

容器核心技术及SDN实践

田琪&闫国旗

促进软件开发领域知识与创新的传播



ArchSummit
全 球 架 构 师 峰 会

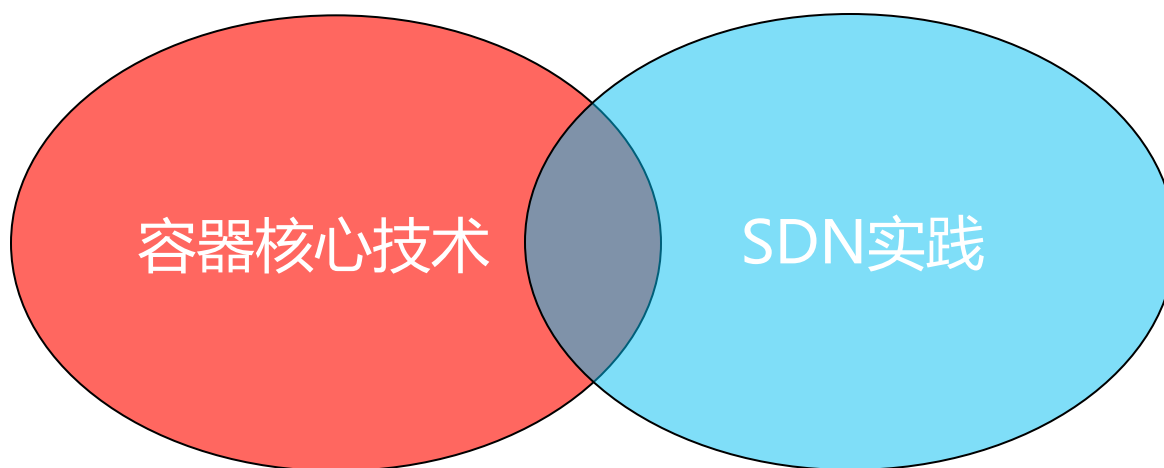
【深圳】2015年7月17日-18日

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

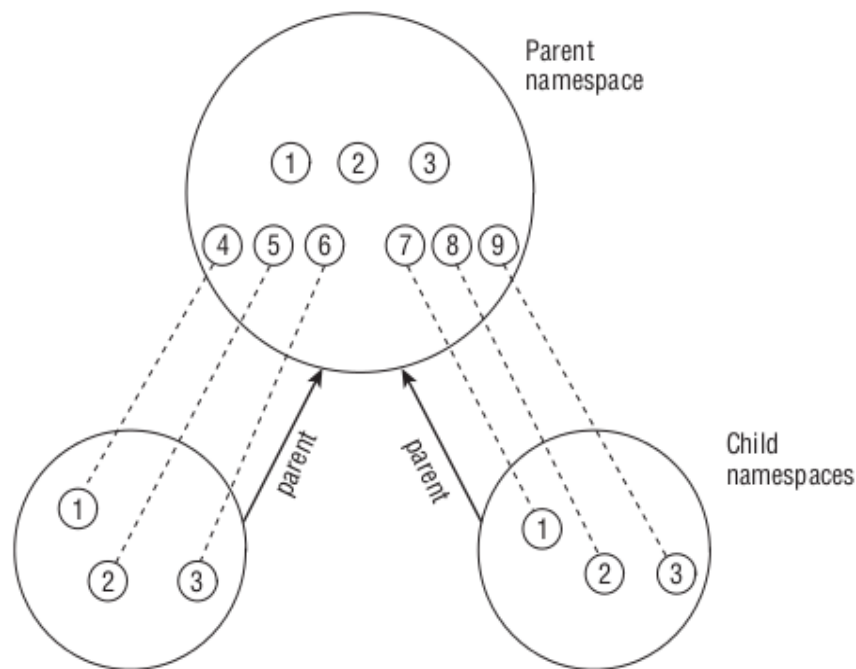


Part 1

- 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

- 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

- 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 */ );
```

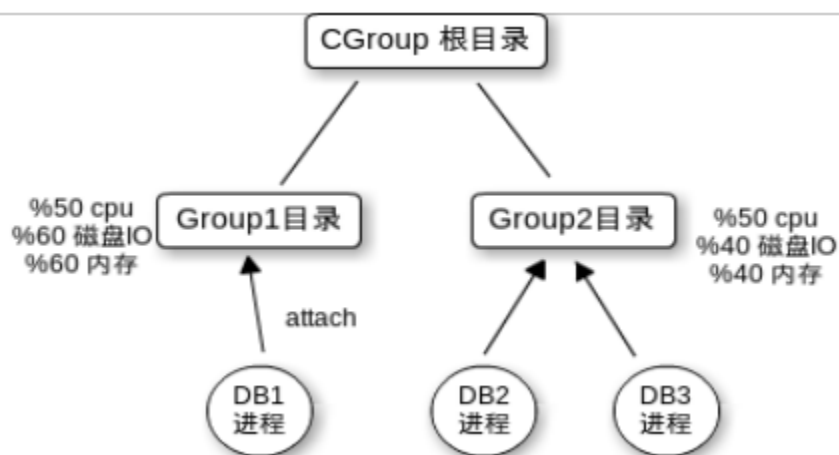
- 加入当前进程到新建namespace中

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

- 改变当前进程的namespace

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


- 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



block子系统 CFQ策略文件

blkio.weight_device
blkio.weight
blkio.leaf_weight_device
blkio.leaf_weight
blkio.time[_recursive]
blkio.sectors[_recursive]
blkio.io_merged[_recursive]
blkio.io_queued[_recursive]
blkio.io_wait_time[_recursive]
blkio.io_serviced[_recursive]
blkio.io_service_time[_recursive]
blkio.io_service_bytes[_recursive]

cpu 子系统相关文件

cpu.shares
cpu.cfs_quota_us
cpu.cfs_period_us
cpu.rt_runtime_us
cpu.rt_period_us
cpu.stat

block子系统 throttle策略文件

blkio.throttle.read_bps_device
blkio.throttle.write_bps_device
blkio.throttle.read_iops_device
blkio.throttle.write_iops_device
blkio.throttle.io_service_bytes
blkio.throttle.io_serviced

block子系统框架产生文件

blkio.reset_stats

cpu accounting 子系统相关文件

cpuacct.usage
cpuacct.stat
cpuacct.usage_percpu

security 子系统文件

devices.allow
devices.deny
devices.list

cpuset 子系统相关文件

cpuset.cpu_exclusive
cpuset.cpus
cpuset.mem_exclusive
cpuset.mem_hardwall
cpuset.memory_migrate
cpuset.memory_pressure
cpuset.memory_pressure_enabled
cpuset.memory_spread_page
cpuset.memory_spread_slab
cpuset.mems
cpuset.sched_load_balance
cpuset.sched_relax_domain_level

CGroup 框架相关文件

cgroup.clone_children
cgroup.event_control
cgroup.procs
notify_on_release
release_agent
tasks
cgroup.sane_behavior

memory 子系统相关文件

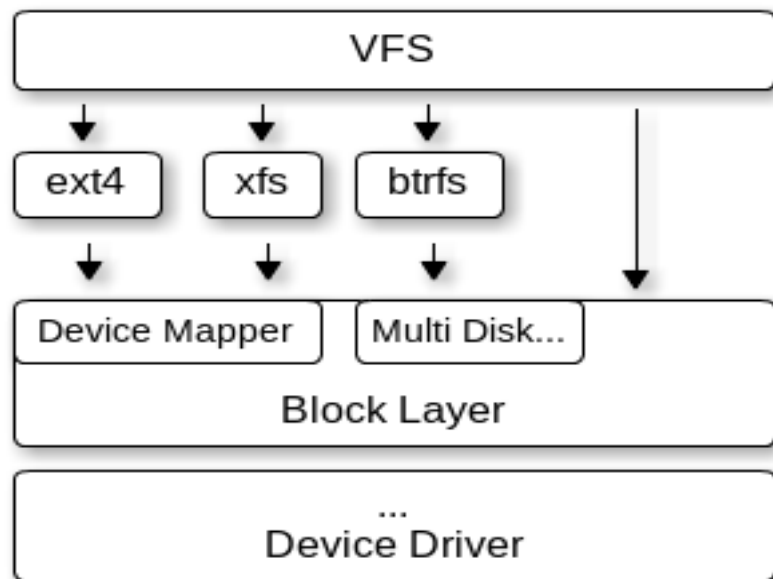
memory.failcnt
memory.force_empty
memory.limit_in_bytes
memory.max_usage_in_bytes
memory.move_charge_at_immigrate
memory.numa_stat
memory.oom_control
memory.pressure_level
memory.soft_limit_in_bytes
memory.use_hierarchy
memory.swappiness
memory.usage_in_bytes
memory.stat

hugetlb 相关文件

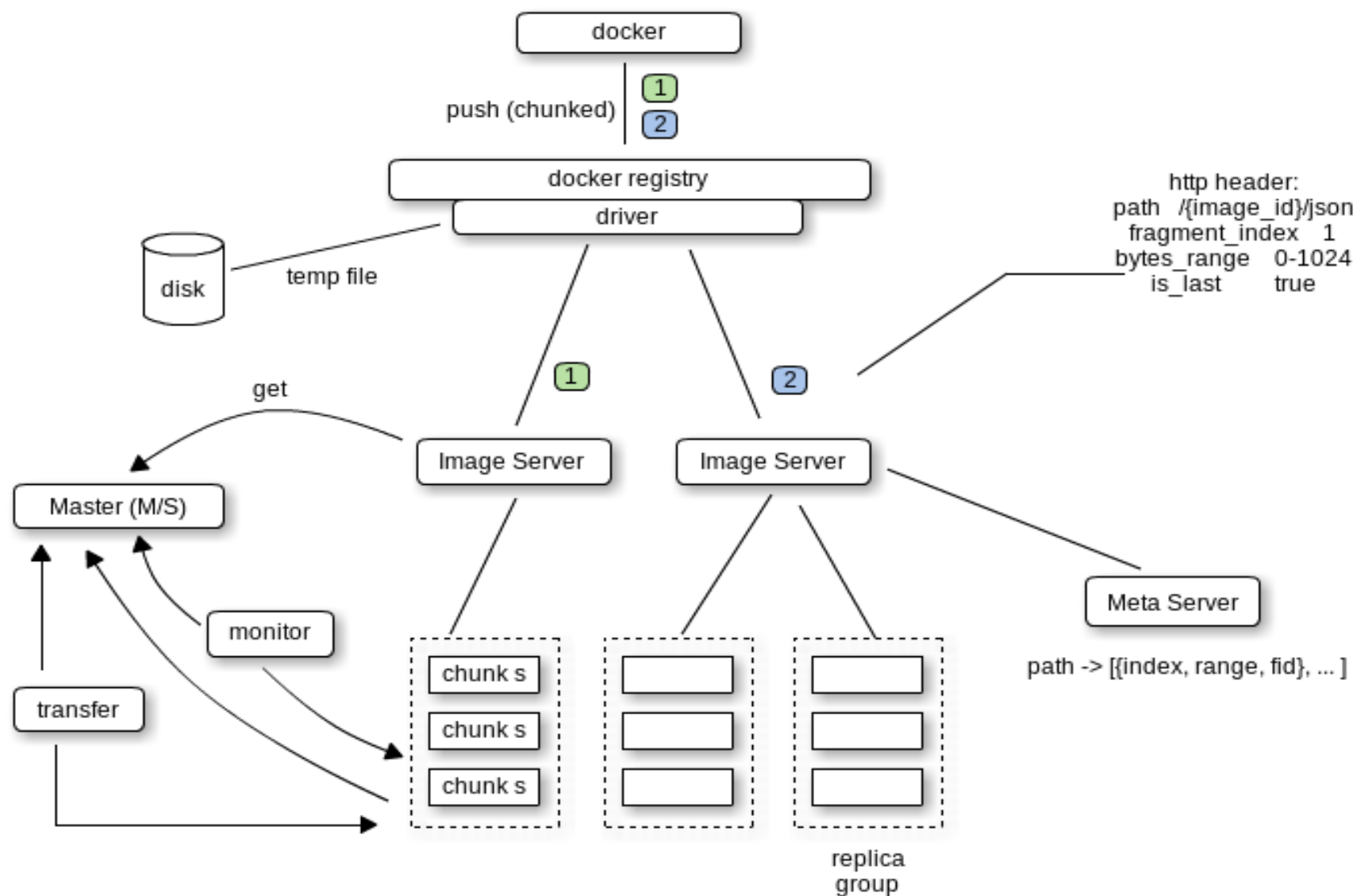
hugetlb.2MB.failcnt
hugetlb.2MB.limit_in_bytes
hugetlb.2MB.max_usage_in_bytes
hugetlb.2MB.usage_in_bytes

- 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



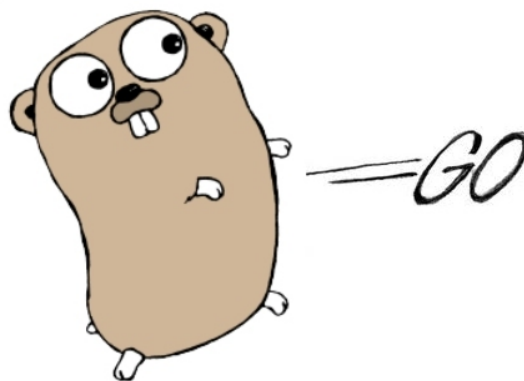
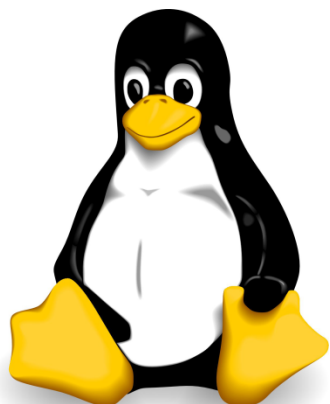
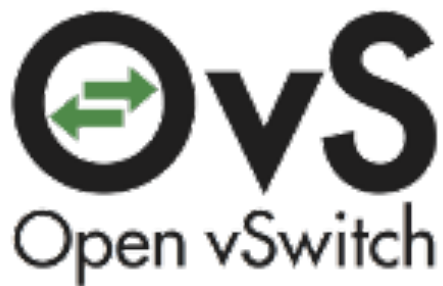
小结

Part 2

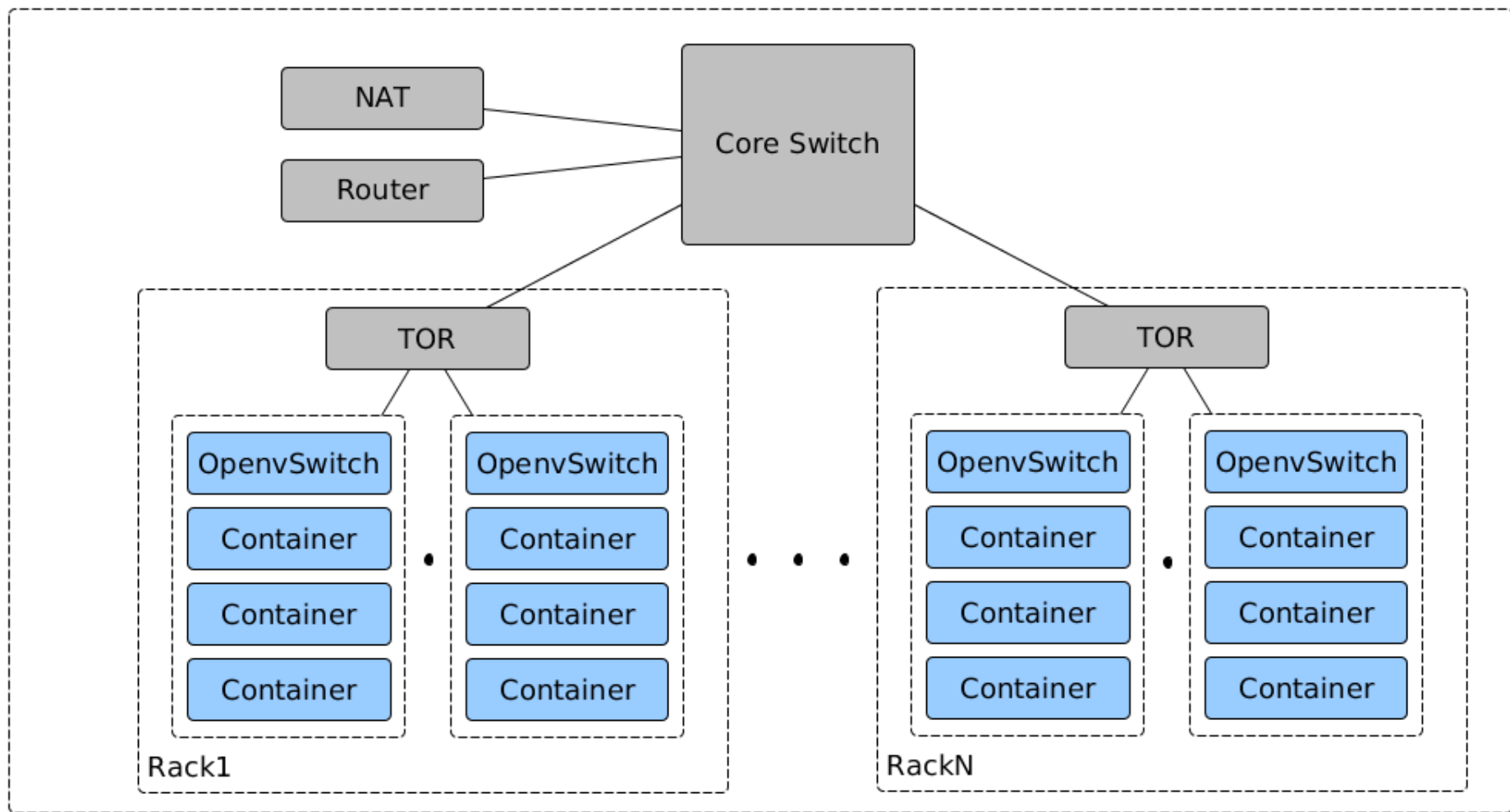
- 网络隔离
- 网络拓扑自定义
- IP资源动态管理与分配
- 网络流量的精细化运营
- 业务和基础网络的融合

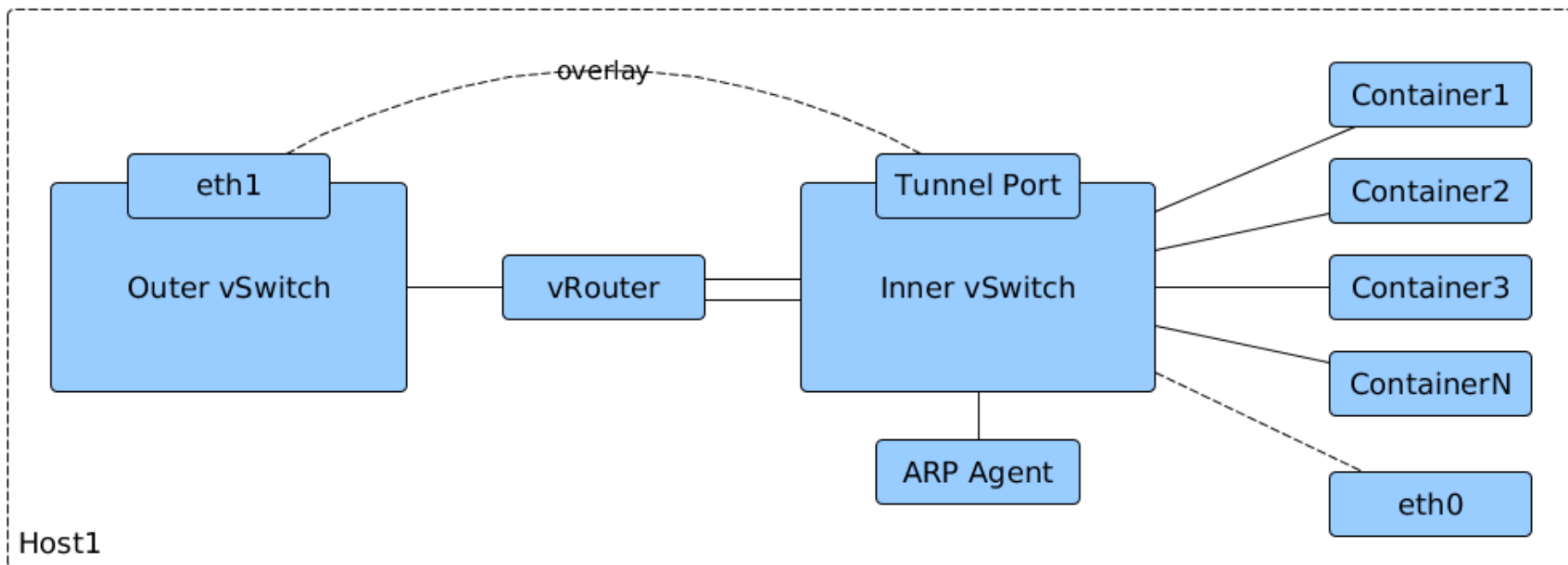
- 项目周期紧迫
- 落地经验不足
- 传统网络架构的平滑过渡
- 技术的实现

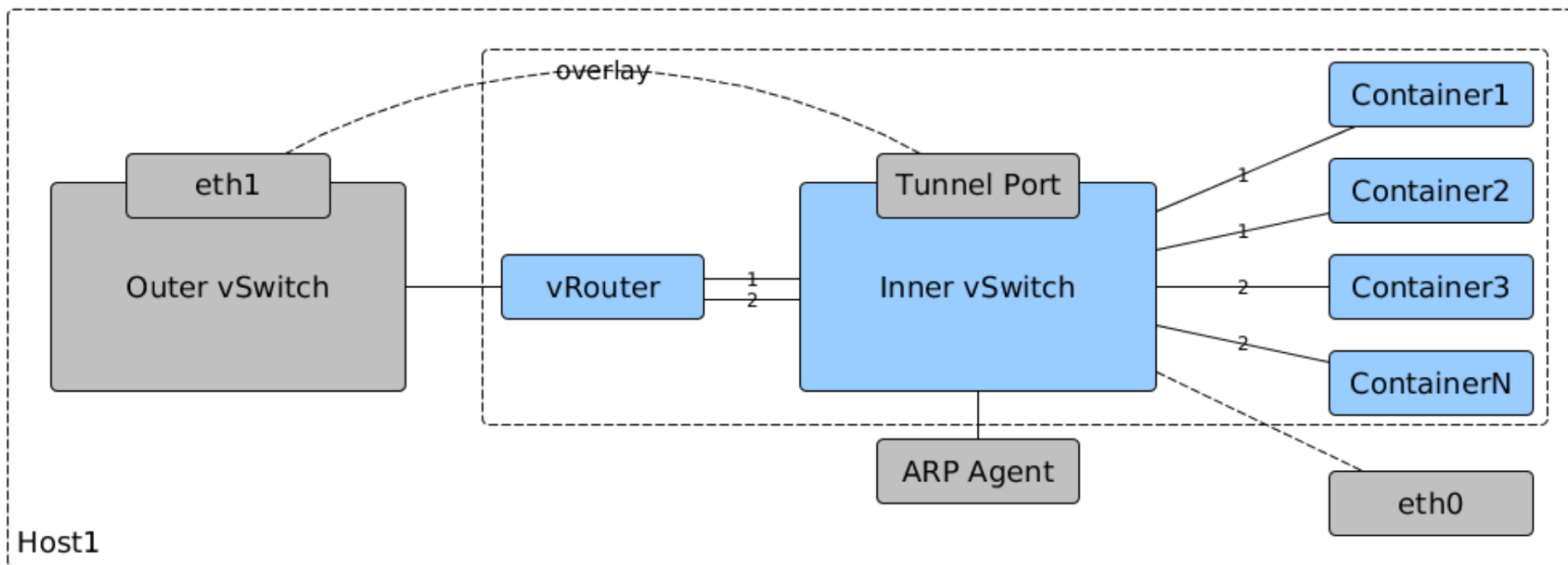
- 充分利用已有资源
- 大系统小做，快速迭代
- 避免过度依赖硬件设备
- 方案设计要冗余



- 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



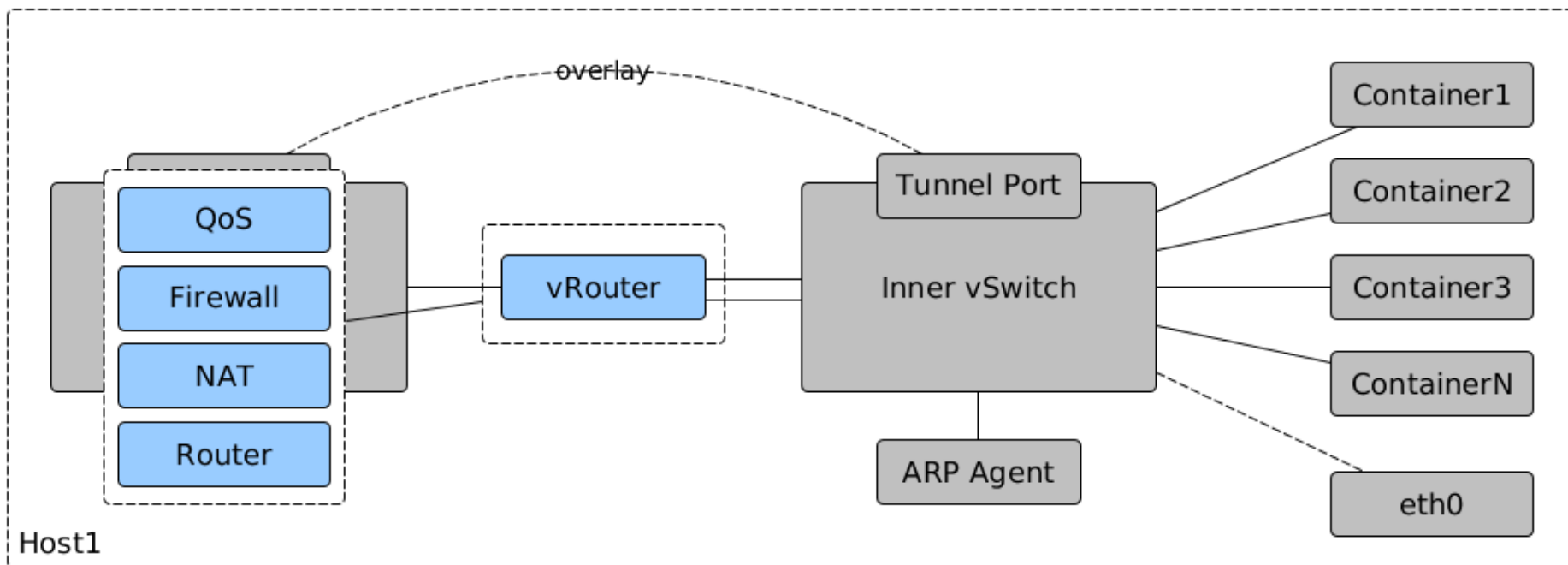




Container1~N: 计算实例

vRouter: 虚拟路由实例

Inner vSwitch: 虚拟二层交换机支持Openflow协议

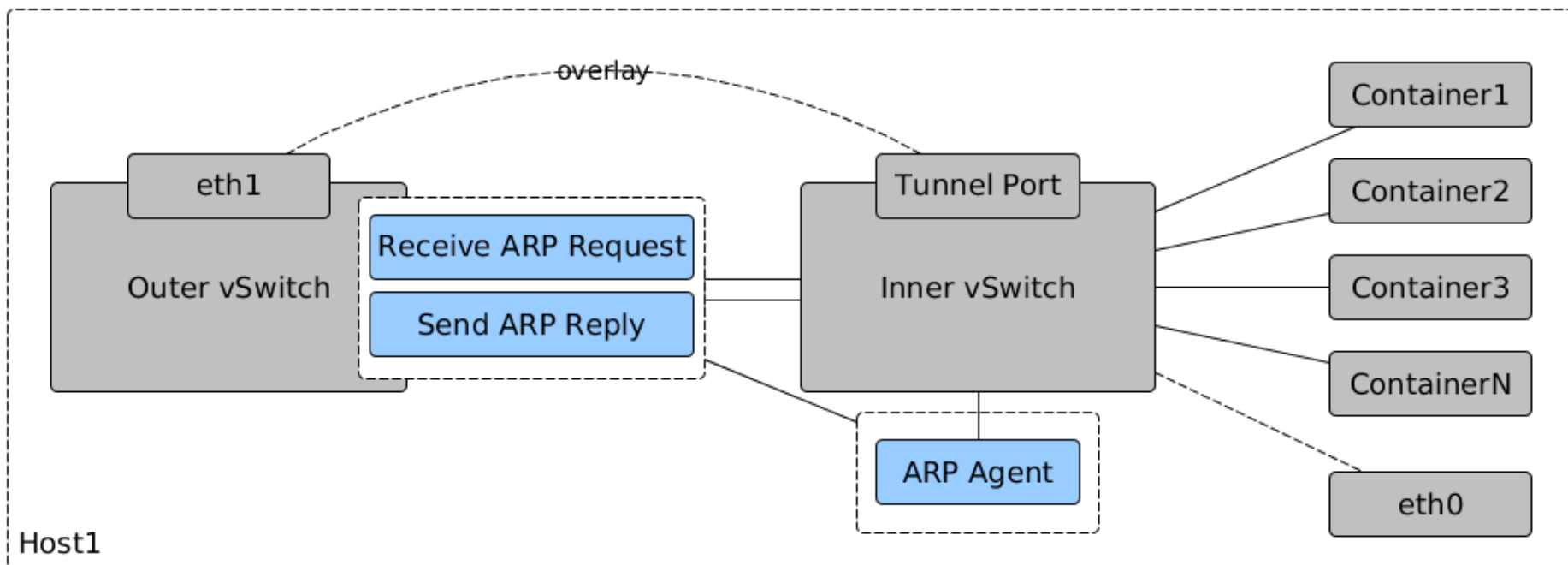


QoS: 带宽、速率限制

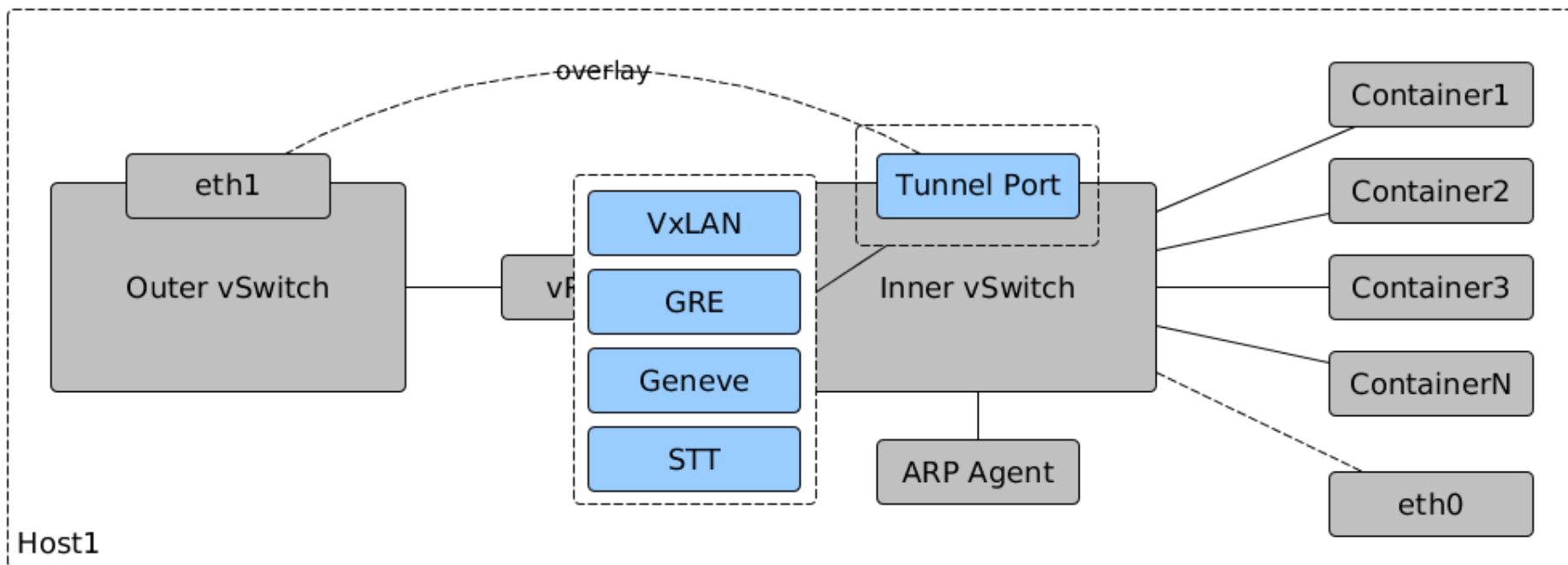
Firewall: 按自定义规则过滤上行及下行流量

NAT: 源地址、目标地址转换

Router: IP包转发



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



Tunnel Port: 虚拟端口、用于对数据包进行隧道封装

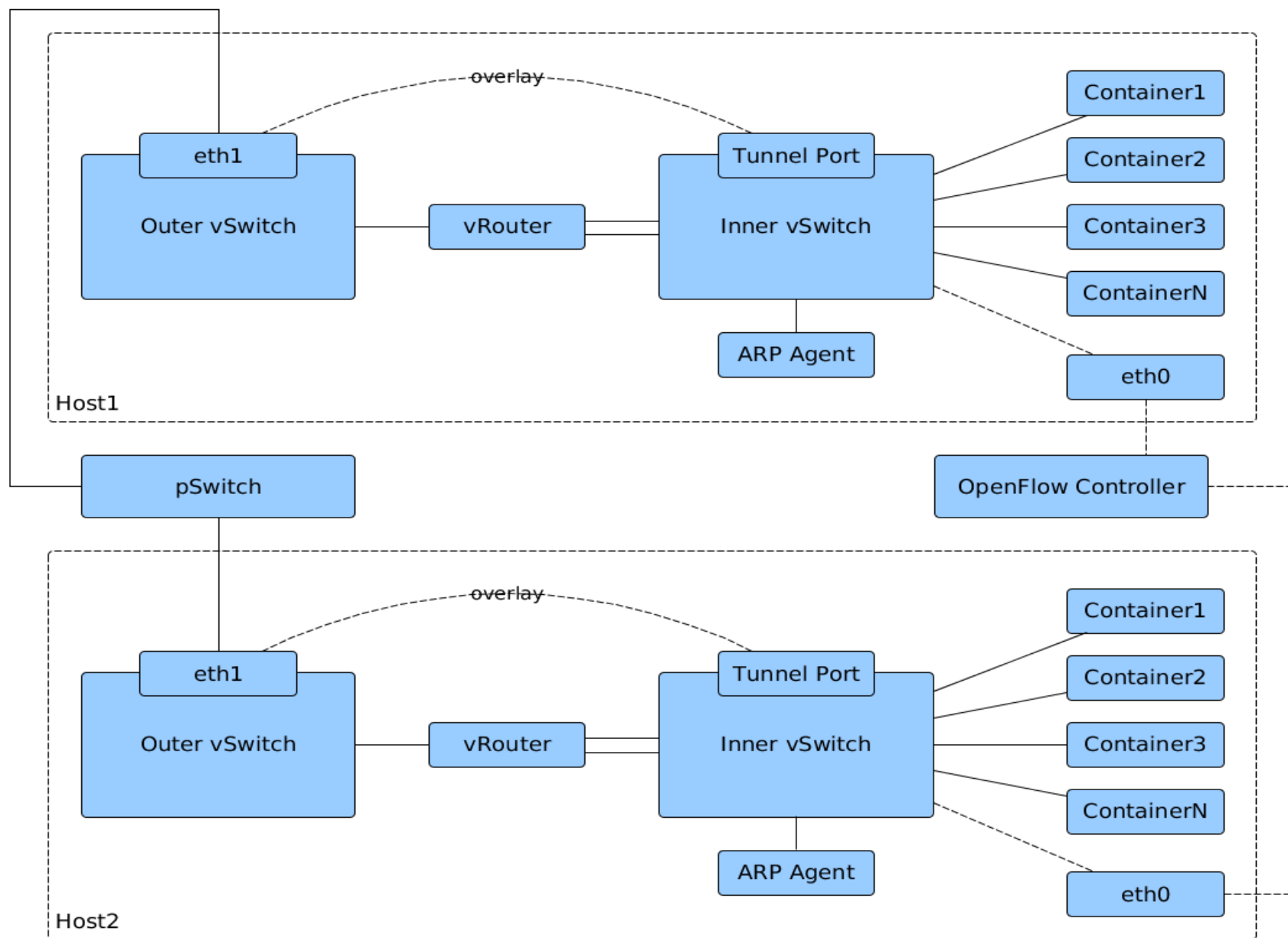
VxLAN: Virtual Extensible LAN(supported)

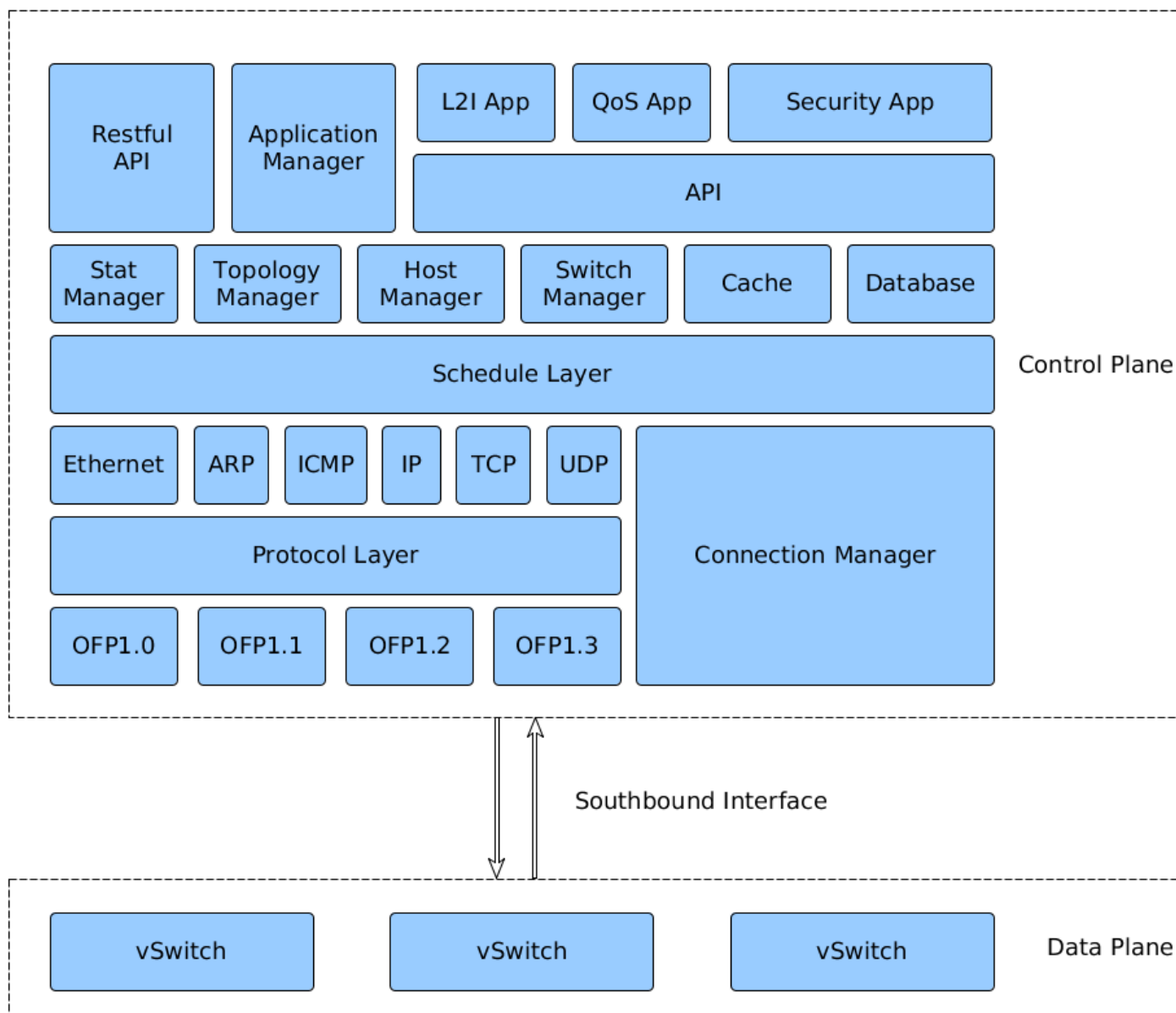
GRE: Generic Routing Encapsulation(supported)

Geneve: Generic Network Virtualization Encapsulation(3.18)

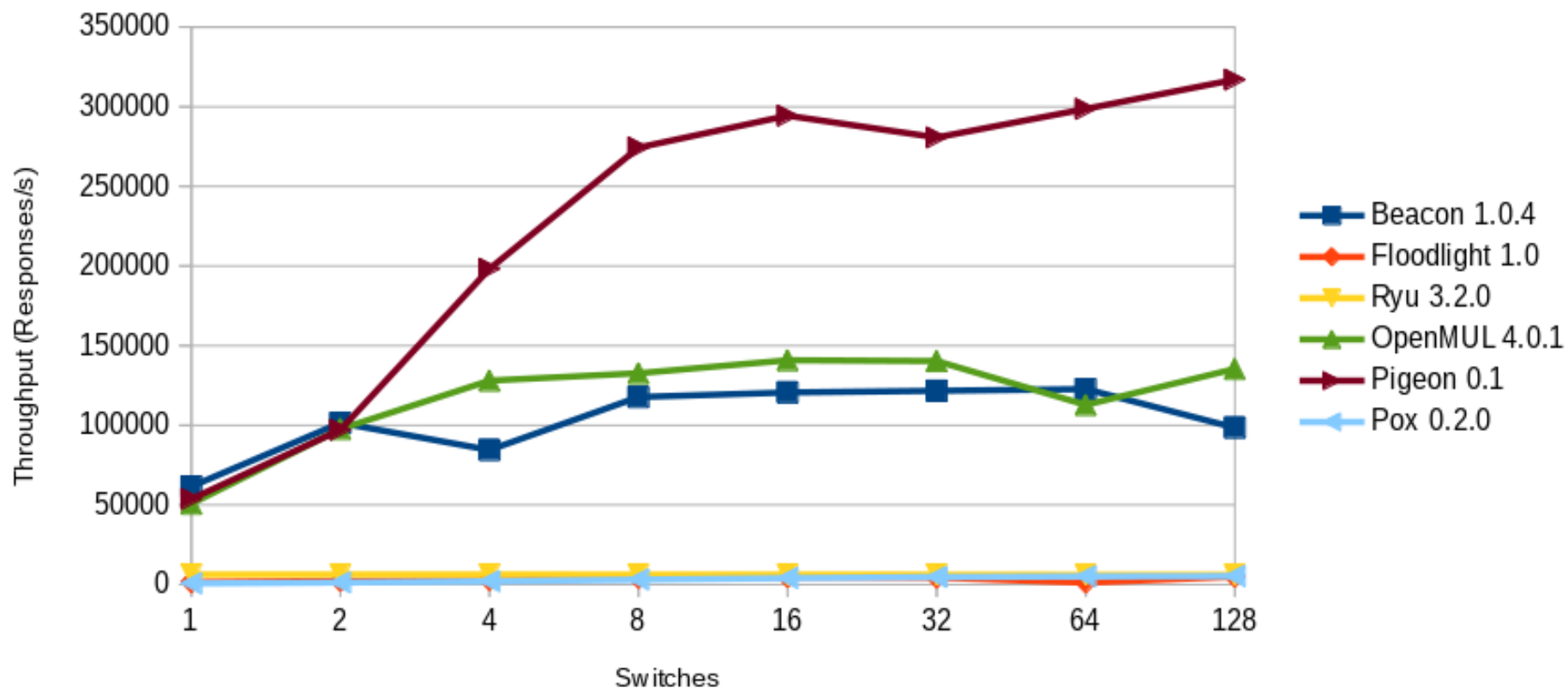
STT: Stateless Transport Tunneling(patch)

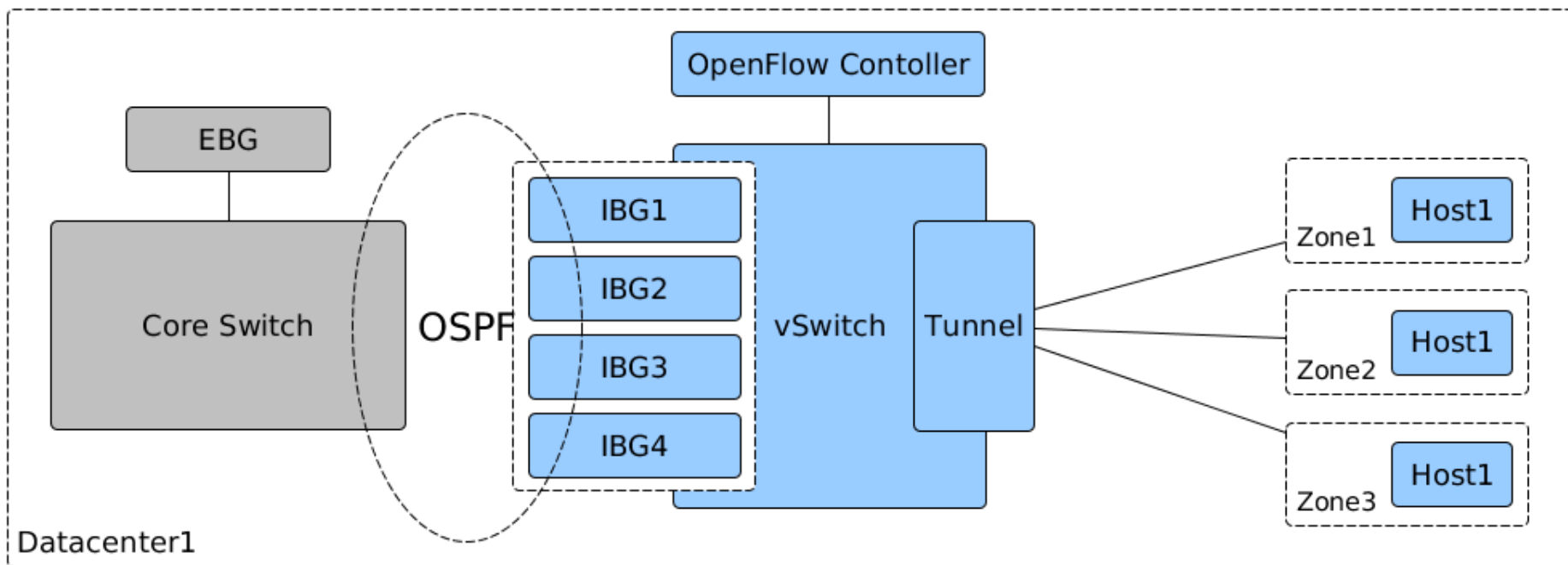
拓扑 – Multiple Host





Openflow Controller Throughput Test





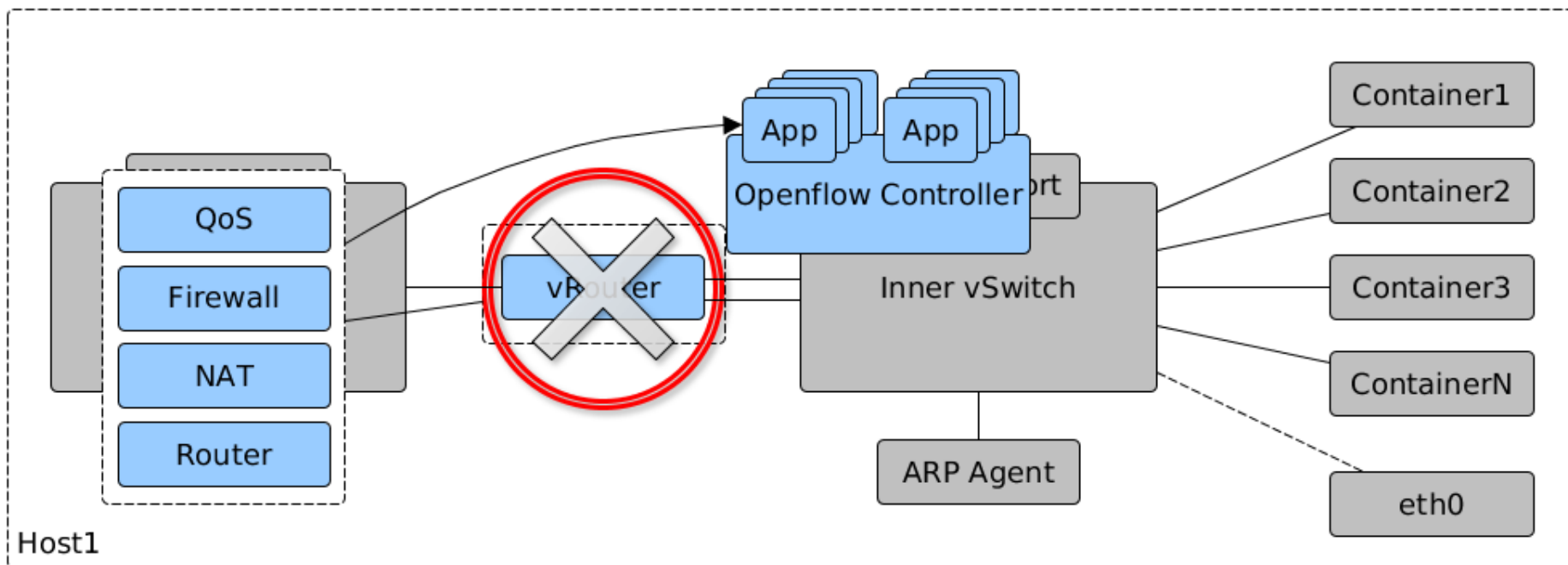
IBG: 内部边界网关

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



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

InfoQ^{ueue}

专注中高端技术人员的
社区媒体



EGO^{ueue} EXTRA GEEKS' ORGANIZATION
NETWORKS

高端技术人员
学习型社交网络



StuQ^{ueue}

实践驱动的
IT职业学习和服务平台



极客邦科技

InfoQ | EGO | StuQ

让技术人学习和交流更简单