

WOTA

51CTO

World Of Tech 2017

全球架构与运维技术峰会

2017年4月14日-15日 北京富力万丽酒店

ARCHITECTURE



出品人及主持人：

张立刚

1号店技术部
电商云平台技术总监

云服务架构

YY 游戏私有云建设历程

虎牙信息
刘亚丹



刘亚丹

虎牙直播

基础运维负责人

分享主题：

YY游戏私有云建设历程

个人简介

虎牙信息-基础运维负责人

9 年互联网运维经历

5 年私有云 IaaS 和 PaaS 实践

参与编写《自主实现SDN虚拟网络和企业私有云》



大纲

1

- 需求背景

2

- YY私有云1.0现状

3

- 2.0平台方案选型

4

- 网络，存储，计算集成实践方案

需求背景

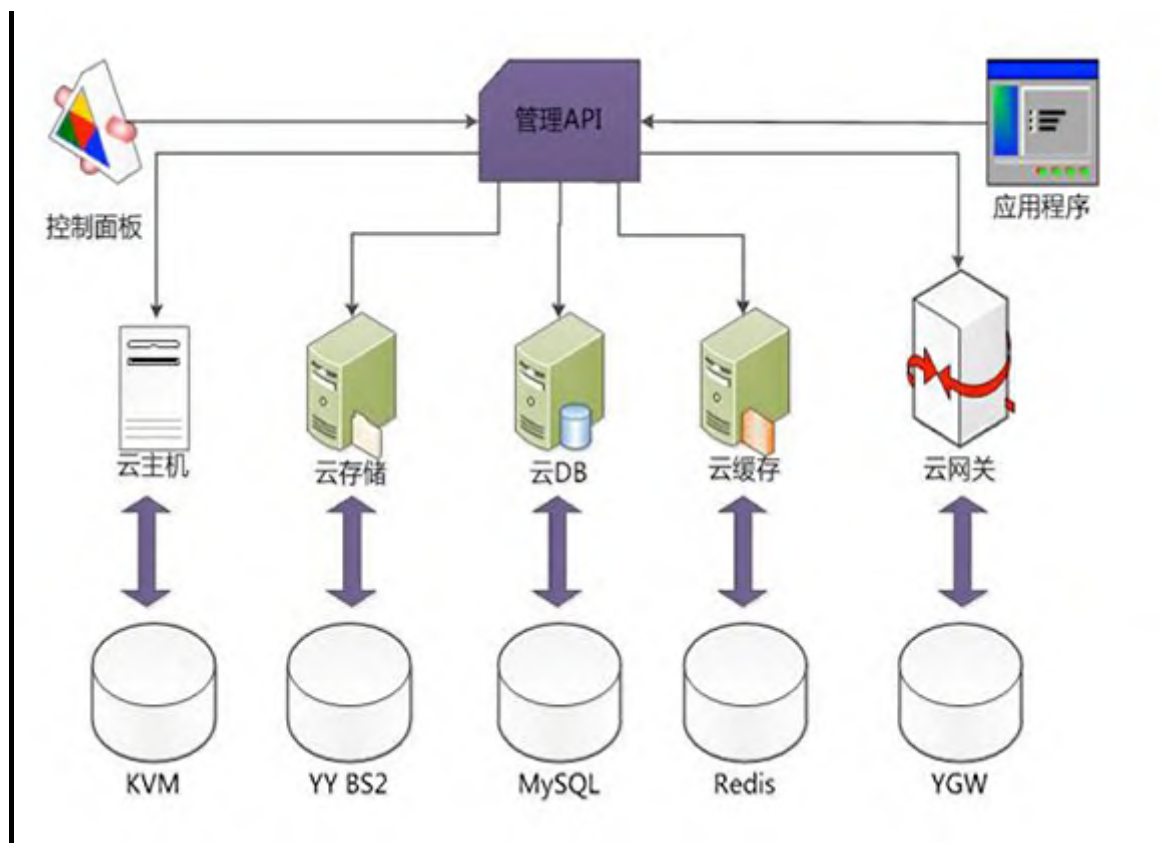
页游和 Web 类应用适合虚拟化

YY游戏运营平台：联运和独代游戏，具备一定的规模经济性

页游特点：开服多，周期短，云化提高资源利用率和回收效率

Web类应用特点：无状态，弹性伸缩，故障隔离

建设历程-YY 私有云1.0平台现状



- 1.0 核心功能
 - 云主机-本地存储
 - 云网关 (LVS+NAT)
 - 云DB, 云缓存
 - 云存储 (对象存储)
 - 云监控
- Cloud1.0的不足
 - 租户网络未隔离
 - 可控性差

建设历程-基于OpenStack 的1.0平台网络架构

Privoder Network

- flat+vlan 模式

典型网络配置支持

- 公网模式：VM 双网卡，eth0绑定电信、联通(Flat)；eth1绑定内网 IP（flat+vlan）
- 云网关模式：eth0绑定内网 IP（flat+vlan）

建设历程-OpenStack 上踩过的坑

稳定性



架构组件复杂

代码质量不高

可维护性



安装部署复杂对运维要求极高

无法平滑升级

扩展性



L3-Agent性能瓶颈

Dnsmasq & Keystone MQ 等性能瓶颈



建设历程-OpenStack上踩过的坑

- 从全局来看，OpenStack存在的问题包括：
 - 项目庞大复杂，过多的模块、扩展、功能让这个产品很难部署使用，出错时也难以调试。
 - 发布质量一般。目前是按时间周期发布，每个版本的维护时间过短，且旧版本不支持无缝升级到新版本。整个项目不太注重生产环境的可用性。
 - 项目的代码量仍然在持续增长，但是不注重代码质量和功能完整性。很多部分实现的功能成为遗留代码。
 - 可控性差，对运维的素质要求非常高。
 - 如下是我们在生产环境中面临的具体问题：
 - 由于Neutron租户网络存在l3-agent瓶颈，无法实现生产环境可用的租户隔离网络。
 - Metadata服务存在明显的性能问题。Metadata服务网络结构复杂，容易导致VM创建或者重启失败。
 - 直到IceHouse版本，仍然不能在线无缝升级。升级不方便加上旧版本的维护时间短，不适合生产环境大规模使用。
 - 授权体系一直是软肋。默认的policy.json不安全，也不支持全局动态配置权限。
- 自身的Bug影响稳定性，上游社区不重视。已知的Bug包括：
 - Keystone性能Bug。
 - Neutron定时任务执行稍有异常就清空OVS配置。
 - Neutron DHCP 服务器（Dnsmasq）性能问题。
 - 在某些情况下，动态迁移（Live Migration）不支持按Region迁移。
 - 动态迁移的实现没有考虑超时和网络不通时如何进行恢复。

建设历程-YY 私有云2.0-选型思路

➤ 关键需求:

- 提供高性能, 可扩展的 VPC 网络
- 云管理平台自主可控

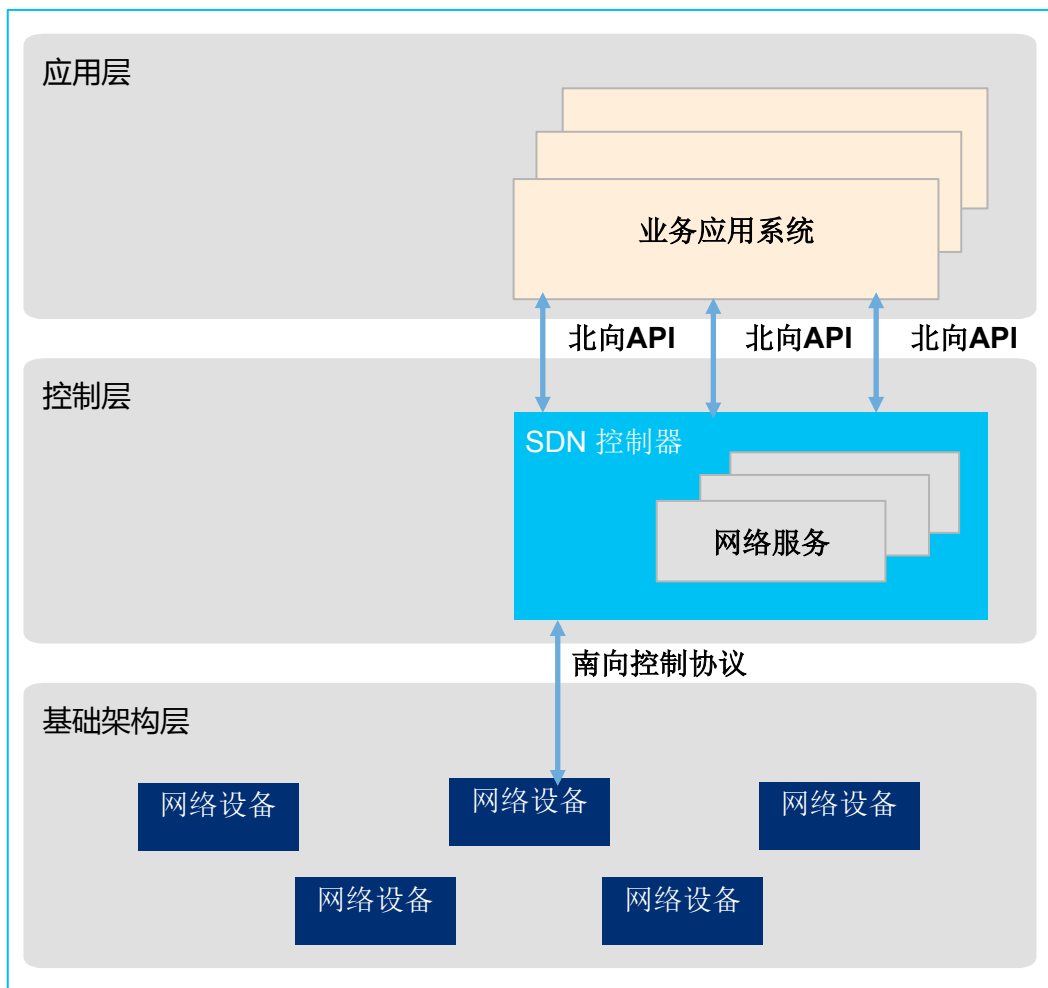
➤ 基本策略: 以开源软件为基础, 自主研发为核心, 专业厂商为关键补充

- 虚拟计算: KVM
- 虚拟存储: Ceph
- 虚拟网络: 借助专业厂商的专门网络硬件和控制器, 实现 VXLAN 网络

建设历程-YY 私有云2.0-方案选型

- 虚拟网络的灵魂：SDN 控制器
 - 工作机制复杂，开发工作量巨大
- 基于硬件的 VXLAN 和集中式组合网关
 - 实现 VPC, NAT, floatingIP, 子网, 路由等功能
- 云管理平台
 - 实现对计算, 存储, 网络, 以及云数据库等模块的调用, 调度, 监控和反馈

建设历程-YY 私有云2.0-方案选型



➤ 虚拟网络的灵魂：SDN 控制器

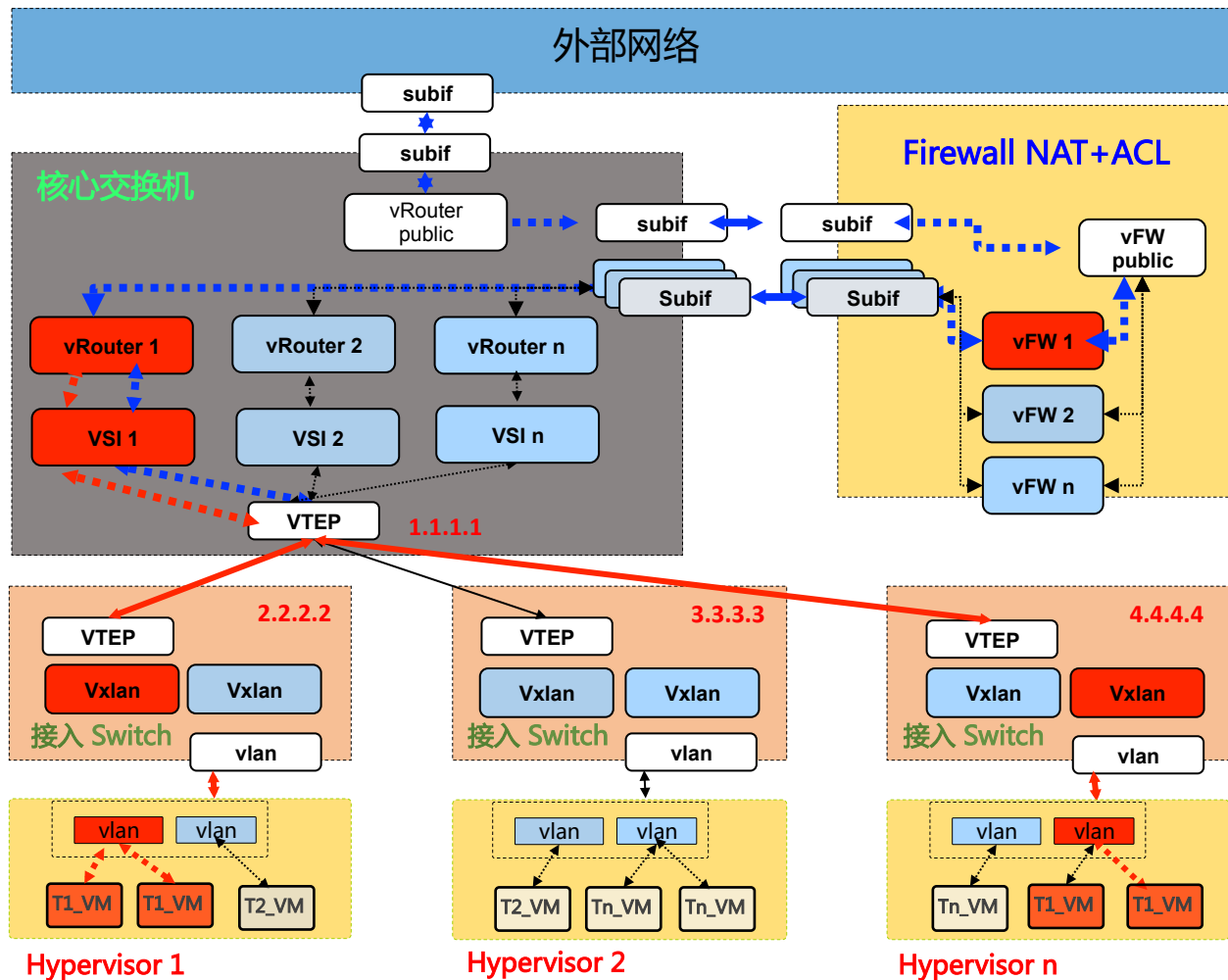
➤ 基本结构

➤ 工作机制

➤ 选型方案：

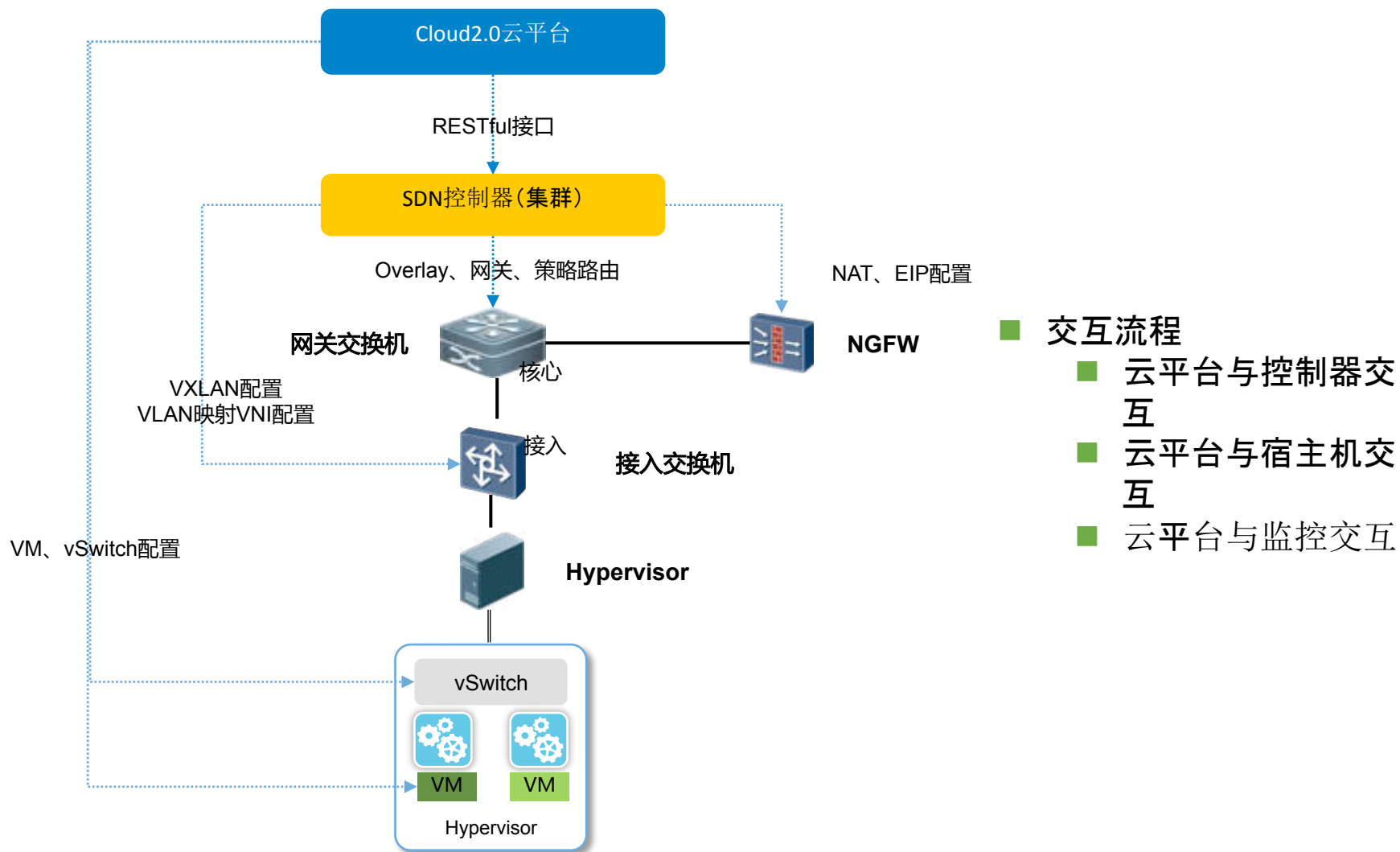
- 简单标准的北向接口
- 复杂的南向接口由厂商实现

建设历程-2.0 Overlay 网络模型

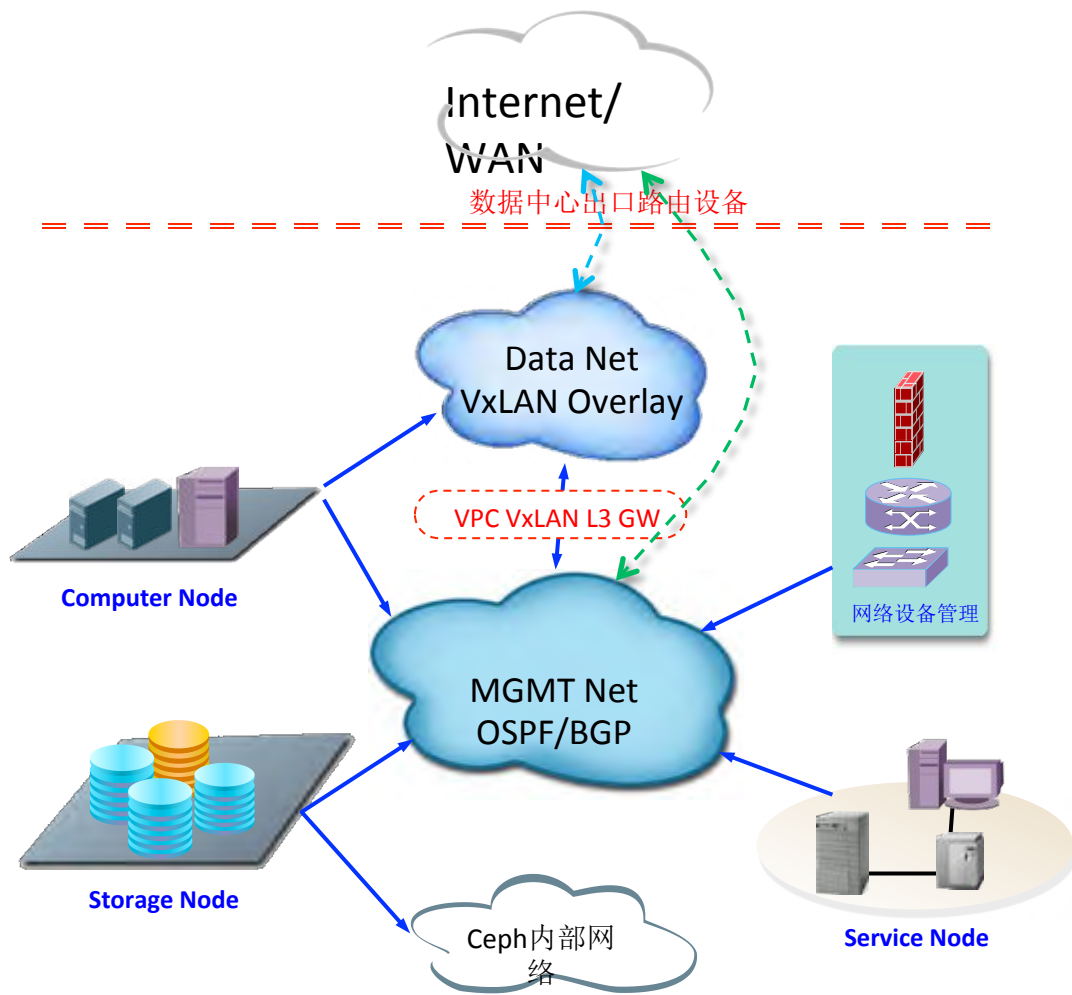


- 基于硬件的 VXLAN 和集中式组合网关
- 东西向流量转发
- 南北向流量转发

建设历程-云平台各层交互流程



建设历程-虚拟网络模型-整体模型

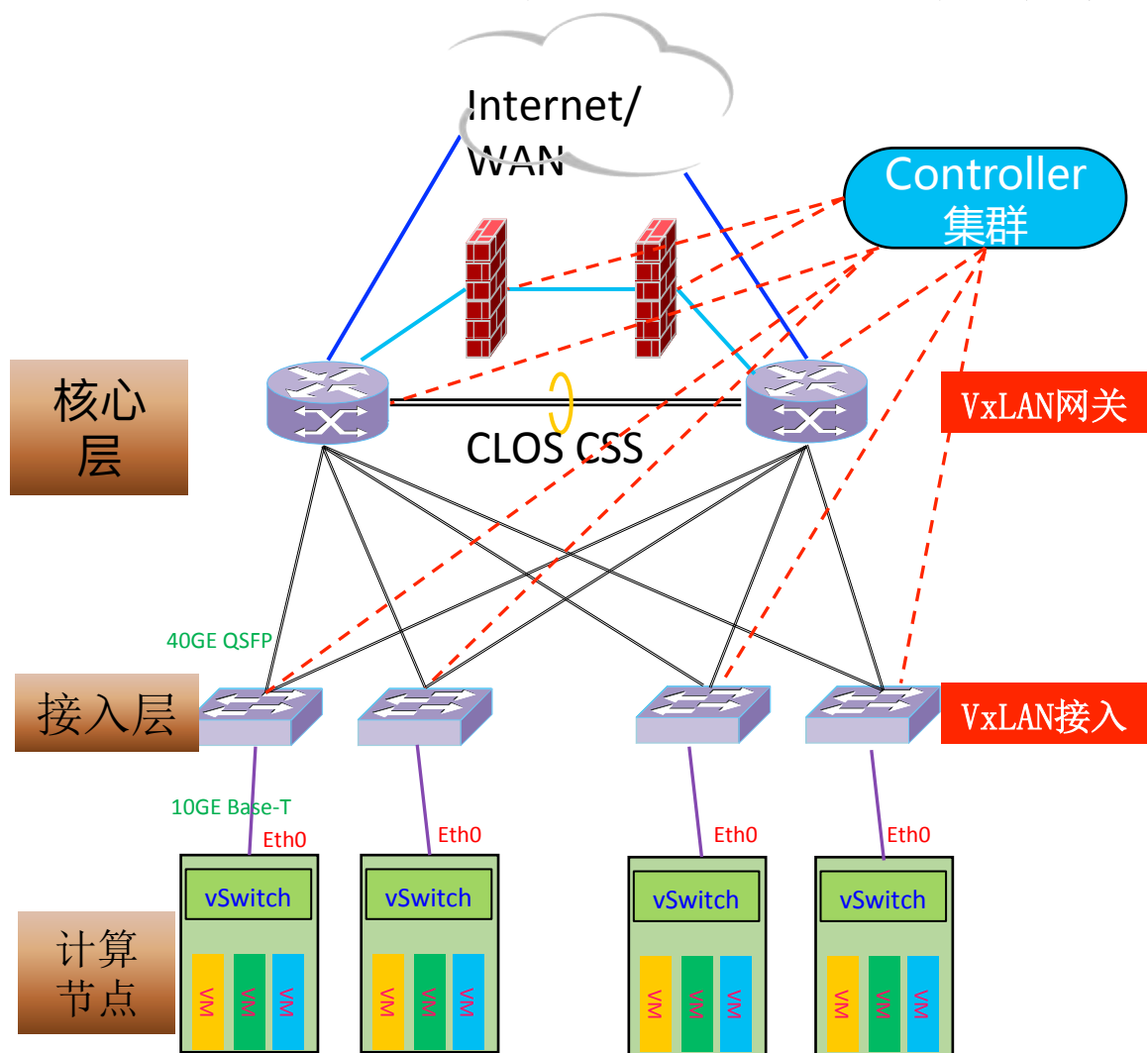


模型描述:

➤ 三个网络

- Data net: Tenant VPC, 利用VxLAN虚拟化技术构建叠加网络
- MGMT net: 组件内部通信、存储访问
- Ceph内部网络

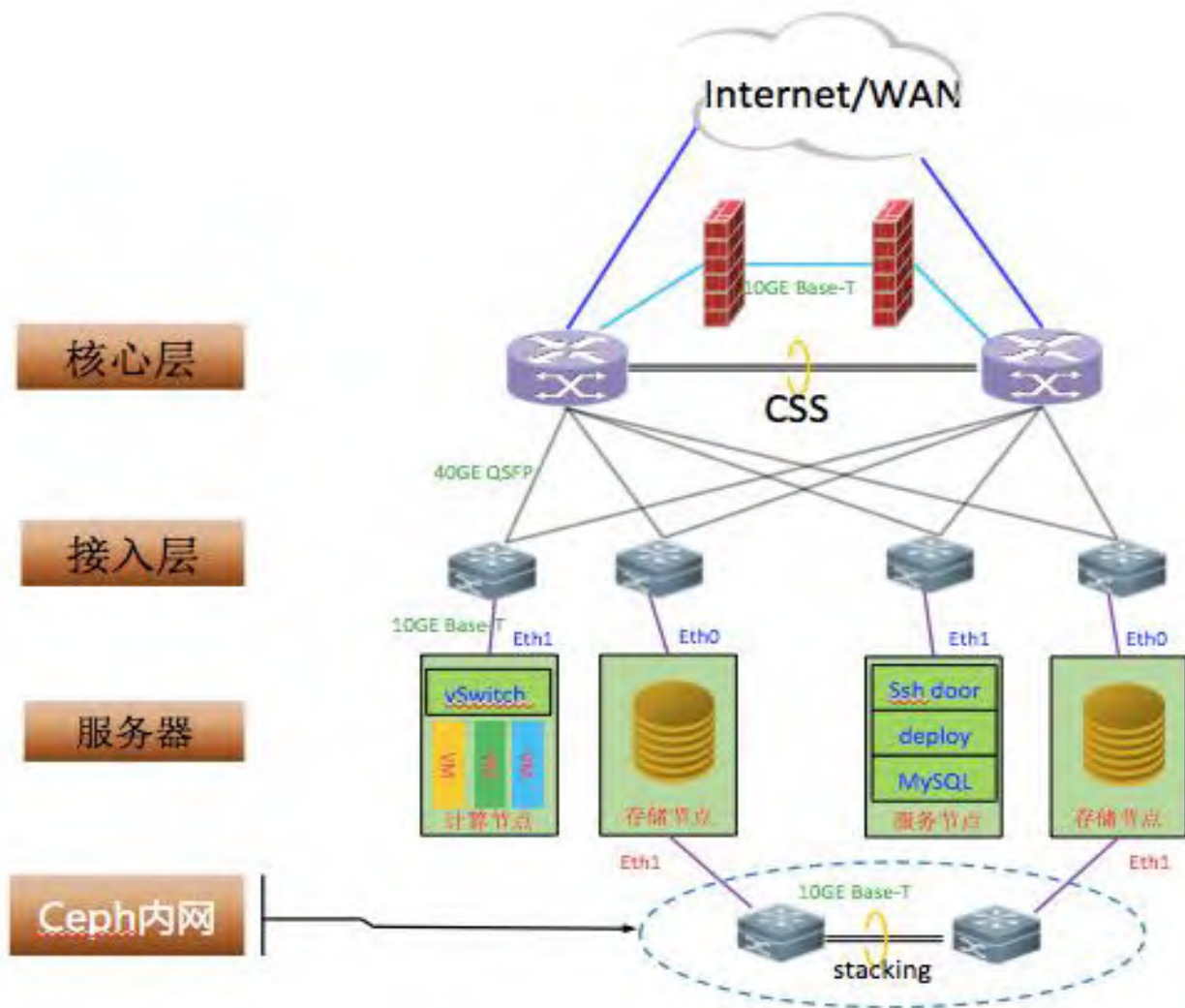
建设历程-网络架构-Data Net



数据网络-架构描述

- SDN Hybrid Switch 构
- Controller 集群集中控制
- 采用核心层-接入层结构
 - 核心 Switch: 采用 CLOS 架构核心交换机, 多级多平面无阻塞交换架构, 双机高可用热部署
 - 接入层: 40GE 双归接入
 - 防火墙: 旁挂核心 Switch, 采用主-被部署

建设历程-网络架构-MGMT Net



管理网络-架构描述:

- 通用L3 Switch构建
- 采用核心层-接入层结构
 - 10GE接入
 - 40GE汇聚
- 与Data net共用核心交换机，核心交换机用作数据中心出口设备
- 接入层 Switch使用40GE双归接入核心 Switch

建设历程-网络设备选型

➤ SDN Hybrid Switch:

- 国内厂家: Huawei, H3C, 锐捷, 中兴, 盛科
- 国外厂家: Cisco, Arista

➤ SDN Controller:

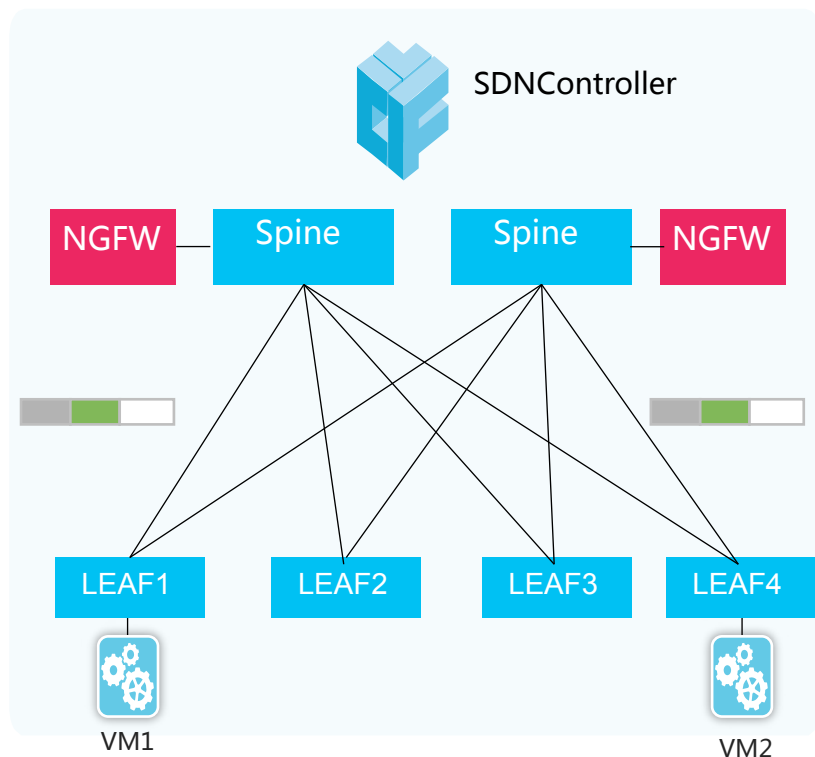
- H3C: VCF Controller
- Huawei: Agile Controller
- 云杉: NSP Controller

➤ NGFW:

- H3C: M9000
- Huawei: USG9500
- Hillstone: X7180

➤ 设备组合:

- H3C: VCFC+M9006+S12508AX-F+S6800
- Huawei: AC+USG9520+CE12808+CE6850
- 云杉+Huawei: NSP+USG9520+CE12808+CE6850

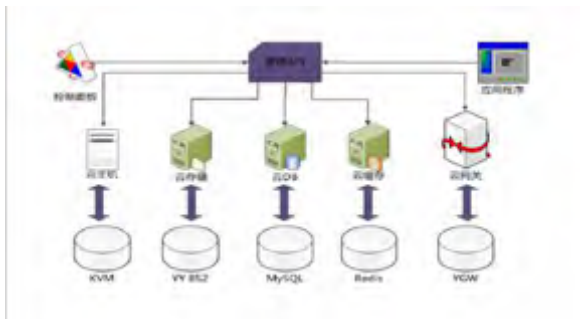


演进路线:

Cloud 1.0

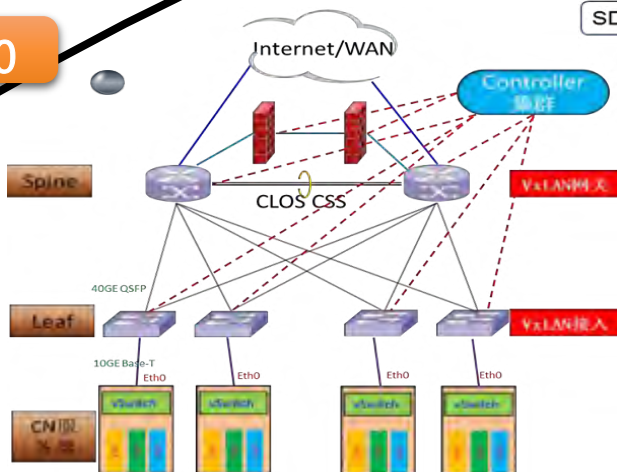
Cloud 2.0

Cloud 2.0 Ext



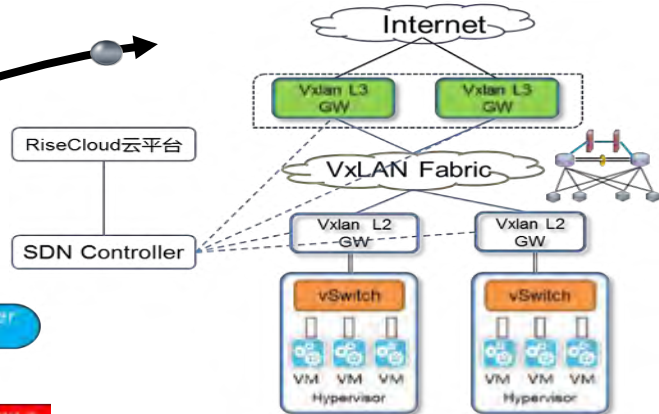
RiseCloud 1.0

- 基于Opensack Provider Network混合网络
- 千兆以太基础网络，VLAN隔离方式



RiseCloud 2.0

- 基于纯硬件VxLAN Fabric的Overlay网络方案
- 集中式硬件L3 VxLAN网关方式
- 万兆以太基础网络，无阻塞Spine-Leaf结构



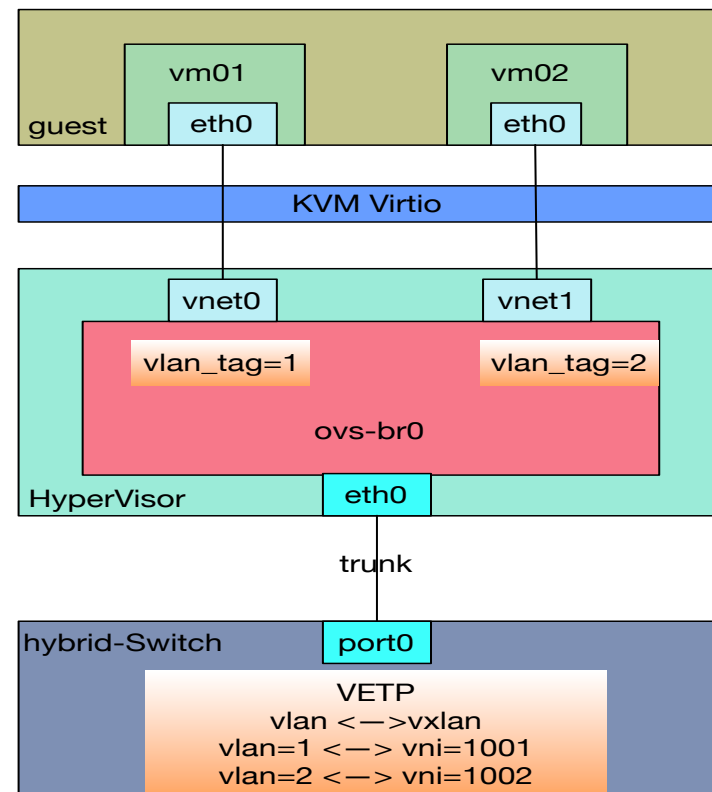
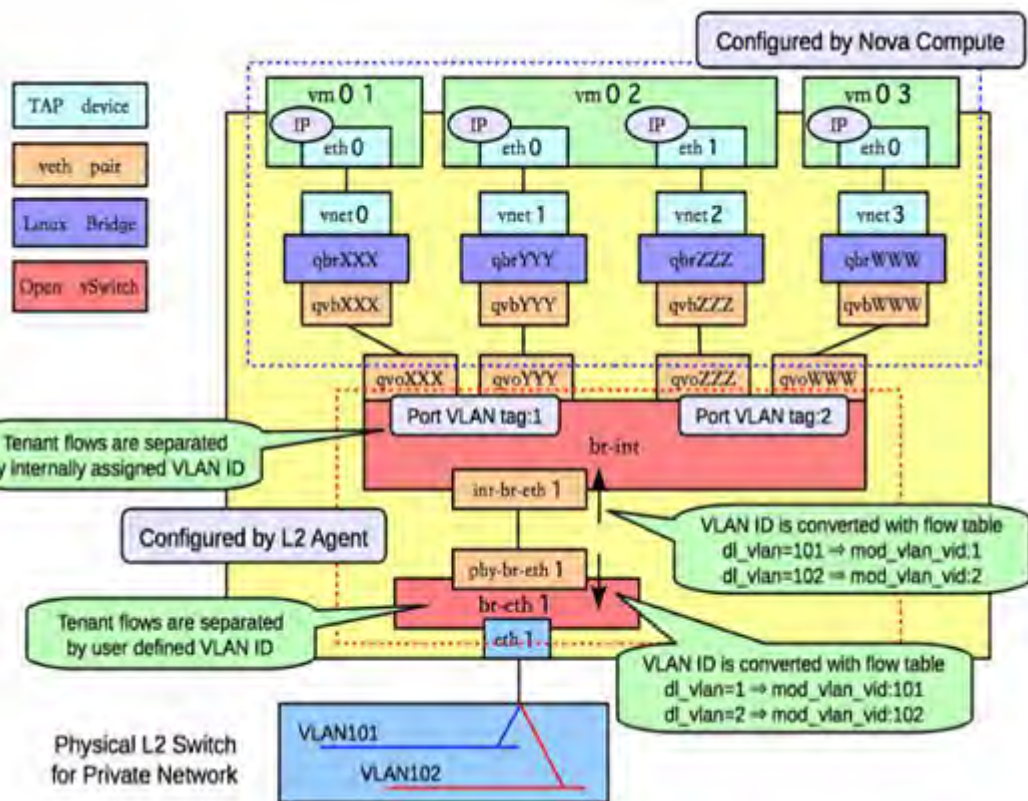
RiseCloud 2.0 Ext.

- 基于混合VxLAN Fabric的Overlay网络方案
- 采用VGW服务器集群作为L3 VxLAN网关
- 真正具有弹性伸缩能力

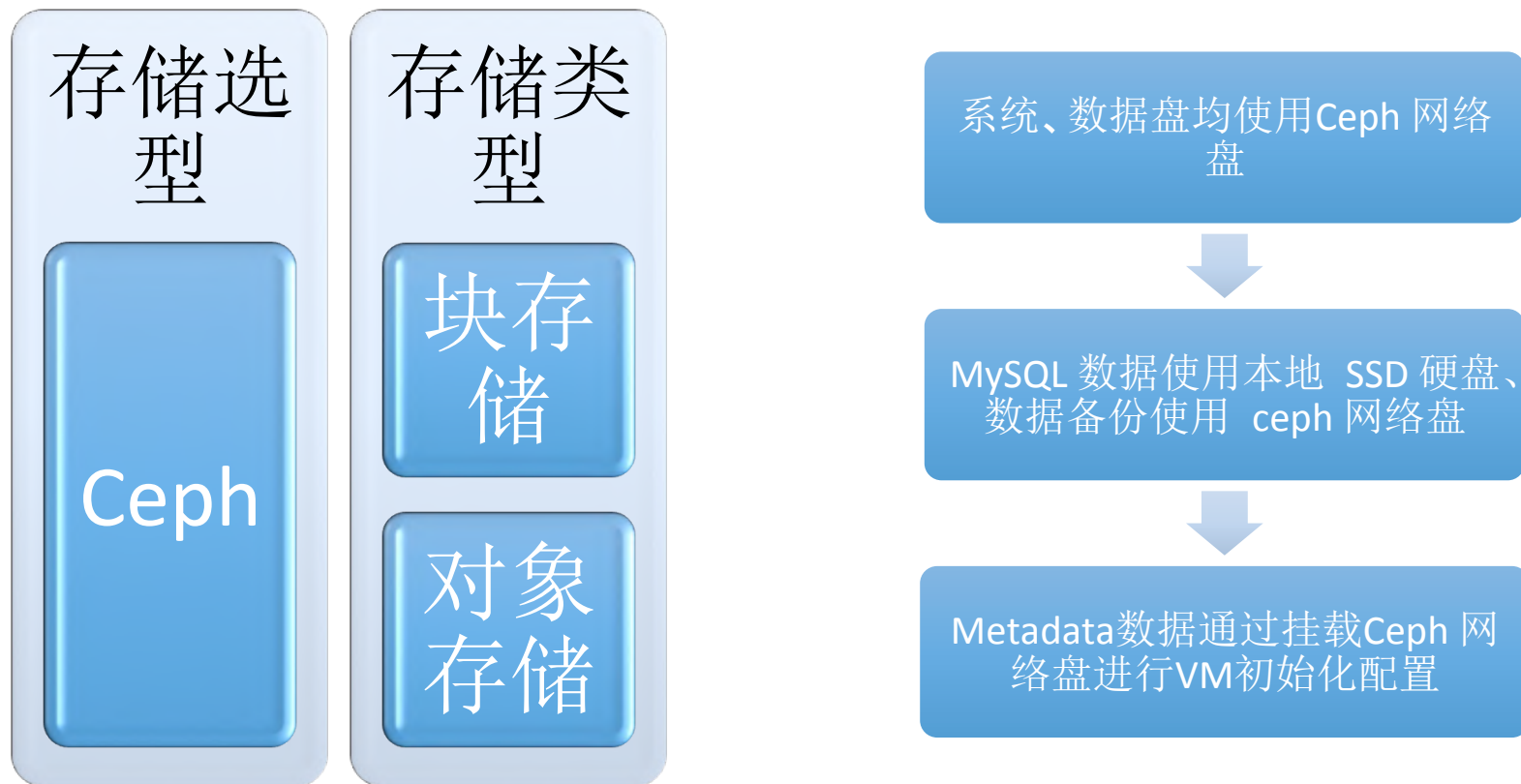
集成方案-计算节点网络结构变化

1.0

2.0



集成方案-虚拟存储方案

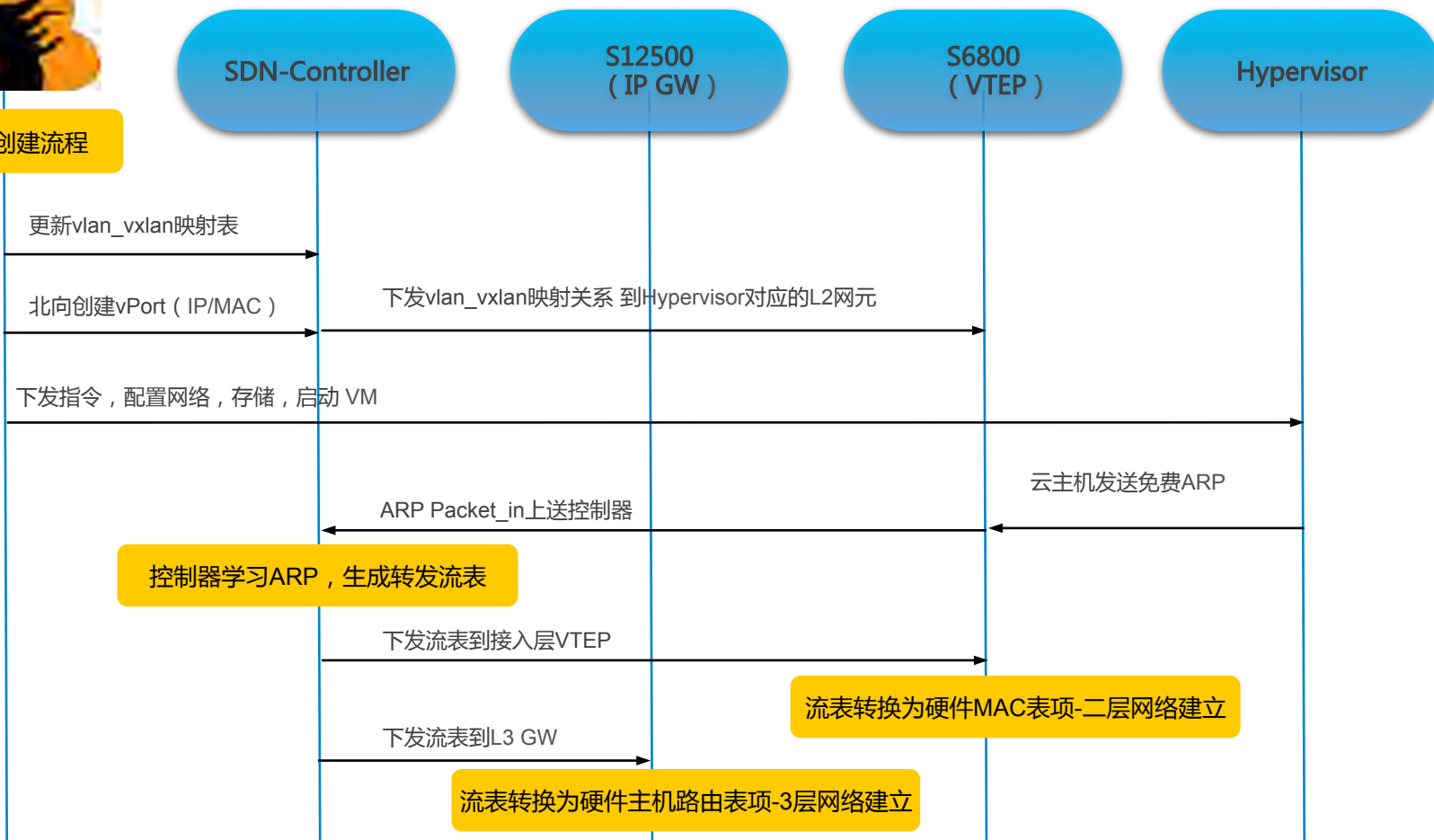


集成方案-VM 创建流程

Cloud2.0云平台



云主机创建流程



运营保障

- 上线和迭代
 - 部分业务测试
 - 功能，稳定性磨合
- 关键组件质量保障
 - 集中式网关的性能，稳定性
 - 集群性能, 稳定性, 容量保障

Thank you !