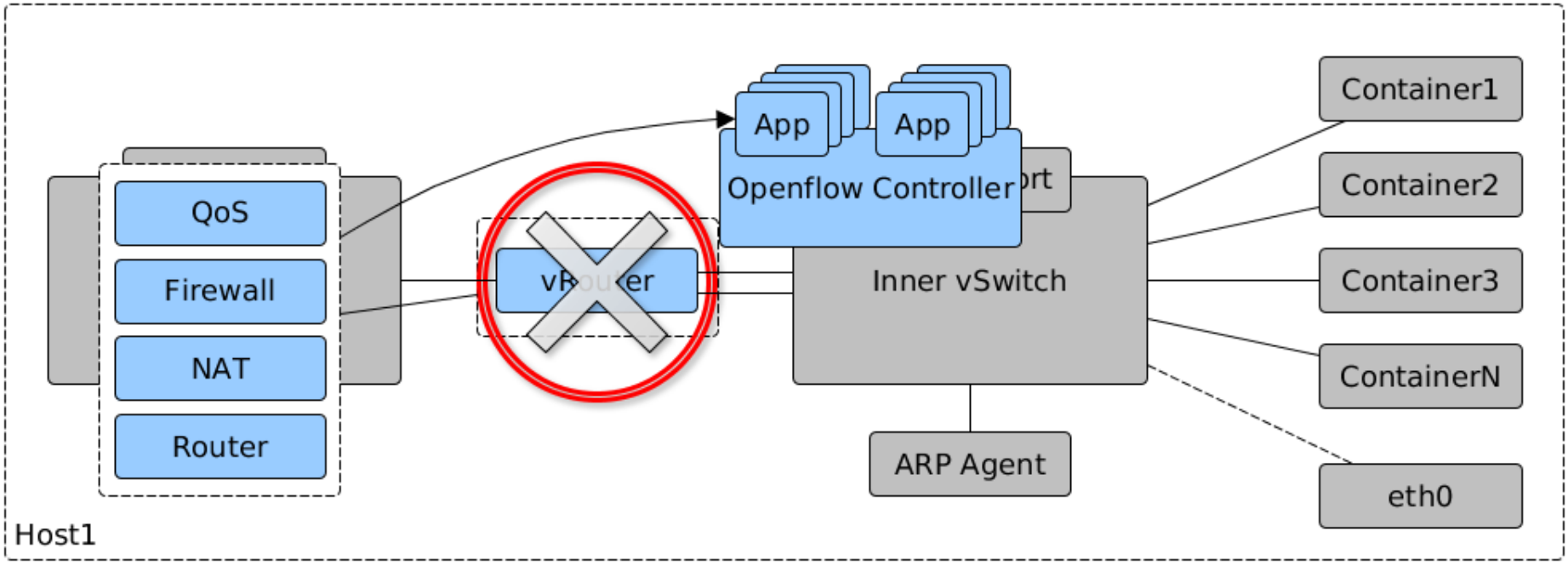
JD.COM 京东



IBG: *内部边界网关*
Zone1~N: 二层隔离的网络*(例如VLAN等)*
OSPF: *Equal-cost multi-path routing*
Quagga: *路由软件，支持OSPF, RIP, BGP等路由协议*

现在进行时 – vRouter

JD.COM 京东



将vRouter的功能以Openflow Controller应用进行交付，统一对网络操作的业务模型。

小结

Q&A

JD.COM 京东