ArchSummit全球架构师峰会北京站2015

百度开放云网络架构之路

杨毅 yangyi@baidu.com

Geekbang.
极客邦科技

整合全球最优质学习资源，帮助技术人和企业成长

Growing Technicians, Growing Companies

InfoQ

专注中高端技术人员的
技术媒体

EGO EXTRA GEEKS' ORGANIZATION NETWORKS

高端技术人员
学习型社交网络

StuQ

实践驱动的
IT职业学习和服务平台

GiT GEEKBANG INTERNATIONAL TRAINING 极客邦培训

一线专家驱动的
企业培训服务

地址：北京市朝阳区洛娃大厦C座6层1607室　　网址：www.geekbang.org

开始之前

≠

百度云（百度网盘）

http://bce.baidu.com
面向企业和开发者的云计算服务平台

http://yun.baidu.com
主要面向个人的云存储

计算与网络	存储和CDN	数据库	安全和管理	应用服务
云服务器 BCC	对象存储 BOS	关系型数据库 RDS	云安全 BSS	简单邮件服务 SES
负载均衡 BLB	云磁盘 CDS	简单缓存服务 SCS	云监控 BCM	简单消息服务 SMS
	内容分发网络 CDN	NoSQL数据库 MolaDB		应用性能管理服务 APM
				问卷调研服务
				移动App测试服务

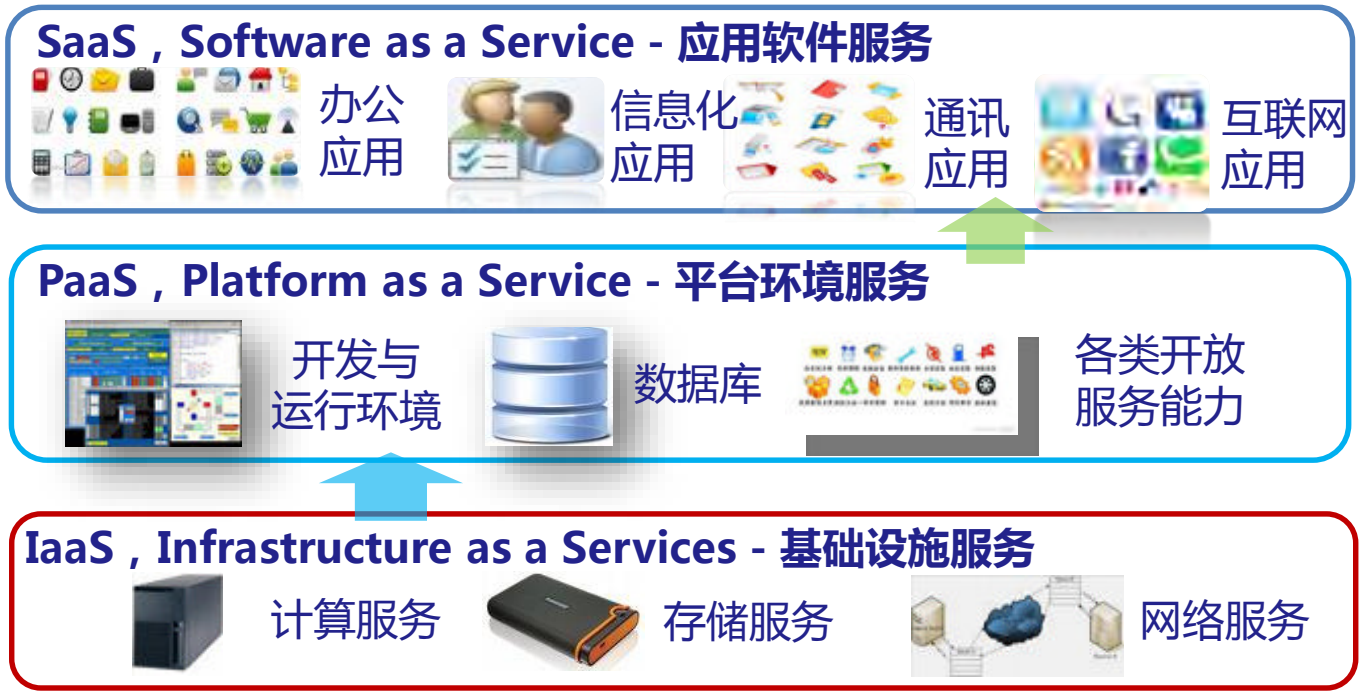
中间件服务	智能多媒体服务	数据分析	网站服务
应用引擎 BAE	音视频转码 MCT	百度MapReduce BMR	云虚拟主机 BCH
队列通知服务 QNS	音视频直播 LSS	百度机器学习 BML	
	人脸识别 BFR	百度OLAP引擎 Palo	
	文字识别 OCR	百度Elasticsearch	

大纲

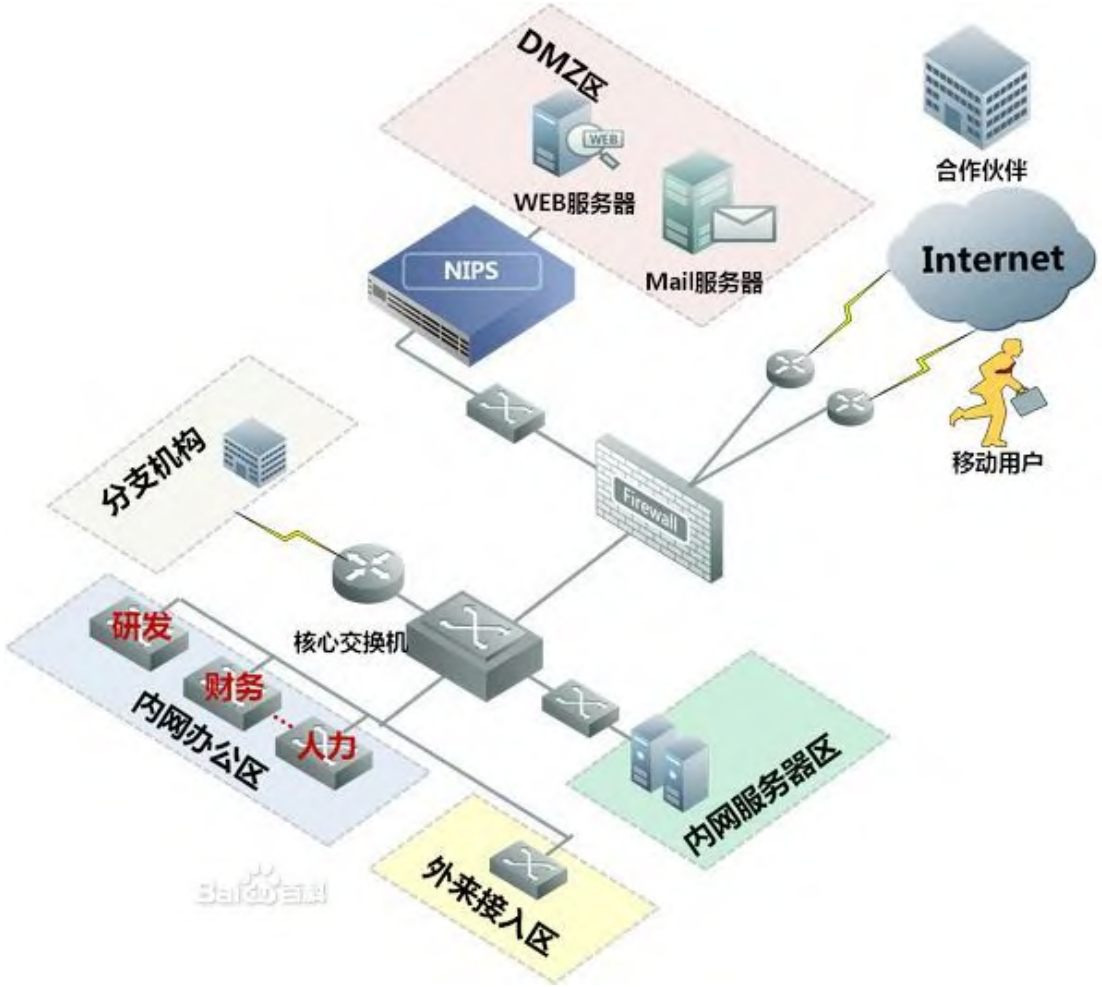
- 背景
- 需求
- 架构
- 发展

背景 – Network as a Service

- IaaS三大基础能力之一

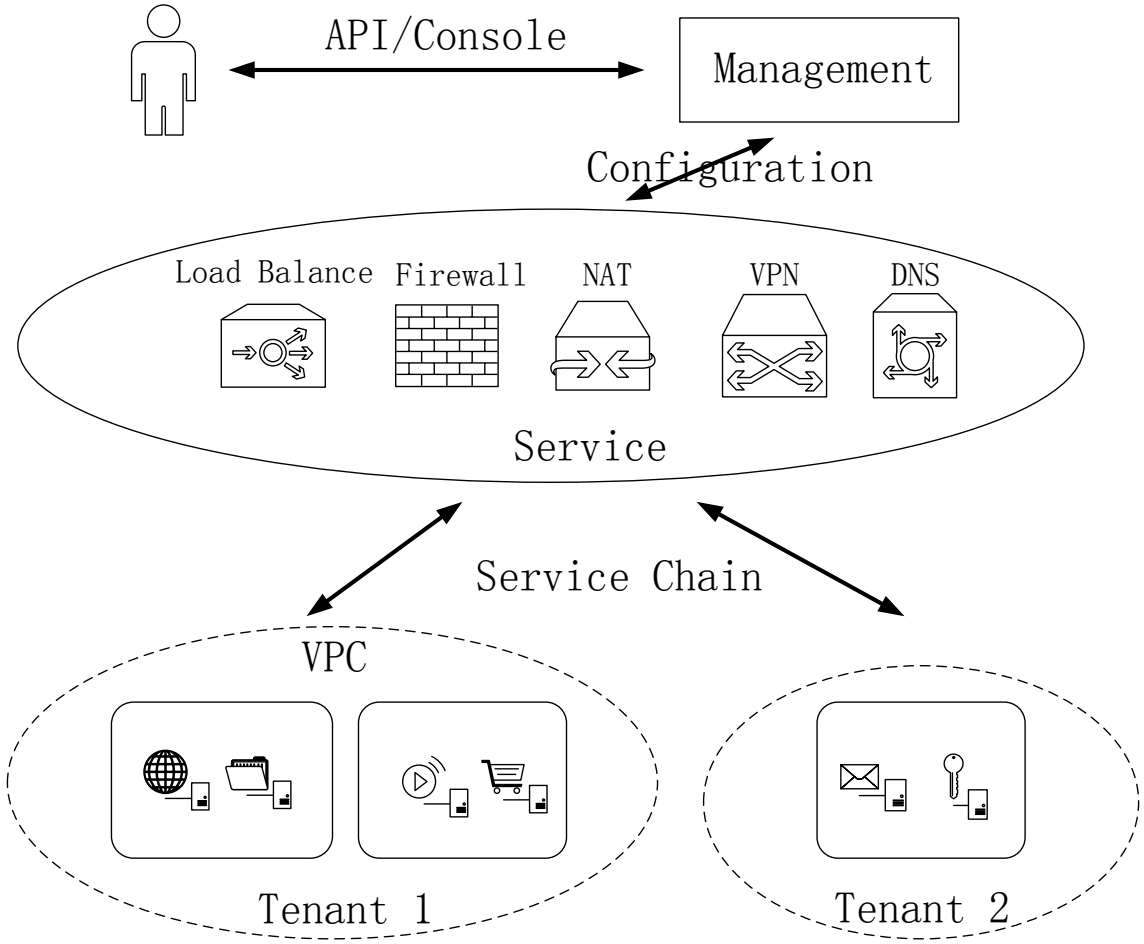


背景 – 传统物理网络



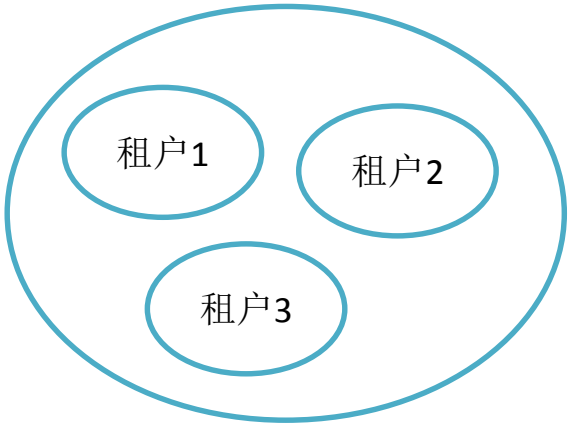
图片来源：http://baike.baidu.com/view/33936.htm

需求 – 多租户共享的云中网络

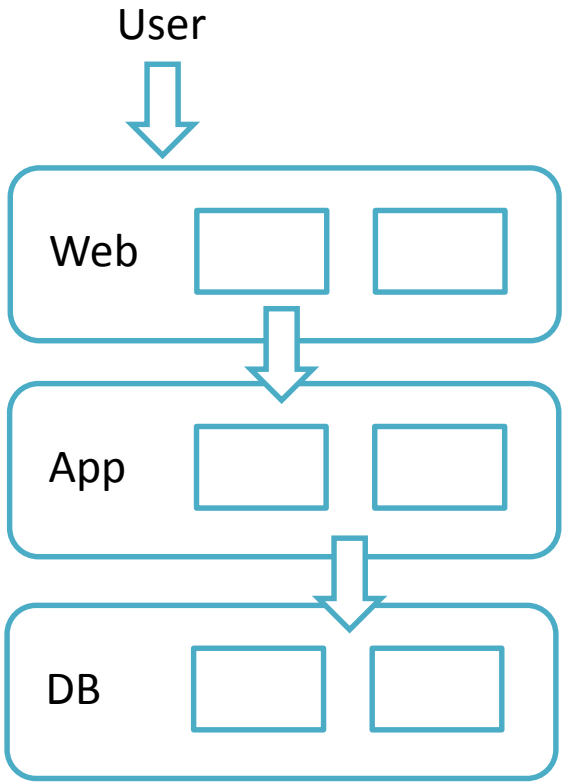


需求 – 用户自定义 Virtual Private Cloud

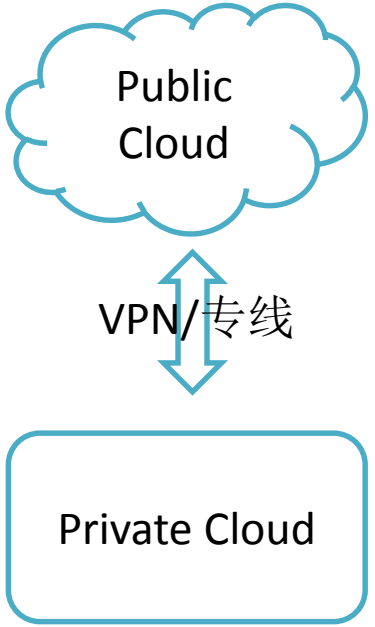
- 场景1：租户隔离



- 场景2：服务分层



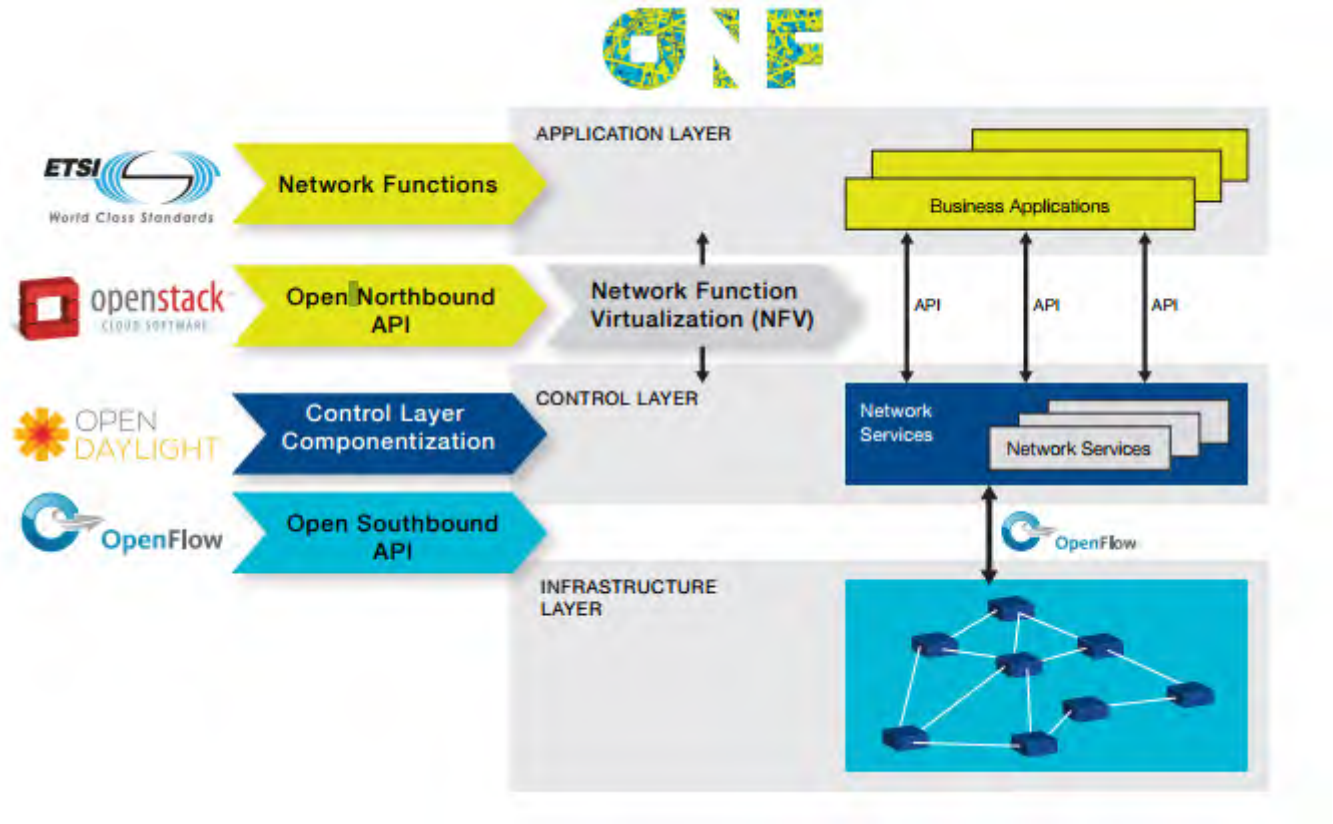
- 场景3：混合云



架构 – 传统物理网络无法满足需求

需求	云中网络	物理网络
配置变更	用户完成，100 q/s	网管完成，1次/天
资源池化	虚机自由迁移 服务自由组合	预先划分IP/路由/内网/外网/DMZ等
多租户	百万级别	vlan隔离，4K
弹性扩展	数据面+控制面	纯数据面，收敛比
稳定性	SLA：99.95%以上	

架构 – 理想美好，现实骨感的SDN

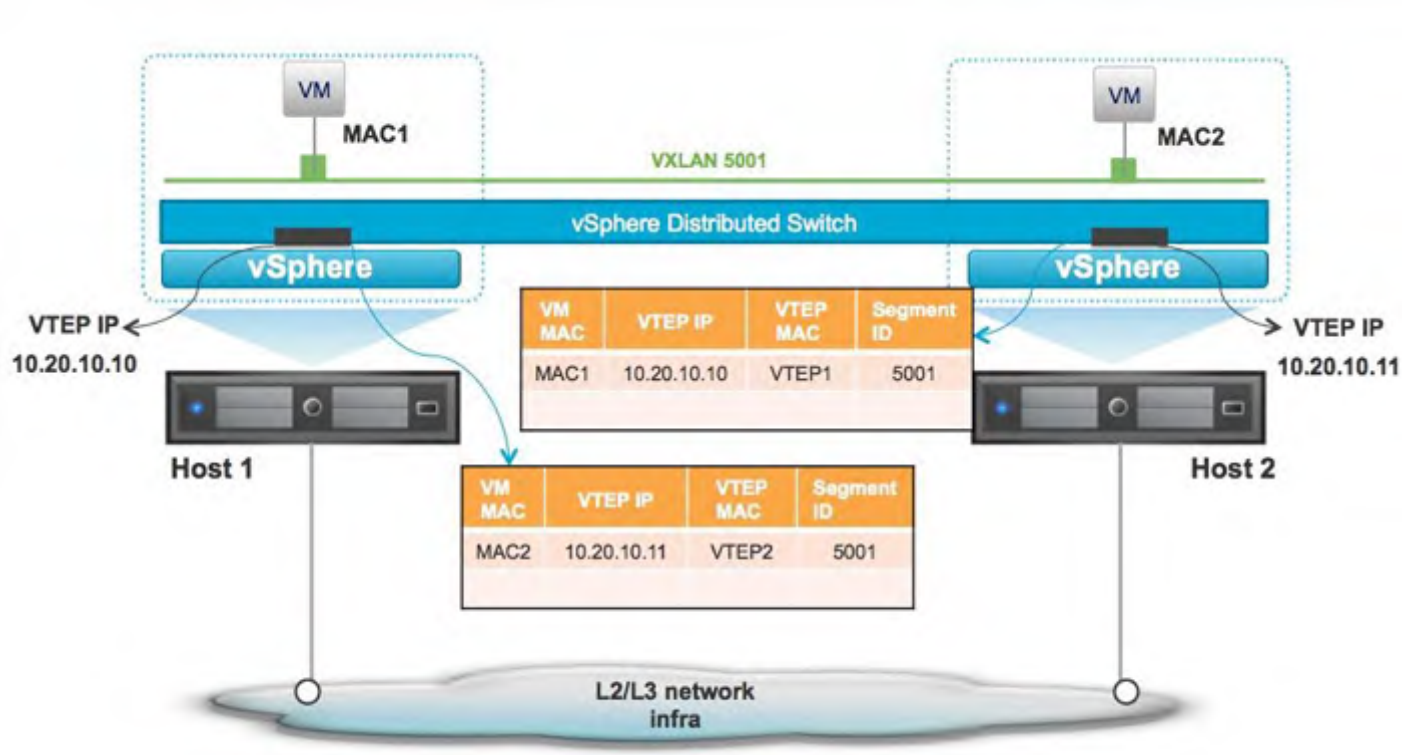


图片来源：https://www.opennetworking.org/images/stories/downloads/sdn-resources/solution-briefs/sb-sdn-nvf-solution.pdf

百度开放云网络架构选型

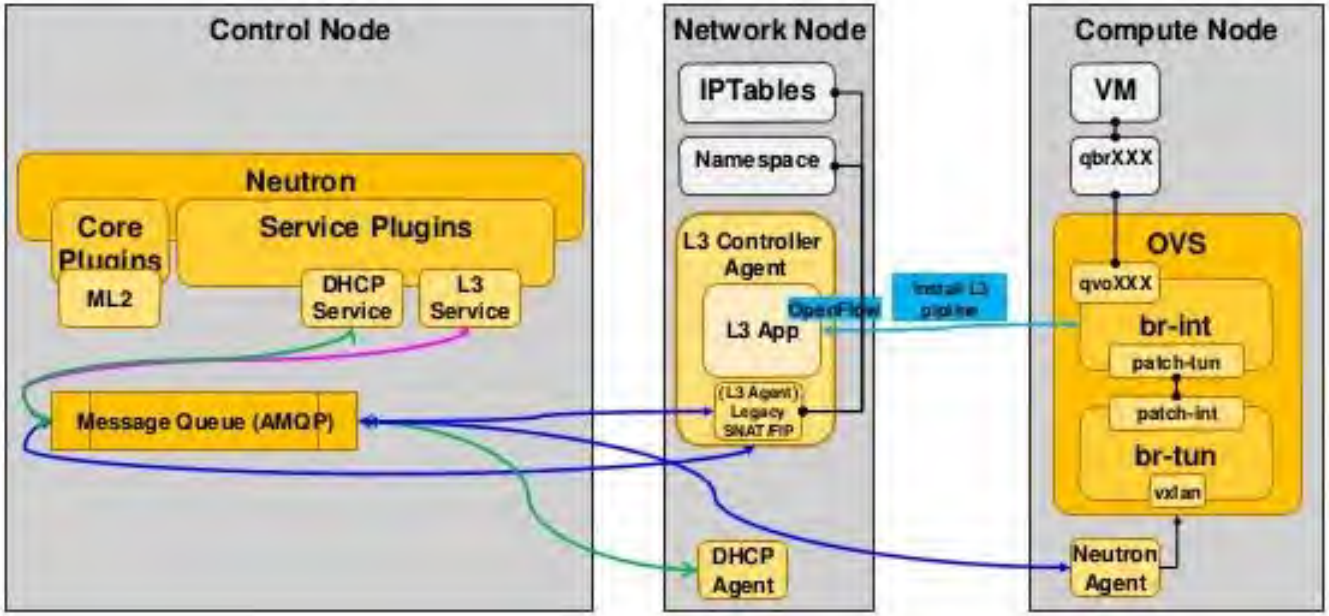
	技术方案	选型思考
网络管理	openstack neutron	标准、兼容
大二层	overlay / vxlan	可扩展、物理无关
虚拟接入	openvswitch / openflow	灵活、可控
物理融合	独立内网核心 统一NAT网关	安全、复用

原理 VxLAN



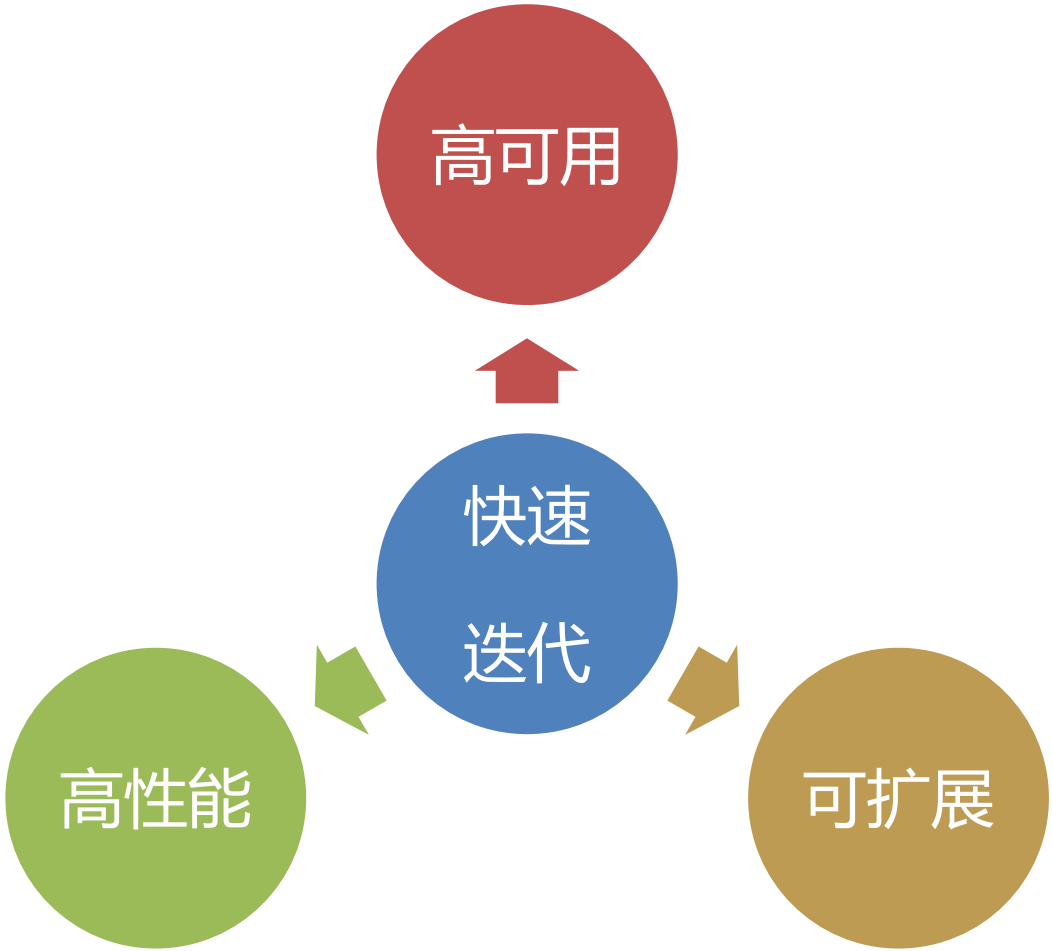
图片来源：https://blogs.vmware.com/vsphere/2013/07/vxlan-series-how-vmotion-impacts-the-forwarding-table-part-6.html

原理 openstack neutron

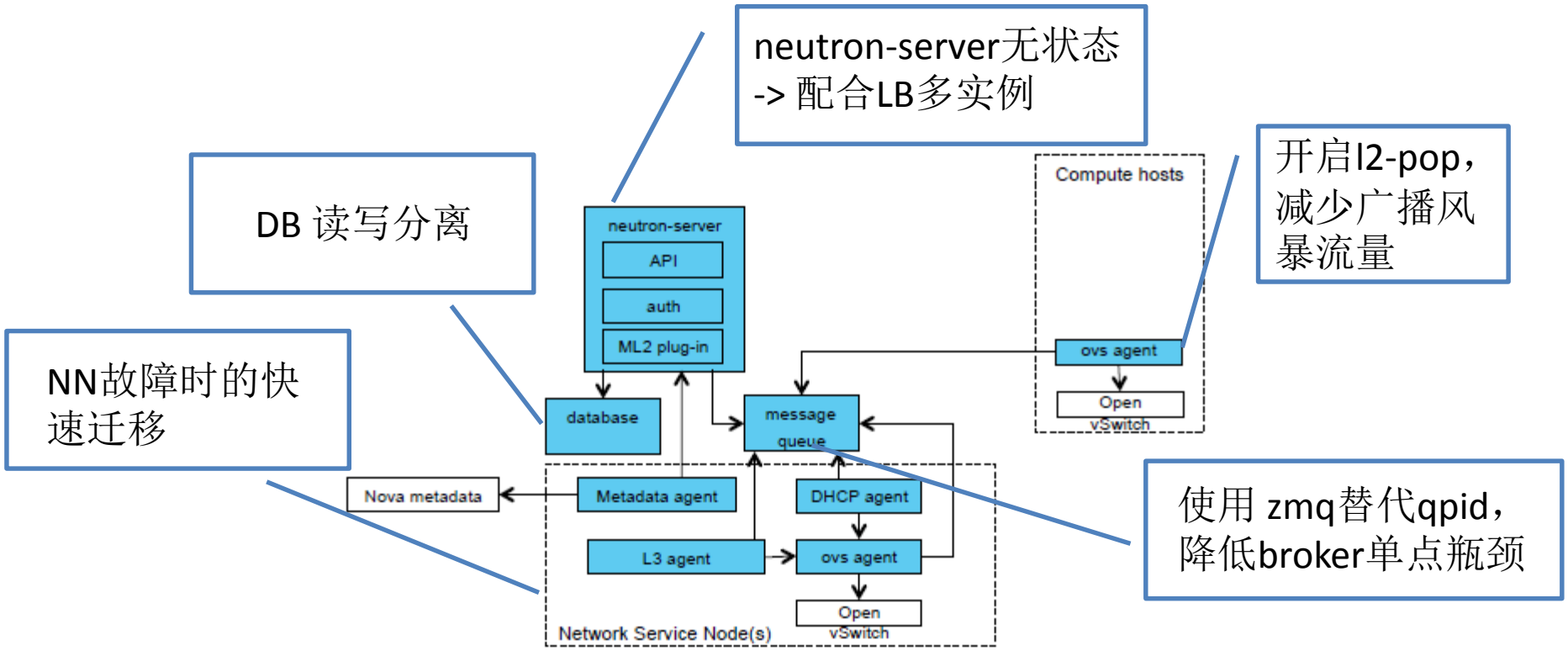


图片来源：http://blog.gampel.net/2015/01/dragonflow-sdn-based-distributed.html

工程实践

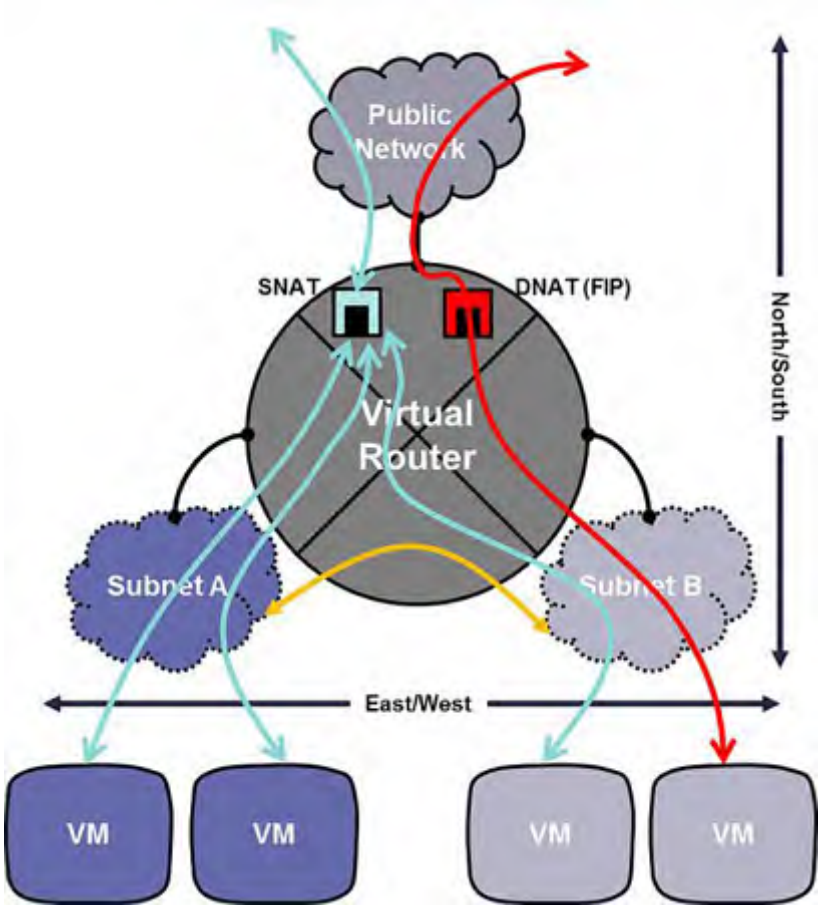


架构之路 – 社区蛮荒时代


neutron-server无状态
-> 配合LB多实例
DB 读写分离
开启l2-pop，
减少广播风
暴流量
NN故障时的快
速迁移
使用 zmq替代qpid，
降低broker单点瓶颈
neutron-server
API
auth
ML2 plug-in
database
message
queue
Compute hosts
ovs agent
Open
vSwitch
Nova metadata
Metadata agent
DHCP agent
L3 agent
ovs agent
Open
vSwitch
Network Service Node(s)

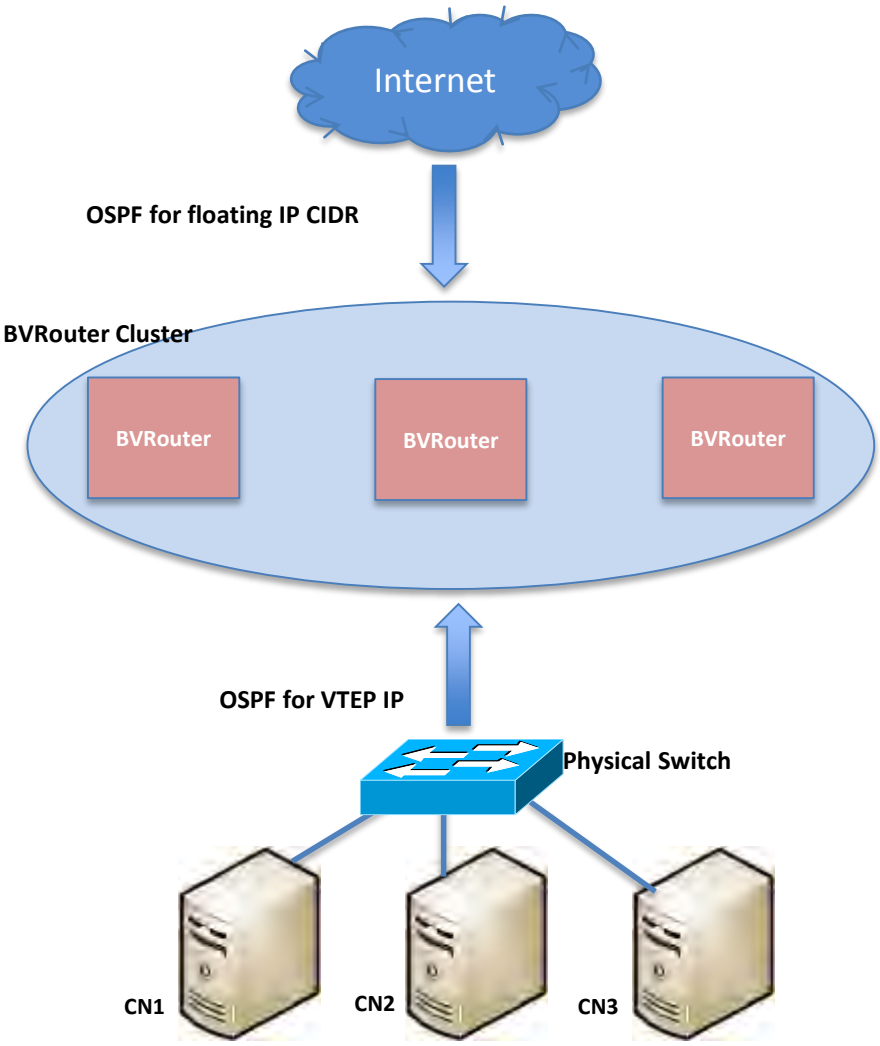
架构 – vRouter的瓶颈

- 单点问题
- 性能瓶颈
- 可扩展性

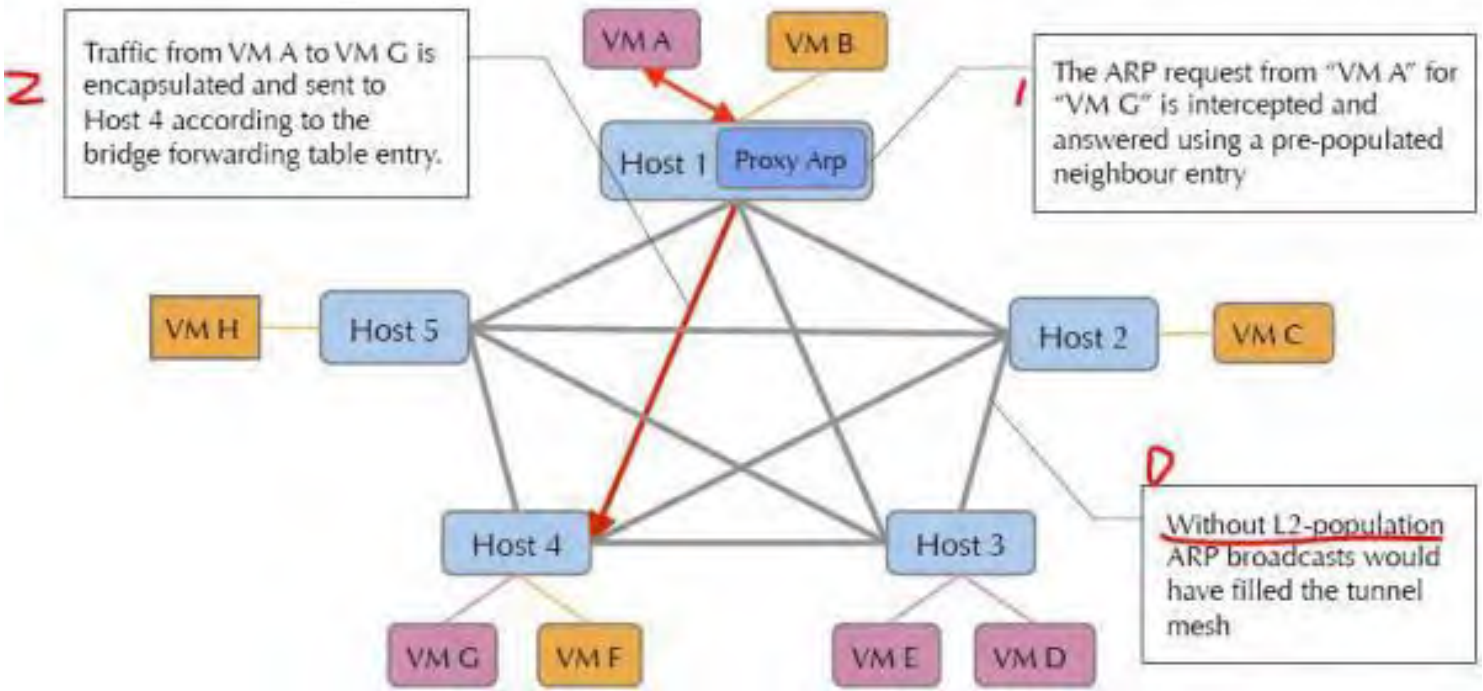


架构之路 – 自力更生时代

- 快速迭代：热升级
  - 断网时间：秒级
- 高可用：dpdk + ECMP实现的BvRouter集群
  - vrouter转发性能：5*N倍
  - NN单机故障损失：分钟级 -> 秒级
  - 单租户规模：100 -> 5000
- 高性能：流表优化
  - ARP代答：去除广播流量
  - 分布式路由：东西流量不过NN



L2 population和ARP代答



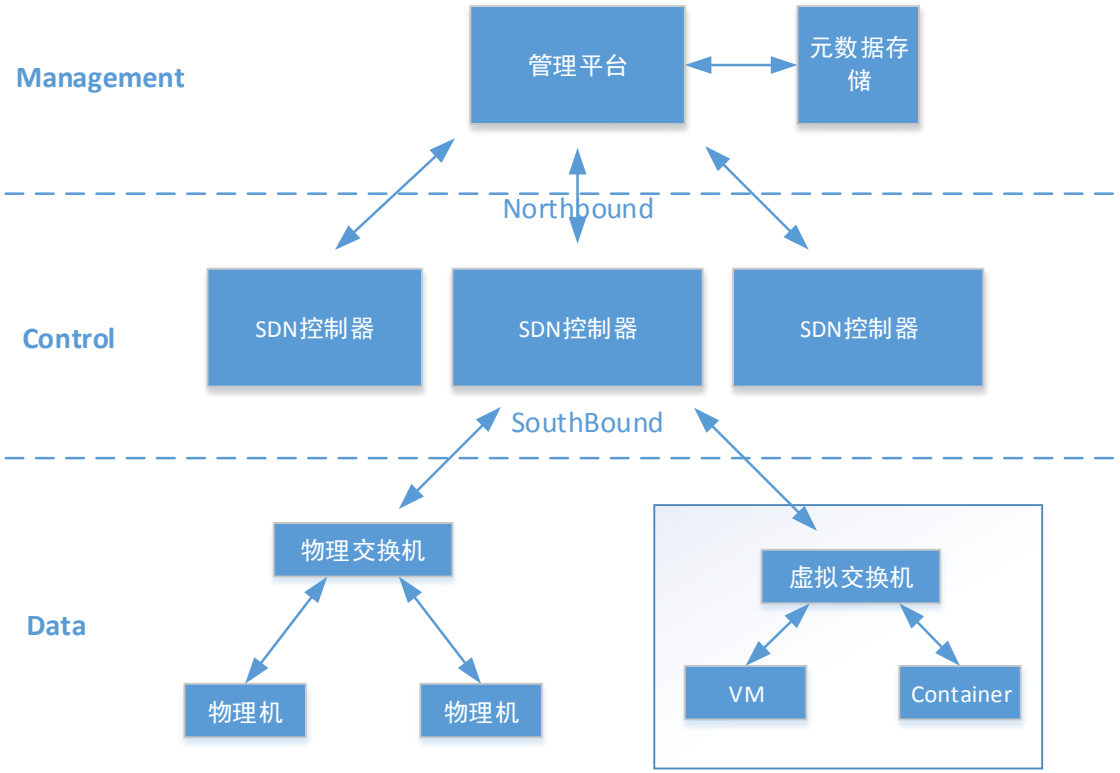
图片来源：http://www.xlgps.com/article/10555.html

可扩展：Neutron不能Scale的关键

- 控制平面
  - 集中而混乱的元数据管理
  - 预配置：单个port up可能需要修改所有计算节点的配置
  - agent重启时的全量sync
- 数据平面
  - 数据通路上的“瓶颈”
  - 单个network消耗的资源：namespace，dnsmasq进程…
  - 软件实现的性能损耗

架构之路 – 软硬结合的分布式时代

- 元数据分层及Sharding
  - 抽象的逻辑元数据
    (loigical network/subnet/ports)
  - 具体的物理位置数据
    (location of ports, tunnel ...)
- 支持差分及按需同步的通信方式
  - 北向RPC
  - 南向openflow
- 混合overlay
  - 物理交换机的容量限制
  - 容器带来的新挑战



总结

- 云计算网络需求复杂
- 传统物理网络技术无法应对
- 新兴SDN技术还不够成熟
- 架构需要逐步演进

Thanks!