SACC 2014中国系统架构师大会
SYSTEM ARCHITECT CONFERENCE CHINA 2014
发现架构之美

# 网易私有云网络虚拟化实践

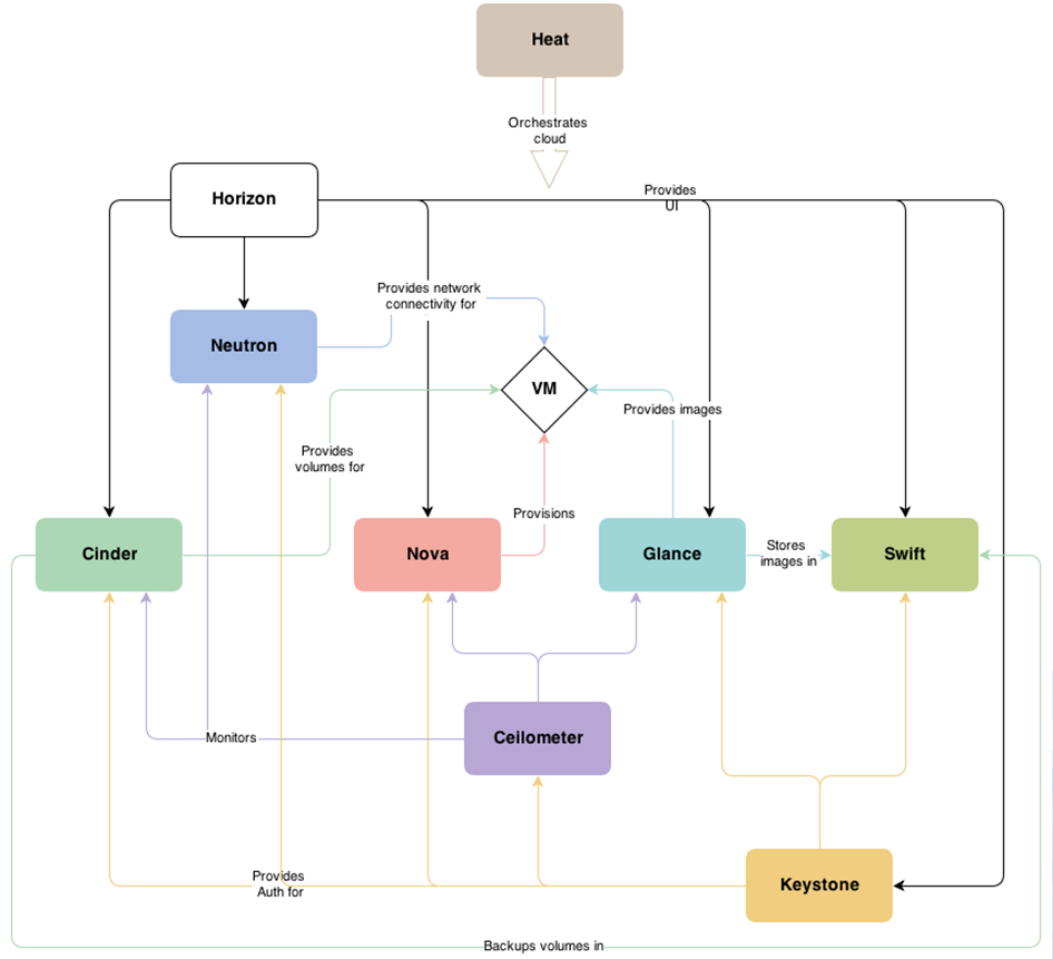
网易杭州研究院
徐城利

# 提纲

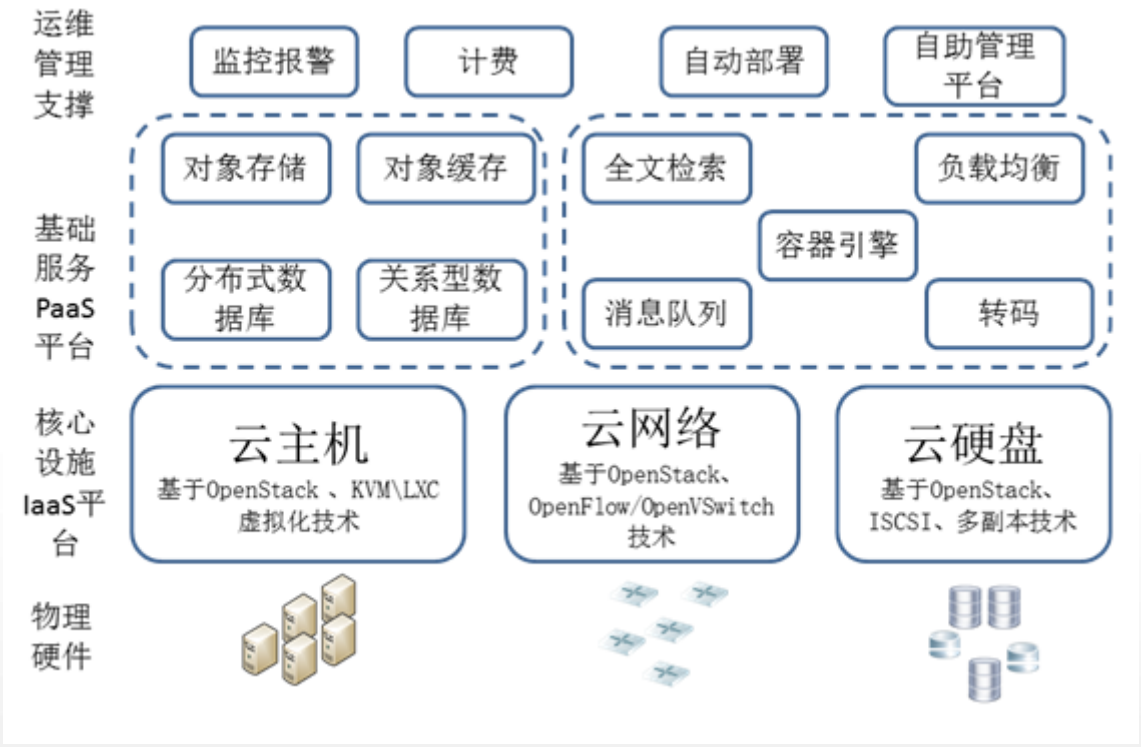
- OpenStack介绍
- 网易私有云简介
- 第一代网络虚拟化服务
- 第二代网络虚拟化服务
- 碰到的问题及一些经验总结

# OpenStack

- 计算 (nova)
- 网络 (neutron)
- 认证 (keystone)
- 镜像 (glance)
- 块存储 (cinder)
- 对象存储 (swift)
- 其他模块



# 网易私有云


运维管理支撑
监控报警
计费
自动部署
自助管理平台
基础服务PaaS平台
对象存储
对象缓存
全文检索
负载均衡
容器引擎
分布式数据库
关系型数据库
消息队列
转码
核心设施IaaS平台
云主机
基于OpenStack 、KVM\LXC虚拟化技术
云网络
基于OpenStack、OpenFlow/OpenVSwitch技术
云硬盘
基于OpenStack、ISCSI、多副本技术
物理硬件

# 网易私有云(Cont'd)

- 12年年底正式上线，已稳定运行近2年
- 部署两个Region
- 100+物理节点
- 已广泛服务于网易公司产品（30+）如：门户、易信、网易新闻客户端、云音乐、云课堂/公开课、云阅读等

# 网易私有云(Cont'd)

- 提高硬件资源使用率

  如CPU: ~10% -> ~50%
- 提高物理资源管理及运维自动化
  - 自助服务
  - 运维人员减少1/2
- 提高基础资源弹性
  - 资源池：快速申请，按需使用
  - 适应业务波动
  - 满足开发、测试等零碎需求

# IaaS平台研发

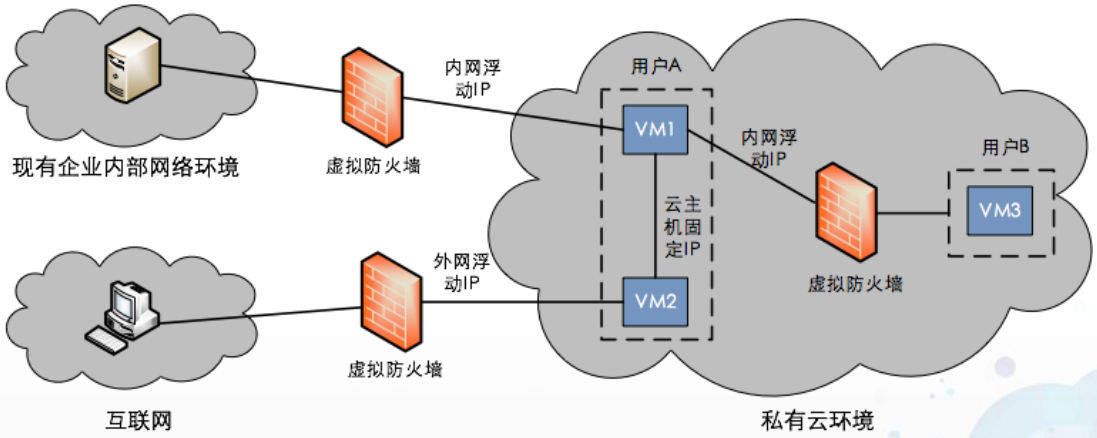
- 三个基础服务
  - 云主机：提供可扩展、安全可靠的弹性计算
  - 云网络：提供动态、安全、灵活的网络服务
  - 云硬盘：提供可扩展、安全稳定的块存储服务
- 研发工作
  - 基于OpenStack的nova/neutron/keystone/glance/cinder开发
  - 对OpenStack作充分的功能、性能、稳定性及异常测试
  - 发现并修复OpenStack社区的bug (提交90+、修复50+)
  - 根据公司需求，研发新功能及优化30+
  - 支持整合网易私有云10多个上层服务

# 云网络演进

- 第一代网络虚拟化服务
  - 基于nova-network
  - 使用Flat DHCP Network Manager
  - 定制开发
- 第二代网络虚拟化服务
  - 基于neutron
  - 使用ML2 + Open vSwitch Agnt
  - 应用SDN (Software-Defined Networking)技术
  - 定制开发

# 第一代云网络

- 基于nova-network，简单可靠
- 固定IP (Fixed IP)、浮动IP (Floating IP)
- 安全组 (Security Groups)
- 租户网络隔离
- QoS



# 第一代云网络(Cont'd)

- 技术选型
  - Flat DHCP Network Manager
  - Multi-host
- 自研功能及优化
  - 内网浮动IP
  - 租户网络优化与隔离
  - 网络QoS

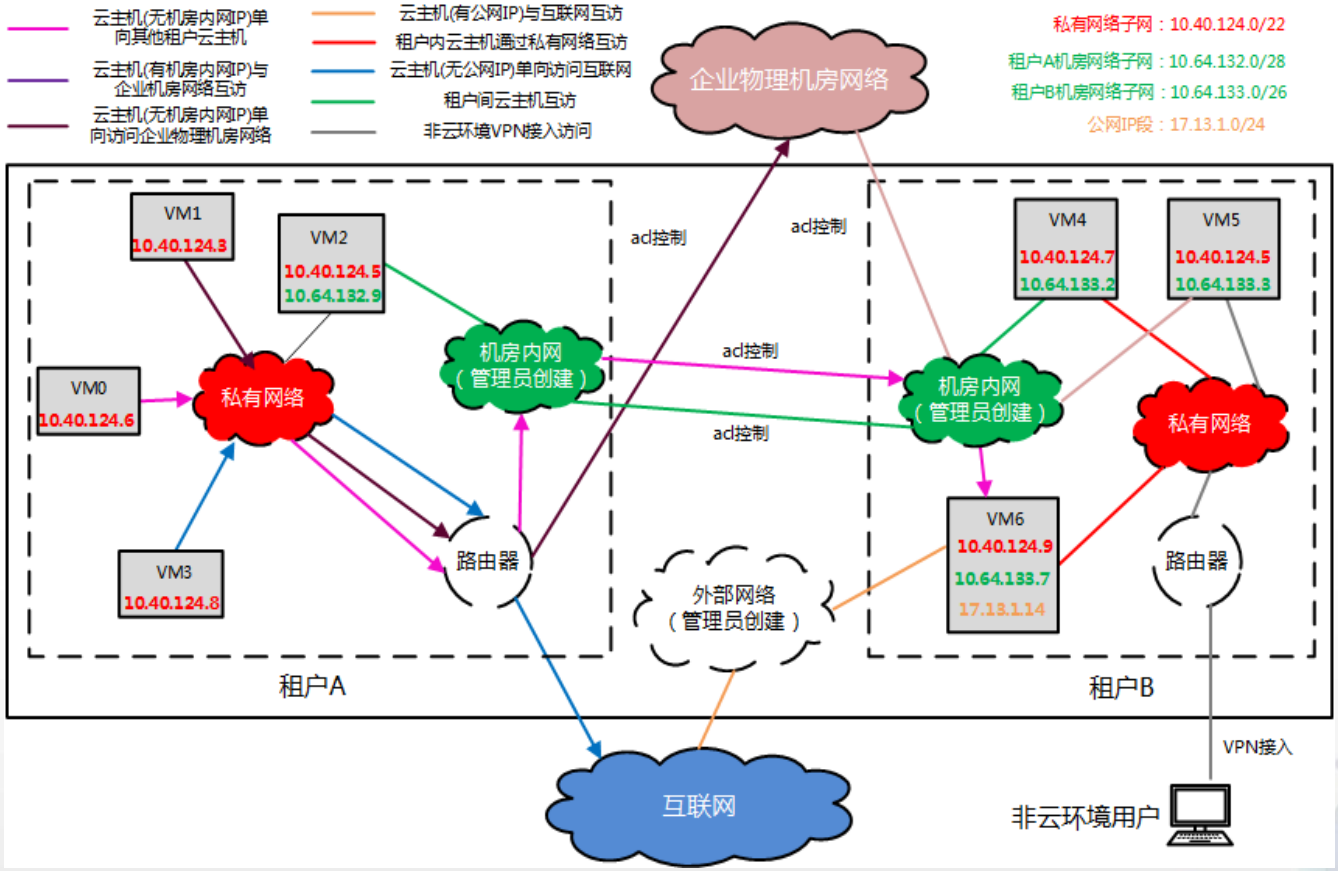
# 不足

- 所有云主机位于同个物理L2
  - 广播域扩大
  - 网络边界
  - IP资源管理
- 云主机和宿主机部分重叠，管理成本高
  - NAT、安全组等数千条iptables规则
  - 路由干扰
- 用户使用习惯上不同
- 技术上扩展性不够

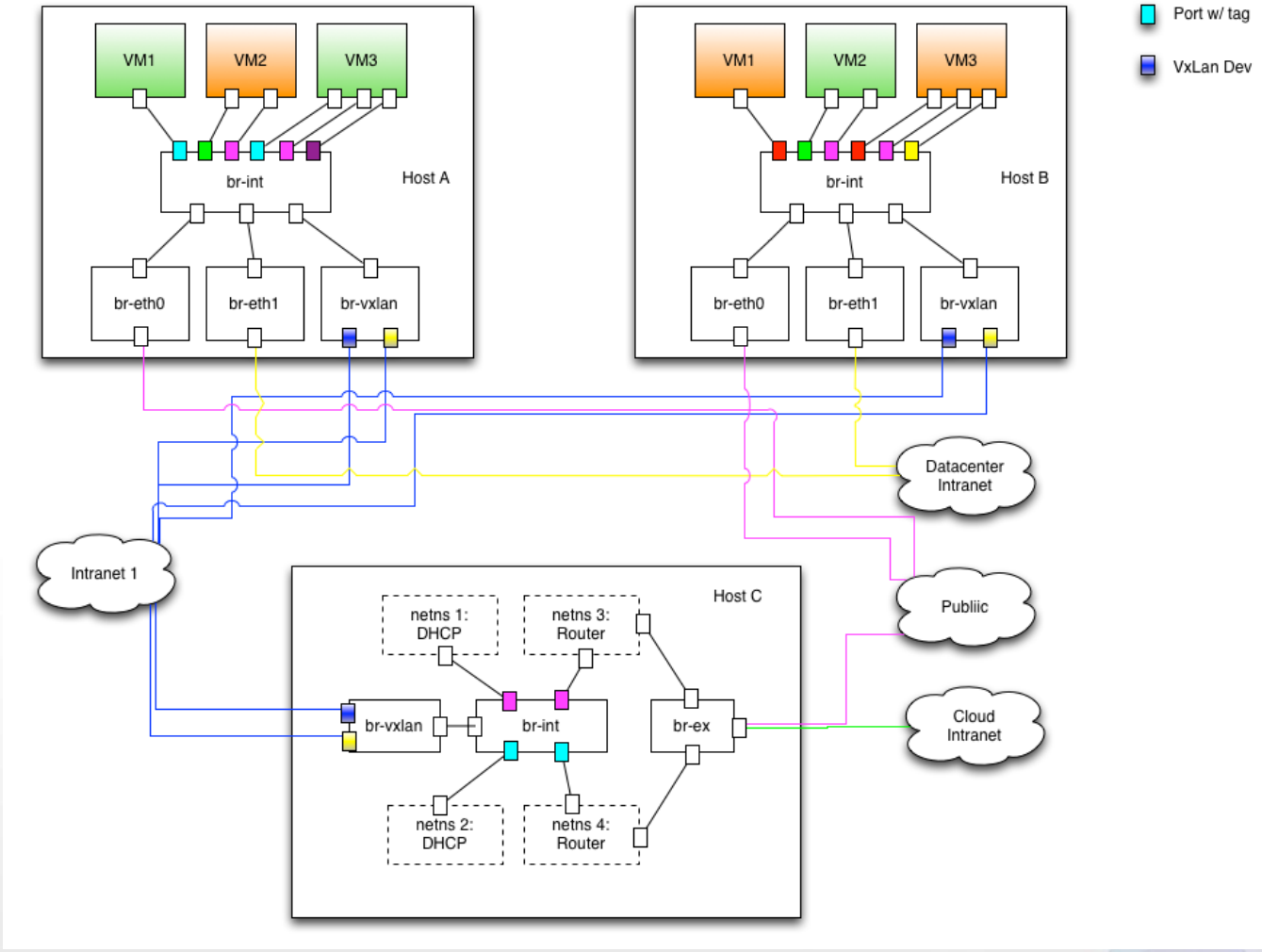
# 第二代云网络

- 基于neutron开发
- 设计上更加合理
  - 资源抽象更彻底
  - 可扩展性高，支持各种网络技术
- 灵活性好，易于实现不同的网络拓扑结构
- 根据业务需求定制
  - 私有网络
  - 机房内网
  - 公网

# 第二代云网络架构


云主机(无机房内网IP)单向其他租户云主机
云主机(有机房内网IP)与企业机房网络互访
云主机(无机房内网IP)单向访问企业物理机房网络
云主机(有公网IP)与互联网互访
租户内云主机通过私有网络互访
云主机(无公网IP)单向访问互联网
租户间云主机互访
非云环境VPN接入访问
私有网络子网：10.40.124.0/22
租户A机房网络子网：10.64.132.0/28
租户B机房网络子网：10.64.133.0/26
公网IP段：17.13.1.0/24
企业物理机房网络
VM1
10.40.124.3
VM2
10.40.124.5
10.64.132.9
VM0
10.40.124.6
VM3
10.40.124.8
私有网络
机房内网
（管理员创建）
路由器
acl控制
acl控制
acl控制
acl控制
VM4
10.40.124.7
10.64.133.2
VM5
10.40.124.5
10.64.133.3
机房内网
（管理员创建）
私有网络
VM6
10.40.124.9
10.64.133.7
17.13.1.14
路由器
外部网络
（管理员创建）
租户A
租户B
互联网
VPN接入
非云环境用户

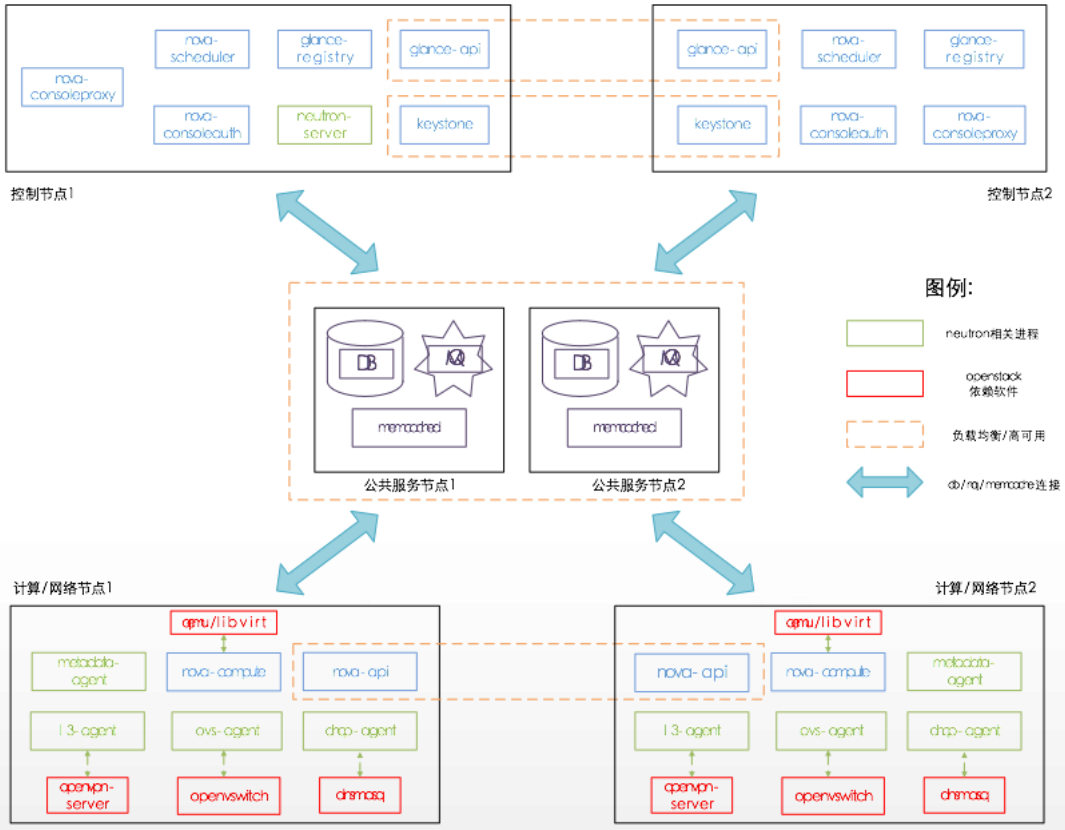
# 网络拓扑示意图


Port w/ tag
VxLan Dev
VM1
VM2
VM3
br-int
Host A
br-eth0
br-eth1
br-vxlan
VM1
VM2
VM3
br-int
Host B
br-eth0
br-eth1
br-vxlan
Datacenter Intranet
Intranet 1
Publiic
Host C
netns 1: DHCP
netns 3: Router
br-vxlan
br-int
br-ex
Cloud Intranet
netns 2: DHCP
netns 4: Router

# 第二代云网络(Cont'd)

- 技术选型
  - VXLAN (Virtual Extensible LAN)
  - Flat网络
  - ML2 Plugin + Open vSwitch Agent
  - L2 Population
- 自研功能
  - 新网络类型SVLAN (Shared VLAN)
  - 支持ACL
  - L3使用两次NAT
  - 网络监控
  - 租户OpenVPN接入
  - L3高可用

# 服务部署


nova-consoleproxy
nova-scheduler
glance-registry
glance-api
nova-consoleauth
neutron-server
keystone
glance-api
nova-scheduler
glance-registry
keystone
nova-consoleauth
nova-consoleproxy
控制节点1
控制节点2
DB
MQ
memcached
DB
MQ
memcached
公共服务节点1
公共服务节点2
图例:
neutron相关进程
openstack依赖软件
负载均衡/高可用
db/mq/memcache连接
计算/网络节点1
计算/网络节点2
qemu/libvirt
metadata-agent
nova-compute
nova-api
l3-agent
ovs-agent
dhcp-agent
openvpn-server
openvswitch
dnsmasq
qemu/libvirt
nova-api
nova-compute
metadata-agent
l3-agent
ovs-agent
dhcp-agent
openvpn-server
openvswitch
dnsmasq

# 问题及总结

- Neutron还在不断的改进中，建议使用新版本，并关注最新功能与bug修复
- 基于neutron的网络服务稳定性上仍有不足，需要完整、严格的测试
- L2 Population在实际使用中容易出问题且后果严重，功能上需要增强
- OVS Agent重启导致网络中断
- 3.9以前内核及低版本iproute容易导致namespace锁死
- 性能上相对原来有所下降，有进一步优化的余地
- 网络问题相对不容易定位和排查，可以定制一些自动化工具
- …

# 后续工作

- 继续提高稳定性
- 功能上的扩展
  - 安全组
  - QoS
  - L3的SNAT支持DVR
  - …
- 性能优化
  - 软件上的调优
  - 尝试一些硬件方案 (OpenFlow交换机等)
- 加强和社区协作

# 联系方式

- xuchengli@corp.netease.com


# Q&A

# THANKS

SequeMedia 盛拓传媒 | IT168 www.it168.com | ChinaUnix | ITPUB