锐捷极简云数据中心技术白皮书

锐捷网络股份有限公司

前　言

摘要

本文介绍极简云数据中心的组网方案及主要原理。

关键字

VXLAN、EVPN、OVERLAY、数据中心、Openstack 对接、VXLAN L2 Mapping、VXLAN L3 Mapping

目 录

1 概述

1.1 背景

随着 IT 业务的不断发展，数据中心承载的业务越来越多，规模越来越大，传统数据中心面临着成本、速度、整合、安全、能源管理等一系列挑战。为了让数据中心变得更加灵活快速响应业务变更需求、同时降低能耗与运营成本，云数据中心随之出现。

传统数据中心面临许多问题：

1）业务无法弹性扩展

传统数据中心的业务网络构建和物理位置绑定（通常业务主机的网关部署在 TOR 或 EOR，业务主机也就绑定在该 TOR 或 EOR 的下联区域内），业务扩容需要提前规划和预留。而通常由于物理资源限制（如机柜空间）和业务不可预测性，前期无法做好充分的规划和预留。导致后续业务扩容可能需要进行较大规模的网络调整并中断业务。

2）业务无法灵活热迁移

业务扩容、业务搬迁、业务热备过程中，为了保持业务对外不中断，需要保障迁移前后业务主机的 IP 不变。由于业务网络构建和物理位置绑定，导致业务只能在区域内实现热迁移，无法灵活跨区域热迁移。

3）业务上线/变更慢

业务上线或变更需要涉及网络资源分配、大量网络配置变更和下发、甚至需要调整既有业务的网络配置。整个过程不仅对人员的技术要求高，需要耗费大量人力、时间，而且容易出错。

4）网络规模受限

传统数据中心用 VLAN 隔离业务网络。业务网络规模受限于 VLAN 4094 的容量限制，无法满足业务快速增长的需求。

5）计算/存储/网络资源无法统一管理

传统数据中心的计算、存储、网络资源都是分开独立管理，无法有效、快速的协同完成业务的部署、上线、变更。

6）网络运维复杂/故障排查困难

传统数据中心的网络运维管理、故障排查基本依赖 ping、trace、网管软件等方式，对外部依赖性高、排查步骤复杂，而且无法基于业务进行精准问题分析。

1.2 方案简介

在数据中心虚拟化、云化的发展趋势下，锐捷网络推出“极简云数据中心解决方案”。采用 VXLAN 网络虚拟化技术解耦业务网络和物理网络，满足业务弹性扩展、灵活迁移、规模受限的需求。通过 SDN 控制器抽象业务网络模型，实现业务网络配置自动化，降低了网络管理人员的技术要求，大幅缩短业务上线/变更的时间。通过 SDN 控制器和云管平台对接，实现云平台对计算、存储、网络资源的统一管理，有效、快速的协同完成业务部署、上线、变更。

运维管理方面，SDN 控制器模拟设备、主机及业务进行路径和连通性探测，提供简单、精准的故障排查方式。同时，SDN 控制器监控整网设备的运行状态、实时性能，直观的、可视化的呈现 Underlay/Overlay 网络运行状态。

2 技术介绍

2.1 架构简介

2.1.1 云化架构

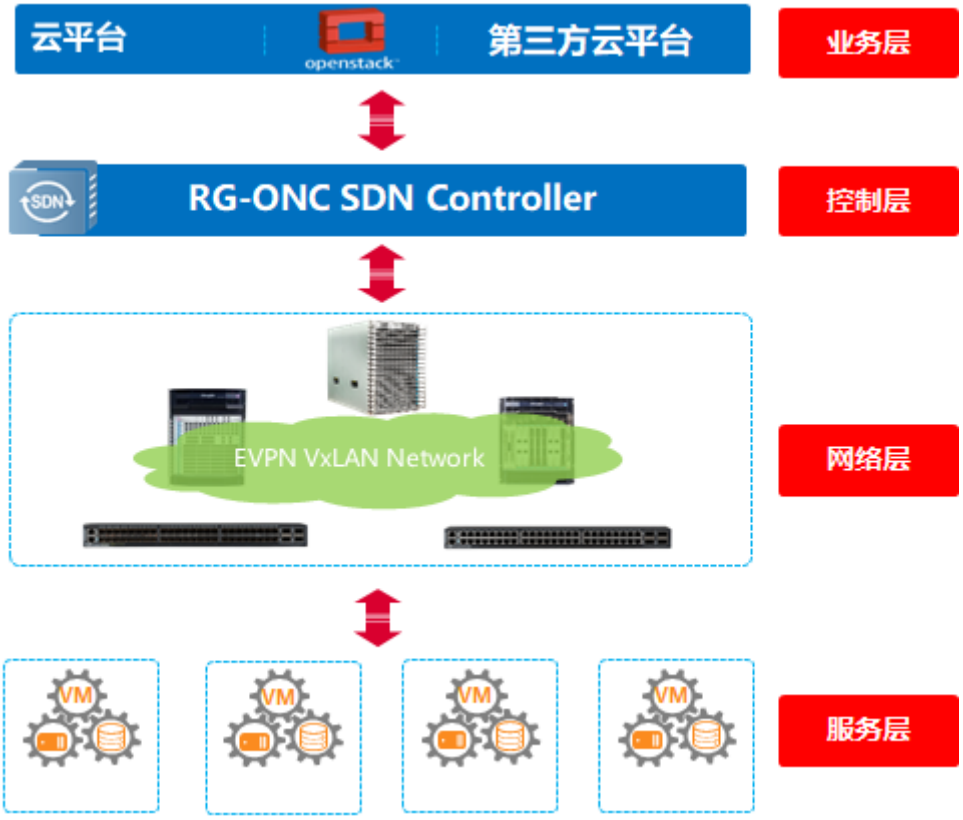


图 2-1 云化数据中心架构图

- 业务层

云平台，负责业务编排和虚拟化资源管理（包括计算资源、网络资源、存储资源的统一管理）。有时也把虚拟化资源管理功能单独部署。

- 控制层

SDN 控制器，管理控制物理网络设备及虚拟网络设备，协同云平台实现业务自动化上线及计算、网络、存储资源的统一管理。

SDN 控制器通过开发的 RESTful API 和云平台对接。对于以 Openstack 为基线的云平台，可通过锐捷网络插件对接 SDN 控制器。

- 网络层

由物理网络设备或虚拟网络设备组成的业务数据转发实体。SDN 控制器通过 Netconf、SNMP、Openflow 等南向协议管理控制这些网络设备实现业务转发。

- 服务层

各类形态的计算资源（如虚拟机、裸金属等）和存储资源。

2.1.2 控制器

图 2-2 控制器架构图

SDN 控制器基于 ODL 平台实现 ,兼容 ONOS 开源控制器平台。北向提供开放的 RESTful API 和 WEB 及云平台对接（通过 RESTful API，第三方平台可灵活扩展新功能），南向通过丰富协议通道和各类网络设备对接。

2.2 硬件 Overlay 架构

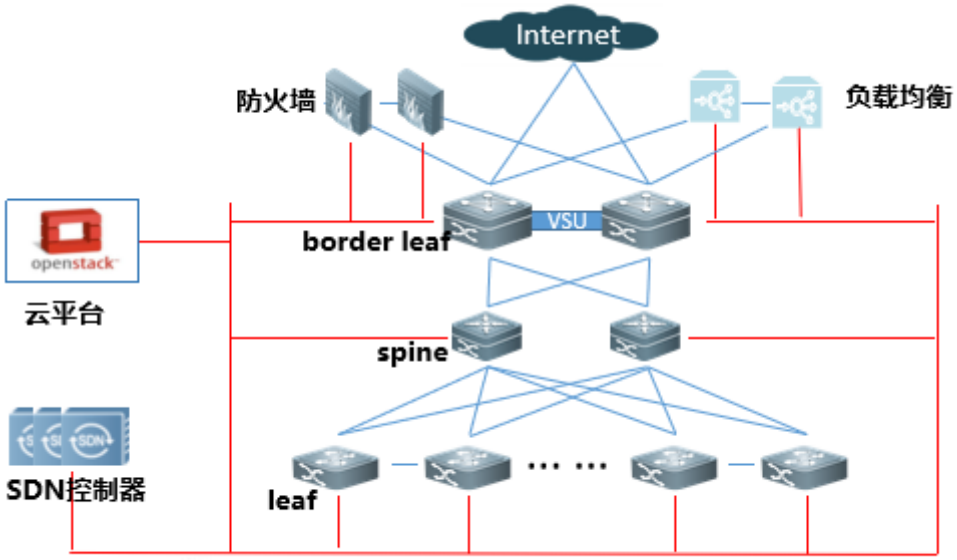


图 2-3 硬件 VXLAN 组网架构图

VXLAN 封装解封装都由硬件交换机实现的组网方案。整体架构如上图。

- **Border Leaf**
  - 作为 VXLAN 三层网关，负责南北向流量和跨 VPC 的流量转发。
  - VSU 堆叠方式部署
  - 防火墙和负载均衡旁挂在 Border Leaf，通过聚合口互联
- **Service Leaf**
  - 数据中心内部或数据中心间互访时网络服务的统一调度节点
  - 数据中心内部或数据中心间访问防火墙和负载均衡器时，由 Service Leaf 完成 VXLAN 封装

/解封装

- **Spine**
  - 东西向流量汇聚设备
  - 和 Leaf、Border Leaf 之间通过三层路由口互联，整网通过 ECMP 等价路由进行负载均衡
  - 可以和 Border Leaf 合并，Leaf 直接三层路由口连接到 Border Leaf
- **Leaf**
  - 接入设备。单机、VSU 堆叠或 M-LAG 部署
  - VSU 堆叠方式，通过聚合口和服务器互联。M-LAG 方式，通过 M-LAG 和服务器互联
  - 集中式 VXLAN，Leaf 节点作为 VXLAN 二层网关。分布式 VXLAN，Leaf 节点作为 VXLAN 三层网关
  - Leaf 节点采用 M-LAG 方式时，组成 M-LAG 的 Leaf 节点组配置相同的 VTEP IP，通过 Underlay 的 ECMP 路由实现 VXLAN 流量的负载均衡

2.2.1 集中式 VXLAN

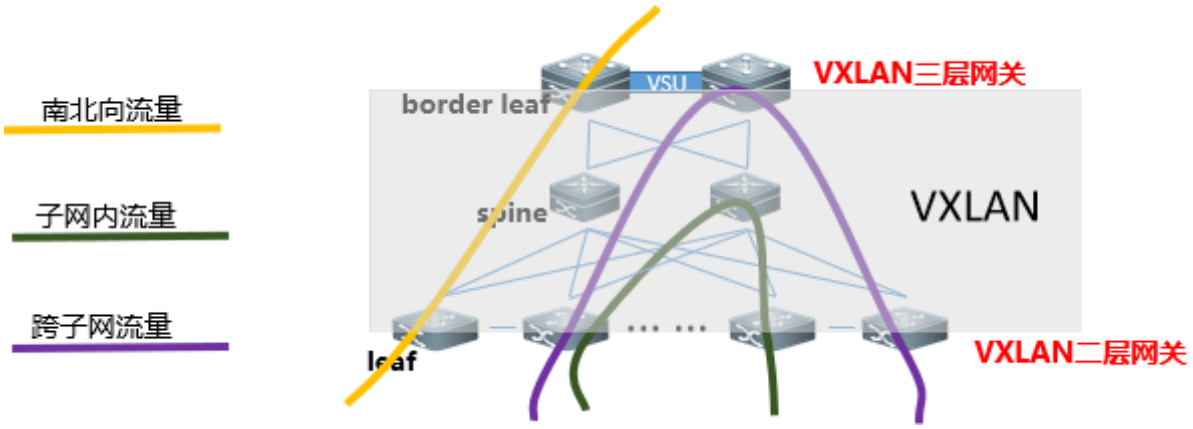


图 2-4 集中式 VXLAN 示意图

- **组网模型**
  - Border Leaf 作为 VXLAN 三层网关（即业务网关），负责跨子网和南北向流量转发
  - Leaf 作为 VXLAN 二层网关，负责业务 VLAN/VXLAN 间的映射、子网内流量转发
  - Leaf 间、Leaf 和 Border Leaf 间通过 EVPN 同步 VXLAN 转发表项
- **流量模型**
  - 南北向流量：出外网的流量在 Leaf 节点经过 VXLAN 封装后，通过 Overlay 转发到 Border Leaf 节点，Border Leaf 进行 VXLAN 解封装后，查找路由表转发到外网。反之，外网进来的流量在 Border Leaf 根据 ARP 表/MAC 表查找到目的 Leaf 节点，经过 VXLAN 封装后转发到目的 Leaf 节点，Leaf 节点进行 VXLAN 解封装后，查找 MAC 表转发到本地主机。如上图黄色线条所示。

- 子网内流量 :子网内访问流量在源 Leaf 节点根据 MAC 表查找到目的 Leaf 节点 ,经过 VXLAN 封装后转发到目的 Leaf 节点，目的 Leaf 节点进行 VXLAN 解封装后，查找 MAC 表转发到本地主机。如上图绿色线条所示。

- 跨子网流量 :跨子网访问流量在源 Leaf 节点根据 MAC 表查找到 Border Leaf(业务网关)，经过 VXLAN 封装后 (源 VNI) 转发到 Border Leaf，Border Leaf 进行 VXLAN 解封装，根据直连路由/ARP 表/MAC 表查找到目的 Leaf 节点，再经过 VXLAN 封装 (目的 VNI) 转发到目的 Leaf，目的 Leaf 节点进行 VXLAN 解封装后，查找 MAC 表转发到本地主机。如上图紫色线条所示。

- ARP：ARP 请求报文在源 Leaf 节点封装成 VXLAN，依次发送到所有 VTEP。Border Leaf 通过该报文学习到主机 ARP。所有目的 VTEP 进行 VXLAN 解封装后广播到本地主机。ARP 响应报文参考“子网内流量”。

集中式 VXLAN 网络结构相对简单，容易管理，但所有的子网外流量都通过 Border Leaf 转发，对 Border Leaf 要求较高，适用中小规模数据中心。

2.2.2 分布式 VXLAN

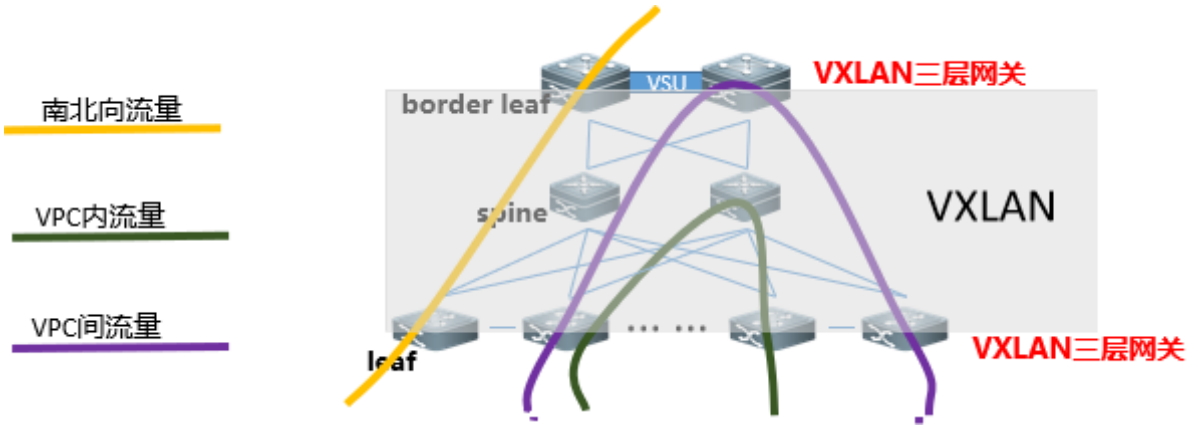


图 2-5 分布式 VXLAN 示意图

- **组网模型**
  - Border Leaf 作为 VXLAN 三层网关（即业务网关），负责跨 VPC 和南北向流量转发
  - Leaf 也作为 VXLAN 三层网关（即业务网关），负责业务 VLAN/VXLAN 间的映射、VPC 内流量（包括 VPC 内同二层网络、跨二层网络流量）转发
  - Leaf 间、Leaf 和 Border Leaf 间通过 EVPN 同步 VXLAN 转发表项
  - Spine 节点不感知 VXLAN 隧道，只作为 VXLAN 报文的转发节点
  - 采用对称式转发模型。每个二层网络对应一个 L2 VNI，每个 VPC 对应一个 L3 VNI。二层网络内通过 L2 VNI 互访，VPC 内不同网络通过 L3 VNI 互访。
- **流量模型**
  - 南北向流量：出外网的流量在 Leaf 节点经过 VXLAN（**L3 VNI**）封装后，通过 Overlay 转

发到 Border Leaf 节点，Border Leaf 进行 VXLAN 解封装后，查找路由表转发到外网。反之，外网进来的流量在 Border Leaf 根据**路由表**查找到目的 Leaf 节点，经过 VXLAN（**L3 VNI**）封装后转发到目的 Leaf 节点，Leaf 节点进行 VXLAN 解封装后，查找 MAC 表转发到本地主机。如上图黄色线条所示。

- VPC 内流量：VPC 内同二层网络访问流量在源 Leaf 节点根据 MAC 表查找到目的 Leaf 节点，经过 VXLAN（**L2 VNI**）封装后转发到目的 Leaf 节点，目的 Leaf 节点进行 VXLAN 解封装后，查找 MAC 表转发到本地主机；VPC 内跨二层网络访问流量在源 Leaf 节点根据路由表查找到目的 Leaf 节点，经过 VXLAN(**L3 VNI**)封装后转发到目的 Leaf 节点，目的 Leaf 节点进行 VXLAN 解封装后，查找 MAC 表转发到本地主机。如上图绿色线条所示。

- VPC 间流量：VPC 间访问流量在源 Leaf 节点根据**静态路由**查找到 Border Leaf，经过 VXLAN 封装后（**源 VPC 的 L3 VNI**）转发到 Border Leaf，Border Leaf 进行 VXLAN 解封装，根据路由表查找到目的 Leaf 节点，再经过 VXLAN 封装（**目的 VPC 的 L3 VNI**）转发到目的 Leaf，目的 Leaf 节点进行 VXLAN 解封装后，查找 MAC 表转发到本地主机。如上图紫色线条所示。

- ARP 代答：ARP 请求报文达到源 Leaf 节点，源 Leaf 节点学习到源主机 ARP 并通过 EVPN 同步给所有 VTEP，同时源 Leaf 节点直接代答 ARP 请求。

分布式 VXLAN 网络结构相对复杂，但东西和南北向在 Leaf 和 Border Leaf 间分流，均衡 Leaf 和 Border Leaf 的压力，适用于大规模数据中心。

2.2.3 分布式 VXLAN（Service Leaf）

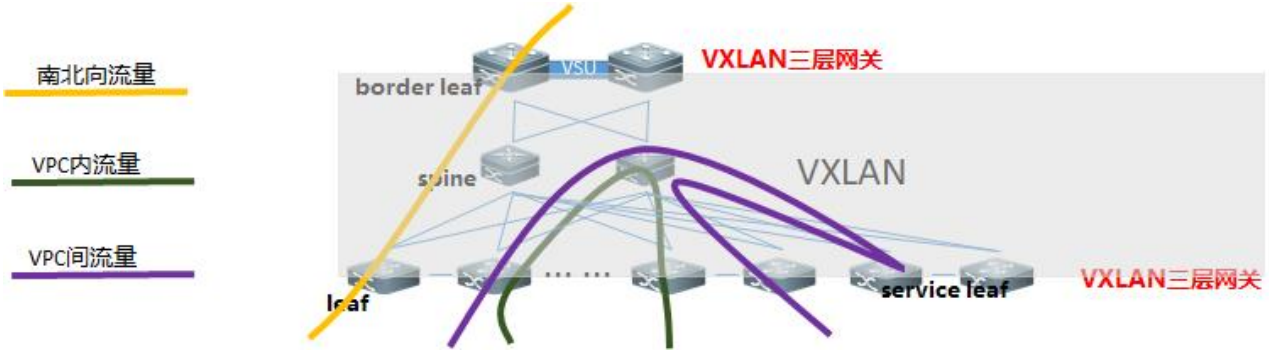


图 2-6 分布式 VXLAN(Service Leaf)示意图

- **组网模型**

  - Border Leaf 作为 VXLAN 三层网关（即业务网关），负责南北向流量转发及数据中心内部无需过墙的跨 VPC 流量转发。

  - Service Leaf 作为数据中心内部或数据中心之间网络互访服务的调度节点，负责跨 VPC 流量转发

  - Leaf 也作为 VXLAN 三层网关（即业务网关），负责业务 VLAN/VXLAN 间的映射、VPC 内流量（包括 VPC 内同二层网络、跨二层网络流量）转发

  - Border Leaf、Service Leaf、Leaf 间通过 EVPN 同步 VXLAN 转发表项

  - Spine 节点不感知 VXLAN 隧道，只作为 VXLAN 报文的转发节点

  - 采用对称式转发模型。每个二层网络对应一个 L2 VNI，每个 VPC 对应一个 L3 VNI。二层网络内通过 L2 VNI 互访，VPC 内不同网络通过 L3 VNI 互访。

- **流量模型**

  - 南北向流量：出外网的流量在 Leaf 节点经过 VXLAN（**L3 VNI**）封装后，通过 Overlay 转发到 Border Leaf 节点，Border Leaf 进行 VXLAN 解封装后，查找路由表转发到外网。反之，外网进来的流量在 Border Leaf 根据**路由表**查找到目的 Leaf 节点，经过 VXLAN（**L3 VNI**）封装后转发到目的 Leaf 节点，Leaf 节点进行 VXLAN 解封装后，查找 MAC 表转发到本地主机。

  - VPC 内流量：VPC 内同二层网络访问流量在源 Leaf 节点根据 MAC 表查找到目的 Leaf 节点，经过 VXLAN（**L2 VNI**）封装后转发到目的 Leaf 节点，目的 Leaf 节点进行 VXLAN 解封装后，查找 MAC 表转发到本地主机；VPC 内跨二层网络访问流量在源 Leaf 节点根据路由表查找到目的 Leaf 节点，经过 VXLAN(**L3 VNI**)封装后转发到目的 Leaf 节点，目的 Leaf 节点进行 VXLAN 解封装后，查找 MAC 表转发到本地主机。

  - VPC 间流量（不过防火墙）：VPC 间访问流量在源 Leaf 节点根据**静态路由**查找到 Border Leaf，经过 VXLAN 封装后(**源 VPC 的 L3 VNI**)转发到 Border Leaf，Border Service Leaf 进行 VXLAN 解封装，根据路由表查找到目的 Leaf 节点，再经过 VXLAN 封装（**目的 VPC 的 L3 VNI**）转发到目的 Leaf，目的 Leaf 节点进行 VXLAN 解封装后，查找 MAC 表转发到本地主机。

  - VPC 间流量(过防火墙):VPC 间访问流量在源 Leaf 节点经过 VXLAN 封装后转发到 Service Leaf，Service Leaf 对报文进行 VXLAN 解封装，根据静态路由转发给东西向防火墙，东西向防火墙进行流量清洗后再转发给 Service Leaf，由 Service Leaf 再经过 VXLAN 封装，然

后根据路由表查找到目的 Leaf 节点，目的 Leaf 节点进行 VXLAN 解封装后，查找 MAC 表转发到本地主机。

- ARP 代答：ARP 请求报文达到源 Leaf 节点，源 Leaf 节点学习到源主机 ARP 并通过 EVPN 同步给所有 VTEP，同时源 Leaf 节点直接代答 ARP 请求。

2.3 防火墙三层旁挂 Border Leaf

2.3.1 Border Leaf 堆叠组网

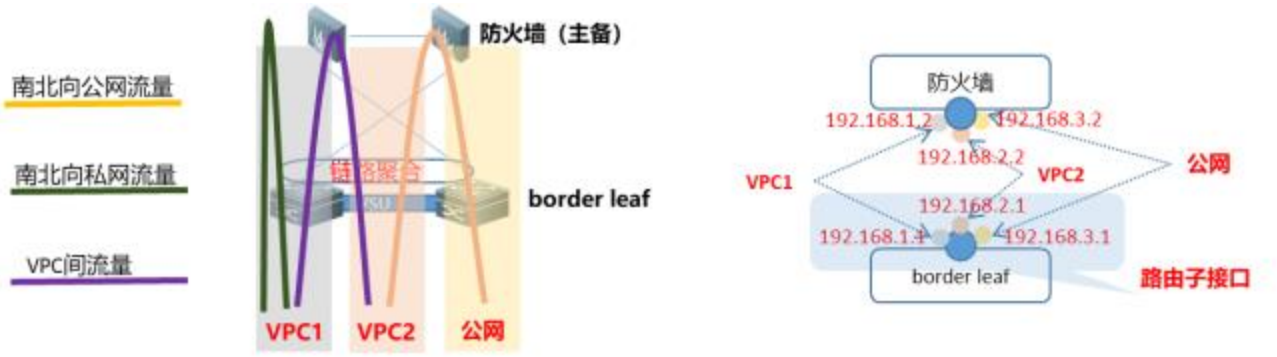


图 2-7 防火墙三层旁挂 Border Leaf（VSU）示意图

- **组网模型**

  - 防火墙旁挂在 Border Leaf，通过聚合口互联，防火墙采用主备方式。

  - 防火墙和 Border Leaf 之间通过三层互访。在 Border Leaf 上每个 VPC 对应一个子接口和防火墙互联（每 VPC 对应一个虚墙，通过各自的路由子接口互联），全局网络对应一个子接口和防火墙互联（如果 Underlay 网络存在多个安全域则每个安全域对应一个子接口和防火墙互联）。

- Border Leaf 上 VPC（出口关联防火墙的 VPC）默认路由指向防火墙（访问其它 VPC 走静态路由，下一跳地址是该 VPC 子接口对应的防火墙地址）。全局网络配置 BGP，防火墙通过 BGP 同步 NAT 的路由到 Border Leaf（也可在 Border Leaf 配置 NAT 的静态路由）。

- **流量模型**

  - 南北向公网流量：南北向公网指的是 VPC 对应的外部网络直接就是公共的外部网络域。主机访问外网流量到达 Border Leaf 后，经过 VXLAN 解封装，根据默认路由，通过 VPC 对应的子接口转发到防火墙。防火墙处理后（如果需要 NAT，则在防火墙进行 NAT 转换），发送到 Border Leaf 的全局网络子接口，Border Leaf 再根据 Underlay 路由表转发流量到外部网络；外网访问主机流量到达 Border Leaf 后，根据路由表，通过全局网络子接口转发到防火墙，防火墙处理后（如果需要 NAT，则在防火墙进行 NAT 转换），发送到 Border Leaf 的 VPC 子接口，Border Leaf 再通过 VXLAN 转发到主机所在的 Leaf 节点。如上图黄色线条所示。

  - 南北向私网流量：南北向私网指的是 VPC 对应的外部网络就是该 VPC 专用的外部网络域，如上图绿色线条所示。

  - VPC 间流量：VPC 间访问流量在到达 Border Leaf，经过 VXLAN 解封装后，根据路由表，通过源 VPC 的子接口转发到防火墙，防火墙处理后，发送到 Border Leaf 的目的 VPC 子接口。如上图紫色线条所示。

2.3.2 Border Leaf 去堆叠组网

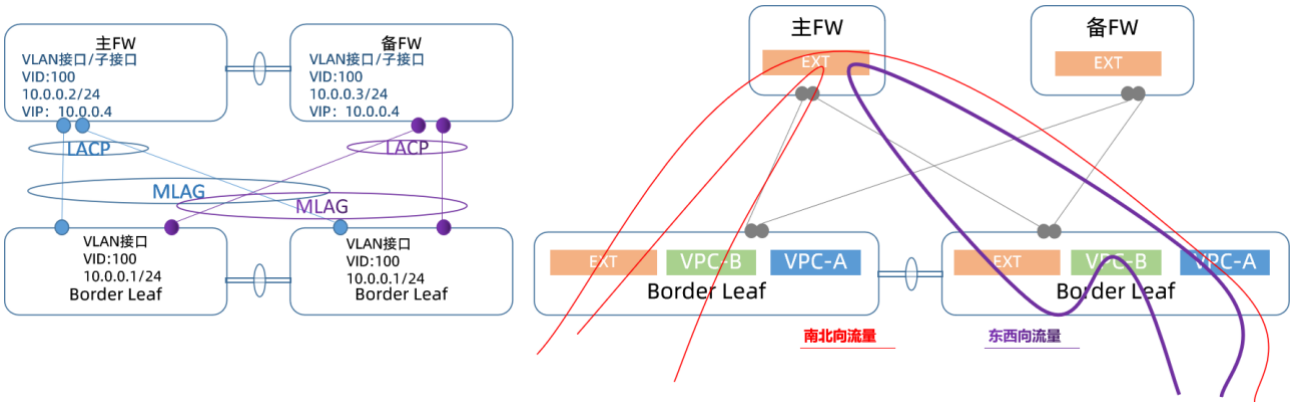


图 2-8 防火墙三层旁挂 Border Leaf（MLAG）示意图

- **组网模型**

  - 防火墙旁挂在 Border Leaf，采用主备模式，主备防火墙各通过一组聚合口和 Border 互联（主备防火墙支持跨设备聚合，也可共同采用一组聚合口）。Border Leaf 采用跨设备链路聚合。

  - 防火墙和Border Leaf之间采用三层路由模式，防火墙配置VLAN接口或路由子接口，Border配置多活 VLAN 接口。Border 、防火墙、外部交换机（或路由器）之间创建外部出口网络，VPC 间互访、VPC 南北向访问都经过外部出口网络。

  - Border Leaf 上 VPC 默认路由指向防火墙（下一跳地址是防火墙在外部出口网络的地址）。同时把 VPC 子网路由泄漏到外部出口网络。

- **流量模型**

  - 南北向流量：主机访问外网流量到达 Border Leaf 后，经过 VXLAN 解封装，根据默认路由，

通过外部出口网络 VLAN 接口转发到防火墙。防火墙处理后（如果需要 NAT，则在防火墙进行 NAT 转换），发回到 Border Leaf，Border Leaf 再根据 Underlay 路由表转发流量到外部网络；外网访问主机流量到达 Border Leaf 后，根据策略路由，通过外部出口网络 VLAN 接口转发到防火墙，防火墙处理后（如果需要 NAT，则在防火墙进行 NAT 转换），发回到 Border Leaf，Border Leaf 根据 Underlay 路由表找对对应的 VXLAN 子接口，再通过 VXLAN 转发到主机所在的 Leaf 节点。如上图红色线条所示。

- VPC 间流量：主机访问外网流量到达 Border Leaf 后，经过 VXLAN 解封装，根据默认路由，通过外部出口网络 VLAN 接口转发到防火墙。防火墙处理后（如果需要 NAT，则在防火墙进行 NAT 转换），发回到 Border Leaf，Border Leaf 根据 Underlay 路由表找对对应的 VXLAN 子接口，再通过 VXLAN 转发到主机所在的 Leaf 节点。如上图紫色线条所示。

2.4 防火墙三层旁挂 Service Leaf

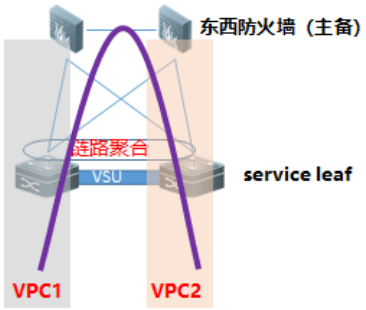


图 2-9 防火墙三层旁挂 Service Leaf（堆叠）示意图

2.4.1 Service Leaf 堆叠组网

- **组网模型**

- 对于南北向和东西向防火墙分别部署场景，南北向防火墙旁挂在 Border Leaf，两者通过聚合口互联；东西向防火墙用类似的方式挂载在 Service Leaf。防火墙采用主备方式部署。

- 东西向防火墙和 Service Leaf 之间通过三层互访。在 Service Leaf 上每个 VPC 对应一个子接口和防火墙互联（每 VPC 对应一个虚墙，通过各自的路由子接口互联），全局网络对应一个子接口和防火墙互联（如果 Underlay 网络存在多个安全域则每个安全域对应一个子接口和防火墙互联）。

- Service Leaf 上过墙 VPC 互访的静态路由指向东西向防火墙，下一跳地址是该 VPC 子接口对应的防火墙地址。

- **流量模型**

- VPC 间流量： VPC 间访问流量在到达 Service Leaf，经过 VXLAN 解封装后，根据路由表，通过源 VPC 的子接口转发到防火墙，防火墙处理后，发送到 Service Leaf 的目的 VPC 子接口。如上图紫色线条所示。

2.4.2 Service Leaf 去堆叠组网

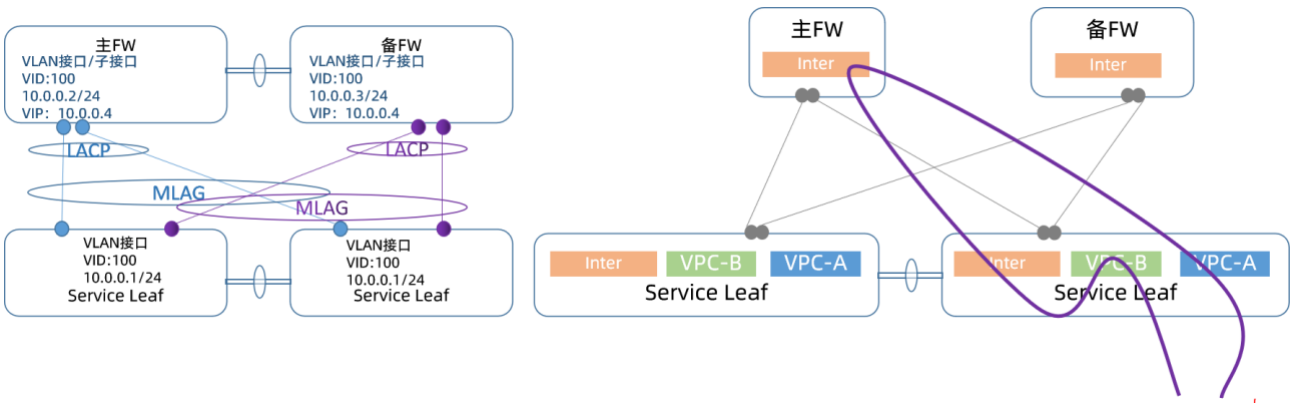


图 2-10 防火墙三层旁挂 Service Leaf（MLAG）示意图

Service Leaf 去堆叠组网防火墙旁挂模型和 Border Leaf 去堆叠完全一致，可参考 Border Leaf 去堆叠的组网模式及 VPC 间流量模型。差别仅在于，Service Leaf 上 VPC 子网互访路由指向防火墙（而不是默认路由）。

2.5 负载均衡器二层邻接 Border Leaf

2.5.1 Border Leaf 堆叠组网

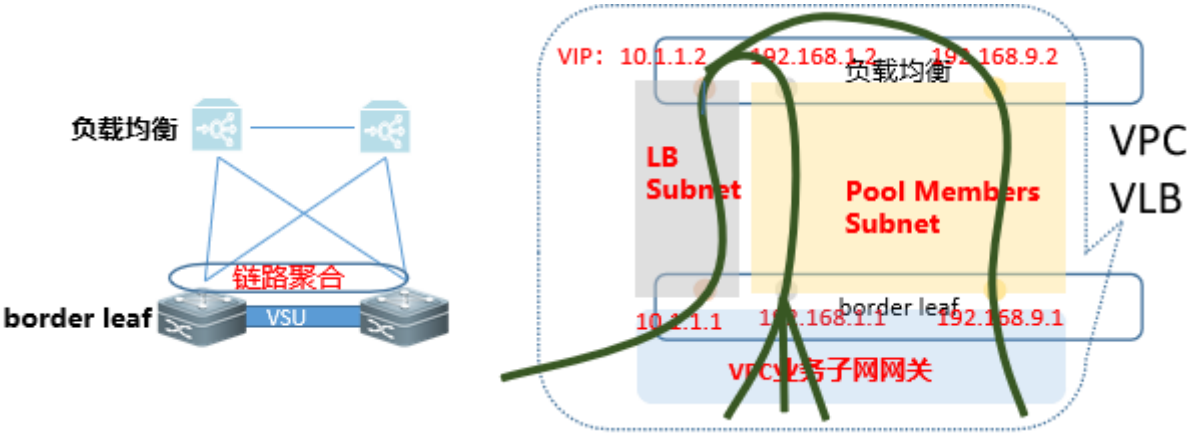


图 2-11 负载均衡器二层邻接 Border Leaf（堆叠）示意图

- **组网模型**

  - 负载均衡器（LB，Load Balancer）旁挂在 Border Leaf，通过聚合口互联。负载均衡器采用二层邻接模式。

  - 负载均衡器启用 SNAT 模式，每 member subnet 一个 SNAT 地址（也可设置多个）。如果

不启用 SNAT，则负载均衡器直接通过 VIP 所在的 subnet 和 pool member 互联。

- **流量模型**

  - 访问 VIP 流量：访问 VIP 流量到达 Border Leaf 后，根据 ARP 表/MAC 表，查找到 LB 对应的端口，通过 Underlay 发送到 LB。LB 监听到 VIP 流量，选择 Pool Member，进行 NAT 后把流量通过二层转发到 Pool Member 在 Border Leaf 的对应网络。Border Leaf 根据 MAC 表查找到对应的目的 VTEP 经过 VXLAN 封装后（**Pool Member 对应网络的 L2 VNI**）发到 Leaf 节点。

  - member 应答流量：member 应答的目的地址是 pool member subnet 在负载均衡对应的 SNAT 地址，类似上述 VPC 内子网内访问流量，member 应答报文发送到负载均衡，负载均衡经过 NAT 后回复给访问源；对于未经过 SNAT 的访问流量（**或业务需要 member 通过其它方式直接识别到访问源**），member 直接应答流量给访问源（**不经过负载均衡**）。

2.5.2 Border Leaf 去堆叠组网

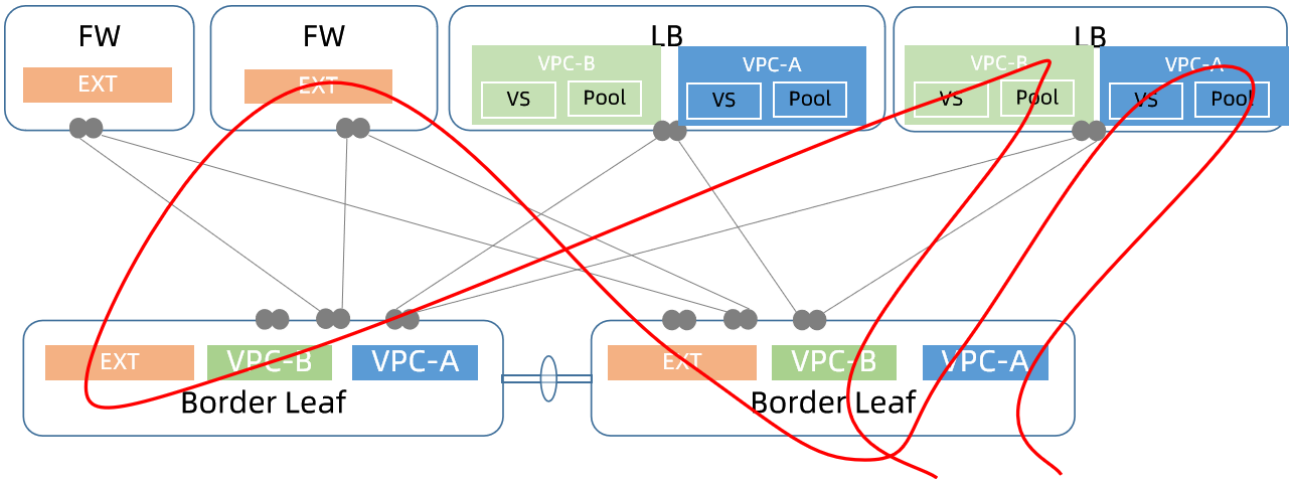


图 2-12 负载均衡器二层邻接 Border Leaf（去堆叠）示意图

LB 通过二层邻接到 Border Leaf，映射到 VXLAN 接口，天然支持多活。组网及流量模型和堆叠组网一致。差别在于去堆叠模式，Border 采用跨设备链路聚合。

2.6 多 Border Leaf（多出口网关组）

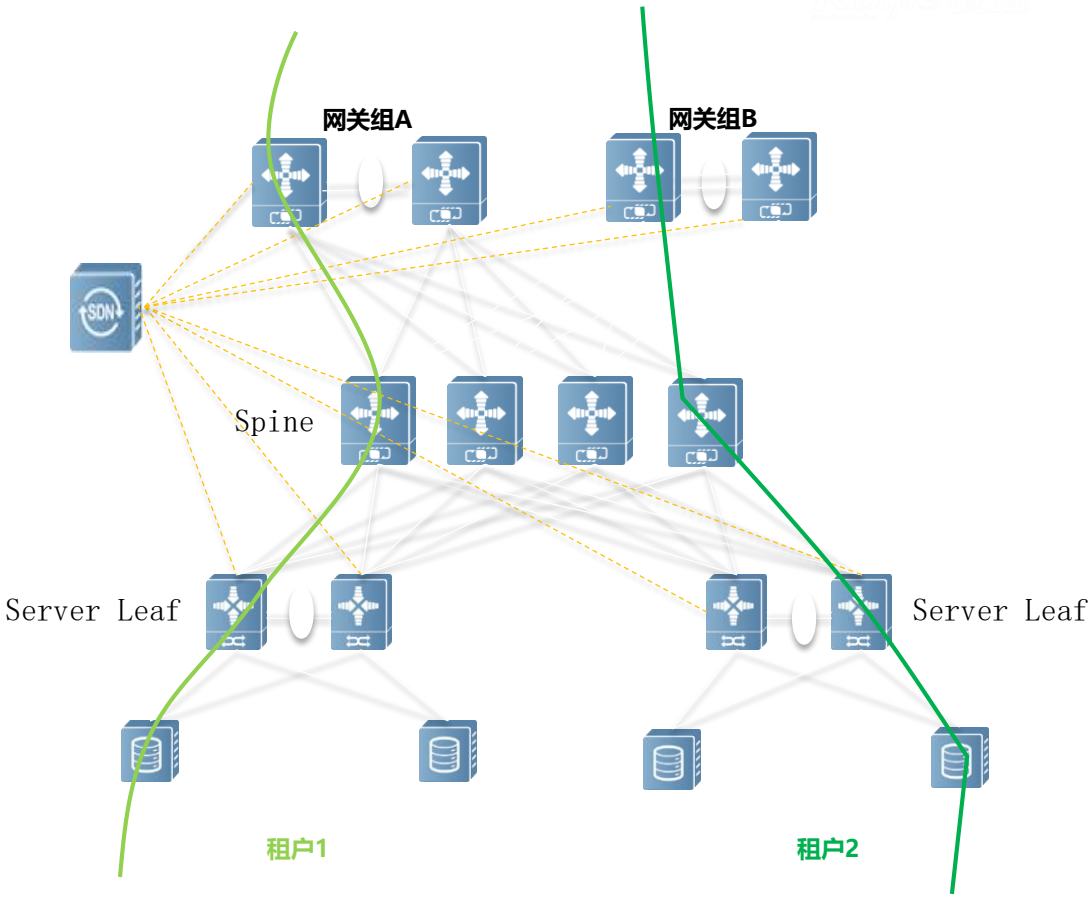


图 2-13 多 Border Leaf 网络架构

- 组网模型
  - SDN 控制器对多个 Border Leaf 设备进行分组，相同 VTEP IP 的 Border Leaf 设备划入同

一个网关组。

- 基于租户选择不同的 Border Leaf 组作为出口网关。

- 流量模型
  - 多个 Border Leaf 组之间物理隔离，不同的租户经过指定的 Border Leaf 组访问数据中心外的网络
  - 南北向流量：VPC 流量在 Leaf 节点进行 VXLAN 封装后，根据路由表将报文转发到 VPC（租户）对应的 Border Leaf，经过 VXLAN 解封装，根据默认路由，通过 VPC 对应的子接口转发到防火墙。防火墙处理后（如果需要 NAT，则在防火墙进行 NAT 转换），发送到 Border Leaf 的全局网络子接口，Border Leaf 再根据 underlay 路由表转发流量到外部网络；外网访问主机流量到达 Border Leaf 后，根据路由表，通过全局网络子接口转发到防火墙，防火墙处理后（如果需要 NAT，则在防火墙进行 NAT 转换），发送到 Border Leaf 的 VPC 子接口，Border Leaf 再通过 VXLAN 转发到主机所在的 Leaf 节点。
  - VPC 间流量：VPC 间访问流量在源 Leaf 节点经过 VXLAN（**源 VPC 的 L3 VNI**）封装后，根据**静态路由**查找到 VPC（租户）对应的 Border Leaf， Border Leaf 进行 VXLAN 解封装，根据路由表查找到目的 Leaf 节点，再经过 VXLAN 封装（**目的 VPC 的 L3 VNI**）转发到目的 Leaf，目的 Leaf 节点进行 VXLAN 解封装后，查找 MAC 表转发到本地主机。

2.7 业务网络编排及自动化

2.7.1 租户/业务网络

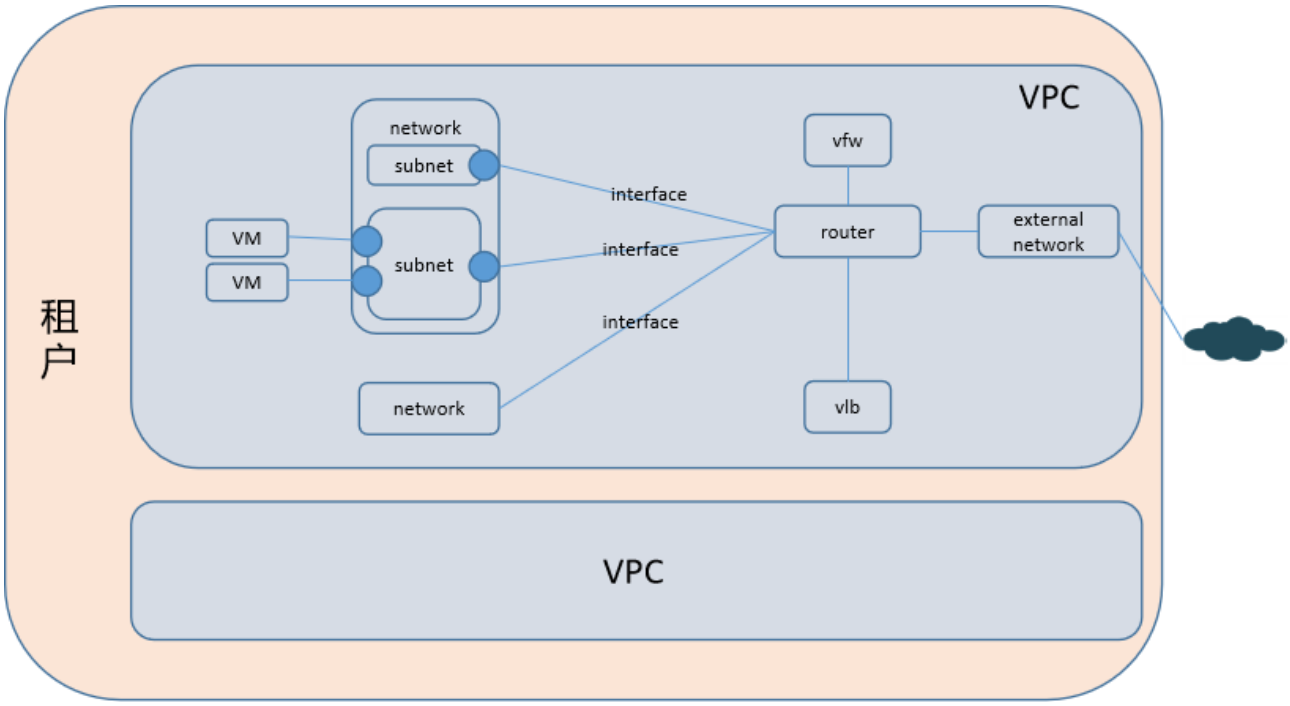


图 2-14 租户/业务网络示意图

租户/业务网络独立于物理网络，是逻辑上的概念。网络架构如上图所示，网络元素说明如下

- **逻辑网络元素**
  - Router：逻辑路由器。负责子网间、子网和外部网络之间的报文转发。
  - Network :二层网络 ,分为内部 network 和外部 network(也成为出口网络)。内部 network 负责子网内报文转发。外部 network 负责外部子网的报文转发。
  - Subnet：子网。Network 所属的子网段。

- Interface：逻辑接口。子网和逻辑路由器连接的接口。
- Port：逻辑端口。network 和外部终端（如：主机、防火墙、负责均衡等）连接的端口。
- 主机：业务主机。通过 port 连入 network。
- Route：业务网络路由。
- VPC：虚拟私有云，即逻辑隔离的私有网络。
- Tenant：租户。每个租户可以有多个 VPC。
- VFW：虚拟防火墙。可以从物理防火墙从虚拟多个防火墙给多个 VPC。通常一个虚拟防火墙只能对应一个 VPC。
- VLB：虚拟负责均衡。可以从物理负载均衡从虚拟多个负载均衡给多个 VPC。通常一个虚拟负载均衡只能对应一个 VPC。
- 安全组：安全访问权限规则组。一组安全访问权限规则（可实现 L4 层的安全访问控制）的集合，应用于指定主机。

2.7.2 逻辑/物理网络映射

租户/业务网络最终是需要映射到物理网络，通过物理网络来实现租户/业务网络的功能定义。租户网络映射到物理网络通过 VLAN/VXLAN 实现二层隔离、通过 VRF 实现三层隔离。租户网络承载在 Overlay 网络上，租户之间互相隔离且独立管理，可基于租户灵活、便捷地规划 IP 网段（租户间 IP 网段可以重叠，

互不影响）。

- **Router**
  - Router 间通过 VRF 隔离，每个 router 对应一个 VRF
  - 分布式 VXLAN，每个 router 还对应一个 L3 VNI
  - Router 承载在所有业务网关上（集中式 VXLAN 的 Border Leaf，分布式 VXLAN 包括所有的 Leaf 和 Border Leaf）
- **Network（内部 network）**
  - Network 通过 vlan/vxlan 隔离，每个 network 对应一个 vid 和 l2 vni
  - Network 承载在所有业务二层互联设备上（VXLAN 方案即所有的 VTEP 和 vSwitch）
  - 虚拟化场景，vSwitch 给接入主机打 vid 标签，VTEP 完成 vid 和 l2 vni 映射
  - 非虚拟化场景，主机接入的 VTEP 设备完成打 vid 标签、vid 和 l2 vni 映射
- **Subnet/Interface**
  - Interface 对应一个 L2 VNI 的 OR 接口（子网同 subnet 定义，接口地址由租户定义）
  - Interface 是关联 subnet 和 router 的元素，承载在 router 对应的物理设备（所有的业务网关）

- **Network/Port**
  - Port 映射到 vSwitch 的虚拟端口和 VTEP 的实际物理端口。一个 VTEP 的实际物理端口可被映射到多个 port
  - Port 映射到 vSwitch 的虚拟端口，则该虚拟端口对应的 VLAN 标签为 network 对应的 vid
  - Port 映射到 VTEP 的物理端口，则该物理端口需要允许 network 对应的 vid。
  - 裸金属场景，一个物理端口只能映射一个 port，且没有 vSwitch 给业务主机打 VLAN 标签，则该物理端口需要给业务主机打 VLAN 标签
- **静态路由**
  - 静态路由直接对应于物理设备的静态路由，承载在 router 对应的物理设备（实际设置时，可设置在 Border Leaf，通过 EVPN 同步到其它 Leaf 节点）
- **VFW（防火墙三层旁挂 Border Leaf）**
  - VFW 对应于物理防火墙上的虚拟防火墙
  - VFW 互联接口，对应于 Border Leaf 和物理防火墙互联端口上的一个子接口
- **External network/出口**
  - 出口网络承载在其实际物理端口所在的 Border Leaf 设备

- 可以映射为外部互联子接口，也可以映射为外部网络路由

2.7.3 业务网络自动化

上文所述，租户/业务网络是逻辑概念，最终是需要映射到物理网络，通过物理网络来实现租户/业务网络的功能定义。而映射的工作是复杂的、不同组网模型不同设备的网络配置还不相同，通过人工实现的效率低下、容易出错。SDN 控制器可根据不同组网模型不同设备角色自动化的把租户/业务网络映射成设备配置，并自动下发到对应设备。下面以业务上线为例，说明 SDN 控制器实现业务网络自动化的流程及原理

- **创建 Network**
  - 控制器自动分配 VID/L2 VNI（或用户指定）
  - 在所有 VTEP 设备生成 VLAN/VXLAN、VLAN/VXLAN 间映射关系、L2 VNI 对应的 EVPN 配置
- **在 Network 下创建 Subnet**
  - 控制器记录相关子网信息
- **创建 Router**
  - 控制器自动分配 VRF，分布式 VXLAN 还分配 L3 VNI
  - 在所有业务网关设备生成 VRF、VXLAN、L3 VNI 对应的 EVPN 配置、路由重分发配置

- **在 Router 下添加 Interface**

  - 控制器根据 Interface 对应的子网和 network 查找到对应的网段信息、L2 VNI 信息

  - 控制器在所有业务网关设备上基于 L2 VNI 生成 OR 口，并把 OR 口添加到 router 所在的 VRF

- **创建过墙的出口网络（external network）**

  - 控制器根据指定的接口名称查找到 Border Leaf 连接防火墙的物理端口

  - 控制器根据指定的本地 IP 地址、VLAN 在 Border Leaf 连接防火墙的物理端口上生成路由子接口，并把该路由子接口添加到 router 所有在 VRF

  - 控制器根据指定的防火墙地址，在 Border Leaf 设置 router 所在 VRF 的默认路由指向防火墙地址

  - 控制器根据指定的 BGP 信息，在 Border Leaf 的全局网络添加 BGP 和防火墙进行 NAT 路由交互。

- **创建无墙的出口网络（external network）**

  - 控制器根据指定的接口名称查找到 Border Leaf 连接外网的物理端口

  - 控制器根据指定的本地 IP 地址、VLAN 在 Border Leaf 连接外网的物理端口上生成路由子接口，并把该路由子接口添加到 router 所有在 VRF。

- 控制器根据指定的外部网络地址，在 Border Leaf 设置 router 所在 VRF 的默认路由指向外部网络地址

- **创建主机/按需下发**

  - 创建主机时，控制器记录主机的 IP、MAC、宿主机和接入位置信息

  - 上述业务流程中，业务网络元素创建时，控制器实时下发配置到相关设备。启用按需下发后，业务网络元素创建时，控制器不会实时下发配置，而是在业务主机上线时，控制器根据业务主机对应的接入位置按需下发网络配置到相应的物理设备。

- **VFW**

  - 本阶段没有云平台的情况下，防火墙的网络配置和规则配置需要通过手工完成

  - 相关物理网络配置通过“过墙的出口网络”完成

- **VLB**

  - 本阶段没有云平台的情况下，负载均衡的网络配置和规则配置需要通过手工完成

  - 相关物理网络配置同一般的 network/subnet。只需创建负载均衡 VIP 和 SNAT 对应的 network/subnet 即可。

- **创建数据中心 L2 互联（VXLAN L2 Mapping）**

  - 控制器根据指定的互联 VNI 和互导 RT，在 Border Leaf 生成 VXLAN 及相关 EVPN 配置

- 控制器根据指定的源 VPC 的 network 查找到对应的 L2 VNI，在 Border Leaf 生成 L2 VNI 和互联 VNI 的 L2 Mapping 配置。

- **创建数据中心 L3 过墙互联（VXLAN L3 Mapping）**

  - 控制器根据指定的互联 VNI 和互导 RT ,分配互联域 VRF ,在 Border Leaf 生成 VRF、VXLAN 及相关 EVPN 配置

  - 控制器根据指定的防火墙，自动分配和防火墙的互联网络和 IP，查找 Border Leaf 连接防火墙的端口，在对应端口上生成互联的路由子接口，并加入互联 VRF

  - 控制器根据指定的本地 router 及远端互访网段 ,在 Border Leaf 的 router 对应 VRF 生成互访静态路由指向 router 对应的防火墙。

  - 控制器根据指定的本地子网，在互联 VNI 对应的 EVPN 通过 route-map 把指定子网路由发布给远端，并在互联 VRF 生成互访静态路由指向对应的防火墙

- **创建数据中心 L3 无墙互联（VXLAN L3 Mapping）**

  - 控制器根据指定的互联 VNI 和互导 RT，在 Border Leaf 生成 VXLAN 及相关 EVPN 配置，并加入租户 VRF

  - 控制器根据指定的本地子网，在互联 VNI 对应的 EVPN 通过 route-map 把指定子网路由发布给远端

- **创建传统数据中心互联（静态路由）**

- 传统数据中心互联即设置静态路由

- **添加静态路由（外部网络）**

  - 控制器根据指定的 router，在 Border Leaf 的 router 对应的 VRF 生成对应的静态路由

- **创建路由器互通（即 VPC 间互通）**

  - 控制器根据指定的源和目的 router/interface，查找对应的 VRF 和子网段，在 Border Leaf 的源和目的 VRF 内分别设置对应的互访静态路由指向各自 VPC 连接的防火墙子接口地址

  - 本阶段只支持过墙的 VPC 互通

2.8 Openstack 平台对接

2.8.1 Neutron plugin 简介

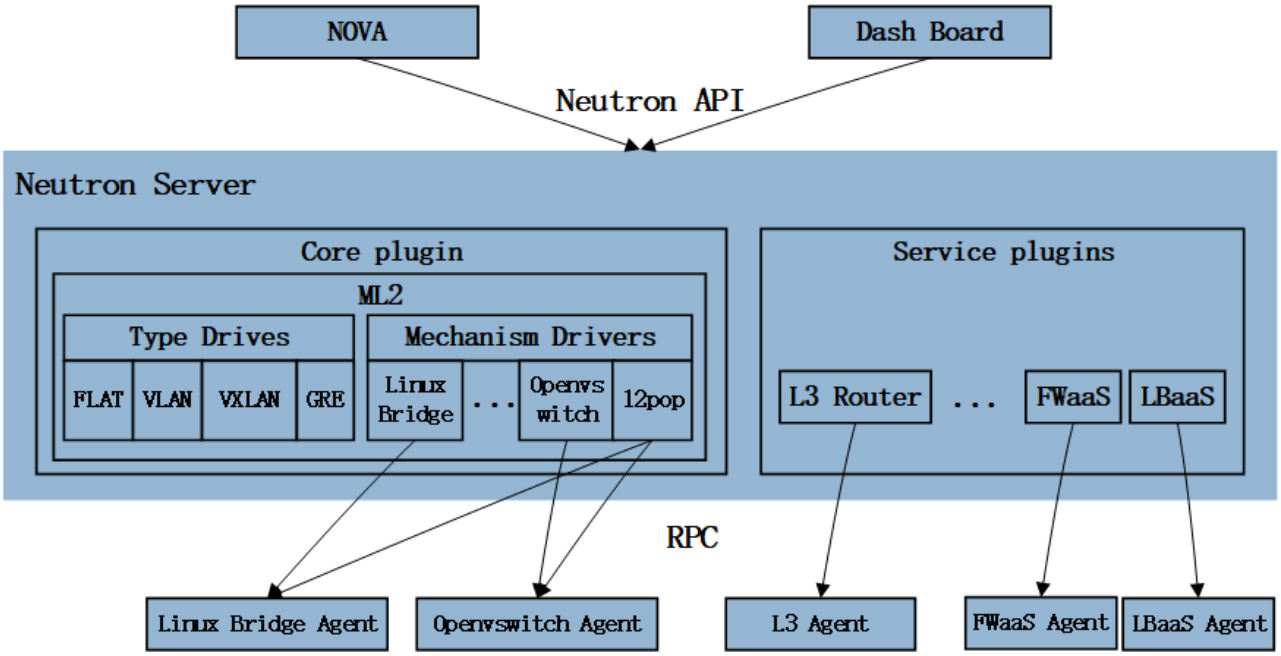


图 2-15 Neutron plugin 示意图

Neutron 是 Openstack 的核心组件之一，实现网络及网络服务的自动化配置管理。Neutron 主要通过以下三大功能模块实现网络服务自动化：

- **Core-plugin**

  实现 L2 网络（包含 network、subnet、port）的资源自动分配和自动管理。

  在 H 版之前，Neutron 存在多个独立的 Core-plugin，每个 Core-plug 都是基于一种网络设备实现 L2 资源的分配、L2 数据库管理，绝大部分代码都是重复的。但是 Neutron 只能挂载一个 Core-plugin，因此在实际部署中，只能支持一种网络设备（比如：使用了 OVS 就不能使用 Linux bridge）。

H 版之后，引入 ML2 作为 Core-plugin，ML2 支持多驱动机制（Mechanism Driver），多种网络设备可以分别实现各自的 Driver 挂载到 ML2。ML2 实现 L2 资源分配、L2 数据库管理等 Core-plugin 的公共功能，Driver 实现和 Agent 或网络设备的交互。

- **Service-plugins**

  实现 L3-router、FWaaS、LBaaS 等扩展网络服务的资源自动分配和自动管理。

  Neutron 支持挂载多个 Service-plugin ,每个 Service-plugin 可实现一种或多种网络服务功能。

- **Agent**

  一种网络设备的核心功能实现或一种扩展网络服务功能实现。

  网络设备的核心功能 Agent，如：neutron-linuxbridge-agent、neutron-openvswitch-agent

  扩展网络服务功能 Agent，如：neutron-l3-agent、neutron-dhcp-agent、neutron-lbaas-agent

2.8.2 Neutron 对接模型

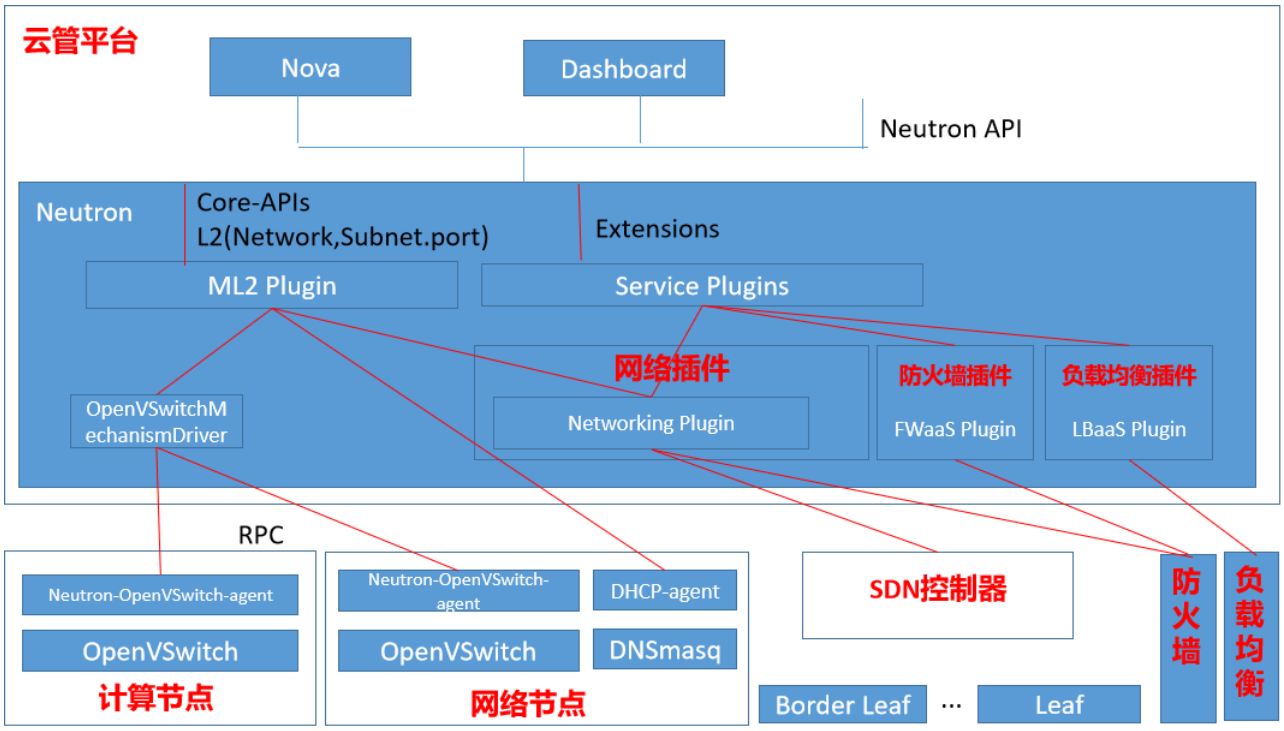


图 2-16 Neutron 对接示意图

- **Networking Plugin（网络插件）**
  - 既实现 ML2 的 L2 网络功能对接，又实现 Service Plugins 中的 L3 router 功能对接
  - 通过 REST api 和 SDN 控制器对接（类 Neutron API），实现业务网络在物理网元（本方案指的是 Border Leaf 和 Leaf 设备）上的自动化配置
  - 通过 REST api 和防火墙对接，实现虚拟防火墙在物理防火墙上的网络接口、SNAT、FloatingIP 配置
  - 实现 Neutron 和 SDN 控制器的数据一致性

- 支持 OpenStack L、M、N、O、P、Q、R

- **OpenVSwitch 插件**

  - 实现 ML2 的 L2 网络功能对接

  - 通过 RPC 和计算节点的 Agent 对接，实现业务网络在计算节点的 vSwitch 上的自动化配置

- **FWaaS 插件**

  - 实现 Service Plugins 中的 FWaaS 功能对接

  - 通过 REST api 和防火墙对接，实现虚拟防火墙在物理防火墙上的安全策略自动化配置

- **LBaaS 插件**

  - 实现 Service Plugins 中的 LBaaS 功能对接

  - 通过 REST api 和负载均衡对接，实现虚拟负载均衡在物理负载均衡上的网络及安全策略自动化配置

- **DHCP Agent**

  - 通过 RPC 和控制节点的 ML2 对接，实现业务网络的 DHCP 服务器自动化配置

2.8.3 业务网络自动化

Openstack 的业务网络模型和 SDN 控制器的业务网络模型类似，自动化过程也相似。

- **创建 Network**
  - Neutron 自动分配 VNI 和 VID
  - 通过 Networking Plugin 调用 SDN 控制器的“创建 Network”
  - 通过 OpenVSwitch plugin 调用计算节点 Agent，生成 VLAN
- **在 Network 下创建 Subnet**
  - Neutron 自动分配业务网关地址（或用户指定）
  - 通过 Networking Plugin 调用 SDN 控制器的“在 Network 下创建 Subnet”
- **创建 Router**
  - 通过 Networking Plugin 调用 SDN 控制器的“创建 Router”
- **在 Router 下添加 Interface**
  - 通过 Networking Plugin 调用 SDN 控制器的“在 Router 下添加 Interface”
- **创建 external network（启用 SNAT）/创建子网/更新 Router 的 Gateway 信息**

- Neutron 分配 VID
- 通过 Networking Plugin 调用 SDN 控制器的创建外部网络/子网，控制器记录相应的信息
- 更新 Router 的 Gateway 信息，通过 Networking Plugin 调用 SDN 控制器的更新 router 接口 ,SDN 控制器处理流程参考“创建过墙的出口网络”。(本功能是定制功能 ,创建 external network 时需要指定防火墙所连接的 Border Leaf 端口。后续会根据外部网络名称匹配控制器预配置的物理外部网络)
- 通过 Networking Plugin 调用防火墙的 API，实现防火墙的网络、NAT 配置

- **创建 external network (不启用 SNAT) /创建子网/更新 Router 的 Gateway 信息**
  - Neutron 分配 VID
  - 通过 Networking Plugin 调用 SDN 控制器的创建外部网络/子网，控制器记录相应的信息
  - 更新 Router 的 Gateway 信息，通过 Networking Plugin 调用 SDN 控制器的更新 router 接口，SDN 控制器处理流程参考“创建无墙的出口网络”。

- **创建主机/按需下发**
  - 创建主机时，Neutron 给主机分配端口
  - 通过 Networking Plugin 调用 SDN 控制器的创建 port，SDN 控制器处理流程参考“创建主机”。

- **创建虚拟防火墙/防火墙策略/防火墙规则**
  - 通过 FWaaS Plugin 调用防火墙实现相关功能
- **创建虚拟负载均衡/创建负载均衡池/创建负载均衡池成员**
  - 通过 LBaaS Plugin 调用负载均衡实现相关功能
- **创建/关联 Floating IP**
  - 通过 Networking Plugin 调用防火墙实现相关功能
- **创建安全组/安全规则**
  - Neutron 分别调用 Networking Plugin 和 OpenVSwitch 插件创建安全组/安全规则
  - 控制器接收到安全组/安全规则消息默认不处理
  - 计算节点的 OVS Agent 接收到安全组/规则后生成对应的流表，在服务器侧实现安全访问控制。不受外部宿主机组网和防火墙规则影响。

2.8.4 IRONIC 简介

IRONIC 是针对裸金属服务器出租场景提供的裸机服务，主要有上线、部署、回收三个阶段。

上线指的是服务器上架、连线完成后。管理员将服务器纳管到 IRONIC。该阶段，通过 IRONIC inspect 功能实现服务器硬件信息及接入位置（上联接入交换机信息）的采集。

部署指的是服务器出租后，根据租户业务部署租户指定的系统镜像和租户网络。该阶段，通过 IRONIC provision 功能实现租户系统和网络的自动化部署。

回收指的是服务器被租户释放后，清理租户的系统镜像和租户网络。该阶段，通过 IRONIC clean 功能清理服务器和网络数据。

IRONIC 部署流程如下：

1）Conductor 节点控制服务器关机

2）Conductor 节点通知 Neutron，为服务器申请 provision 网络的端口并更新 DHCP

3）Conductor 节点准备 PXE 启动环境，设置服务器的 PXE 启动顺序，控制服务器重启

4）服务器启动，通过 PXE 下载并导入 ramdisk，分配到 provision 网络的 IP

5）服务器和 Conductor 建立心跳连接，开始下载租户系统镜像

6）Conductor 节点定时查询下载是否完成

7）下载完成后，Conductor 节点修改服务器的 PXE 启动顺序，并重启服务器

8）Conductor 节点通知 Neutron，释放 provision 网络端口，绑定租户网络端口

9）服务器重启后进入租户系统及租户网络

2.8.5 IRONIC 对接

IRONIC Inspect 网络不受 neutron 管理，通过手工配置物理网络（服务器接入端口的 Native VLAN 需要配置为 Inspect 网络的 VLAN）。

IRONIC Provision 网络受 neutron 管理，可以看成一个专有 VPC，网络创建流程基本同上述的业务网络创建流程。区别的是，交换机对应的物理服务器的 M-LAG 组配置也需要自动化创建。Neutron 创建服务器对应的业务网络端口时，会通告服务器的绑定模式和对应的物理端口组，控制器根据绑定模式（bond4）和物理端口组在对应设备上生成 M-LAG 配置。

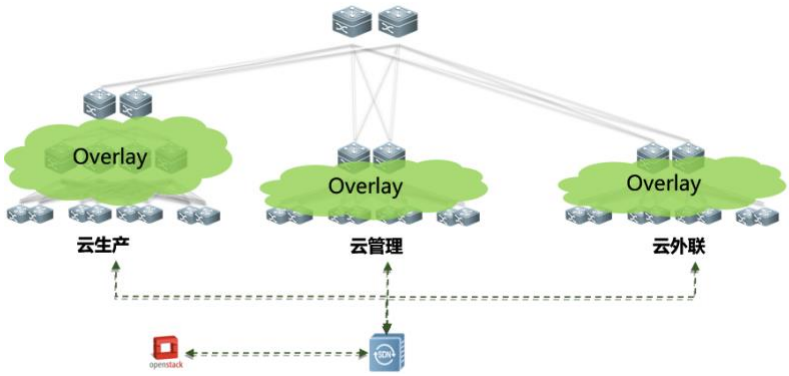
2.8.6 数据一致性对账

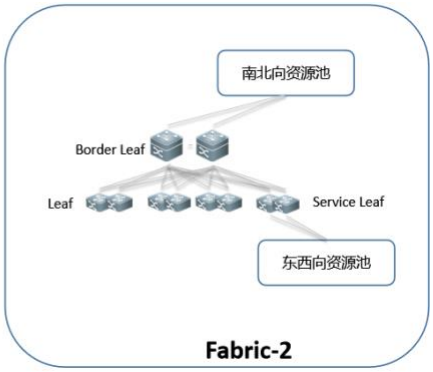
目前实现了两种数据一致性对账的方式：定时方式和手动触发方式。对账处理规则为：如果控制器比云平台上的数据少，则 Networking Plugin 把缺少的数据同步给控制器；如果控制器比云平台上的数据多，则不处理。

定时方式：Networking Plugin 定时（当前为 1 小时）查询 SDN 控制器的数据与云平台上的数据进行对账。

手动触发方式：SDN 控制器上有一个 Openstack 数据一致性对账手动触发开关，当开关开启后， 正常 30 秒内触发 Openstack 进行数据对账（Networking Plugin 定时 30 秒查询控制器一致性对账手动触发开关状态），当 Openstack 完成数据对账后会把手动触发开关关闭。

2.9 多 Fabric 管理





2.9.1 多 Fabric 组网

在 SDN 控制器上创建多个 Fabric，每个 Fabric 都是一个独立 SDN 资源池。每个 Fabric 可以有不同的 Border - Spine - Leaf 架构，可以有不同的南北向/东西向资源池架构。

SDN 控制器基于 Fabric 管理业务网络。业务网络不跨 Fabric。

2.9.2 云平台统一管理

云平台通过一套 SDN 控制器统一管理多 Fabric 时，由于标准 Neutron 插件没有字段标识业务网络和 Fabric 映射关系，需要在 SDN 控制器预置业务网络和 Fabric 映射表。

- 云平台业务网络和 Fabric 映射规则

1）逻辑资源类型+名称前缀映射

SDN 控制器根据逻辑资源名称和类型把逻辑资源映射到对应 Fabric。如：以”DMZ_”开头的逻辑 router 映射到云外联 Fabric；其它 router 映射到云生产 Fabric

2）VLAN/VXLAN 映射规则

SDN 控制器根据 network 的 vlan/vxlan 段映射到对应 Fabric。如 以 VLAN 1-100 的逻辑 network 映射到云外联 Fabric；其它 nework 映射到云生产 Fabric

3）外部出口网络关联

在 SDN 控制器指定 Fabric 创建外部出口网络；云平台创建的外部 network 和外部出口网络的名称一样，则外部 network 关联到该 Fabric。

4）依赖关联

当一个逻辑元素 A 关联到一个 Fabric 时，另一个逻辑元素 B 和 A 产生关系，则 B 也关联到 A 所在 Fabric。如：SDN 控制器在云外联 Fabric 创建一个外部出口网络，云平台创建的外部 network 和外部出口网络名称一样，则外部 network 关联到云外联 Fabric。云平台创建一个 router 关联到外部网络，则这个 router 也关联到云外联 Fabric

- 典型业务流程

在 SDN 配置云外联 Fabric 出口网络 ,名称为 wan_net ;在 SDN 配置 network 的 VLAN 映射 Fabric 规则。

在云平台创建 extern network，名称为 wan_net；SDN 自动对应到云外联 Fabric，在云外联 Fabric 下发配置

在云平台创建 router，关联 wan_net 的 extern network，SDN 自动对应到云外联 Fabric，在云外联 Fabric 下发配置

在云平台创建 network，根据 network 的 VLAN ID，SDN 找到对应的 Fabric，下发配置

2.10 网络运维

2.10.1 物理网络及资源可视

- **物理拓扑**
  - 物理设备通过 LLDP 发现物理链路
  - SDN 控制器通过 SNMP 读取物理设备的 LLDP 信息，计算出整网拓扑

- **设备状态监控及告警**
  - SDN 控制器通过 Netconf 定时读取设备的 CPU、内存、硬盘使用情况，以及 ARP 表项、MAC 表项、VRF 表项、路由表项（静态路由、动态路由）、VNI 表项的容量与使用情况，可视化的展示设备的当前运行状态及历史状态

- SDN 控制器自动对超过阈值的设备进行状态异常告警

- **流量监控及告警**

  - SDN 控制器通过 Netconf/Openflow 监控设备、端口流量，可视化的展示设备的当前流量信息及历史流量信息

  - SDN 控制器自动对超过阈值的端口进行流量异常告警

2.10.2 逻辑网络及资源可视

- Overlay 虚拟拓扑

  - VXLAN 设备通过 VXLAN 隧道建立虚拟链路

  - SDN 控制器通过 Netconf 读取 VXLAN 设备的虚拟链路信息，计算出整网 Overlay 虚拟拓扑

- 租户/业务网络逻辑拓扑

  - SDN 控制器根据租户/业务网络中各网络元素间的关系，计算出每个租户/业务网络的逻辑拓扑

- 网络资源监控

  - SDN 控制器基于租户/业务网络对整网的逻辑网络资源进行管理和统计

2.10.3 物理拓扑与逻辑网络互视

- 物理设备映射逻辑网络
  - 租户/业务网络与物理网络存在映射关系（详见 2.6 节“逻辑/物理网络映射”部分）
  - SDN 控制器根据租户/业务网络与物理网络的映射关系，计算出物理设备所映射的逻辑网络拓扑
- 逻辑网络映射物理设备
  - 租户/业务网络与物理网络存在映射关系（详见 2.6 节“逻辑/物理网络映射”部分）
  - SDN 控制器根据租户/业务网络与物理网络的映射关系，计算出某个租户/业务网络所映射的物理网络拓扑

2.10.4 流量镜像

流量镜像包括本地镜像、远端镜像。

- 本地镜像
  - 本地镜像的被镜像端口和镜像端口必须在同一台设备上，基于 SPAN 实现。SPAN 会话是镜像源端口与目的端口之间的数据流，可以监控单个或多个端口的输入、输出、双向的报文。

端口加入 SPAN 会话后并不影响交换机的正常运行

- SDN 控制器按照指定设备 IP、源端口、目的端口配置镜像会话，报文从源端口被复制到本设备的目的端口

- 远程镜像

  - 远程镜像基于 ERSPAN（Encapsulated Remote Switch Port Analyzer，封装的远程镜像）实现，镜像报文可跨三层传输。被镜像报文通过 GRE（Generic Routing Encapsulation）隧道封装成 IP 报文，并通过以太网转发到任何 IP 路由可达的地方，目的 IP 对应的采集服务器物理位置不受限制

  - SDN 控制器按照指定的源设备 IP、源端口、目的设备 IP 配置镜像会话，报文从源设备-源端口被复制到远端的目的 IP

2.10.5 泛洪抑制

当局域网中存在过量的广播、多播或未知名单播数据流时，就会导致网络变慢和报文传输超时机率大大增加。拓朴协议的执行错误或网络配置错误都有可能导致这种局域网风暴的产生。

针对广播、多播和未知名单播数据流进行泛洪抑制后，当设备端口接收到的广播、多播或未知名单播数据流的速率超过所设定的带宽、每秒允许通过的报文数或者每秒允许通过的千比特数时，设备将只允许通过所设定带宽、每秒允许通过的报文数或者每秒允许通过的千比特数的数据流，超出限定范围部分的数据流将被丢弃，直到数据流恢复正常，从而避免过量的泛洪数据流进入局域网中形成风暴。

2.10.6 VTEP 间路径探测

VXLAN 对业务报文进行了一层封装，无法采用传统方式(如 trace)直观的判断 VTEP 间设备的通路。SDN 控制器通过 Openflow 实现了自动化的路径探测。

1）在 SDN 控制器选择路径探测的源和目的 VTEP，并指定探测流的 DSCP

2）SDN 控制器通知整网设备将具备探测 DSCP 特征的报文上送控制器

3）SDN 控制器通过 Openflow 在源 VTEP 节点下发经过 DSCP 染色的探测流量（以源 VTEP 的隧道 IP 为源 IP，以目的 VTEP 的隧道 IP 为目的 IP，目的端口为 VXLAN 端口的 UDP 报文）

4）物理网络中所有收到该染色报文的设备都将报文上送 SDN 控制器

5）SDN 控制器根据收到报文的设备端口及物理网络拓扑，计算出 VTEP 间路径

6）如 VTEP 间存在多条路径，SDN 控制器递增变换模拟报文的源端口，循环探测。（VTEP 设备的通过报文源端口进行 VXLAN 负载均衡）

2.10.7 设备自动发现

设备自动发现基于 SNMP 协议实现，SDN 控制器遍历某个网段的 IP，将此 IP 作为目标 IP 尝试 SNMP 连接是否正常，若连接成功则表明成功发现一台设备。

2.10.8 主机路径探测

主机路径探测主要是针对主机业务进行路径探测，更精细化的检查 VTEP 间某类业务流量的路径，原理和 VTEP 间路径探测类似

1）在 SDN 控制器指定探测的业务五元组（源 IP、目的 IP、协议、源端口、目的端口）以及源 VTEP，并指定探测流的 DSCP 和业务接入 VLAN（存在重叠地址的网络，需要通过接入 VLAN 确定业务主机）。

2）SDN 控制器通知整网设备将 DSCP 染色报文上送控制器

3）SDN 控制器通过 Openflow 给源 VTEP 节点下发经 DSCP 染色的模拟主机业务流量(业务的源 IP、目的 IP、协议类型、源端口号、目的端口号），VTEP 节点对模拟业务流量进行 VXLAN 封装并转发（封装时会把内层的 DSCP 复制到外层）

4）物理网络中所有收到该染色报文的设备都将报文上送 SDN 控制器

SDN 控制器根据收到报文的设备端口及物理网络拓扑，计算出主机业务路径

2.10.9 主机连通性检测

主机连通性检测包括二层连通性检测（MAC Ping）、三层连通性检测（IP Ping）。

- 二层连通性检测
  - 模拟主机 ARP 学习机制来检测主机间二层连通性

- MAC Ping 的源和目的主机必须属于同一个二层网络

- 三层连通性检测

  - 模拟 Ping 机制来检测主机间三层连通性

二、三层连通检测均基于 Openflow 流表实现：

1）SDN 控制器通过 Netconf 下发报文匹配规则（针对源主机的 ARP 应答报文或 ICMP 应答报文）到设备，控制设备将应答报文上送控制器

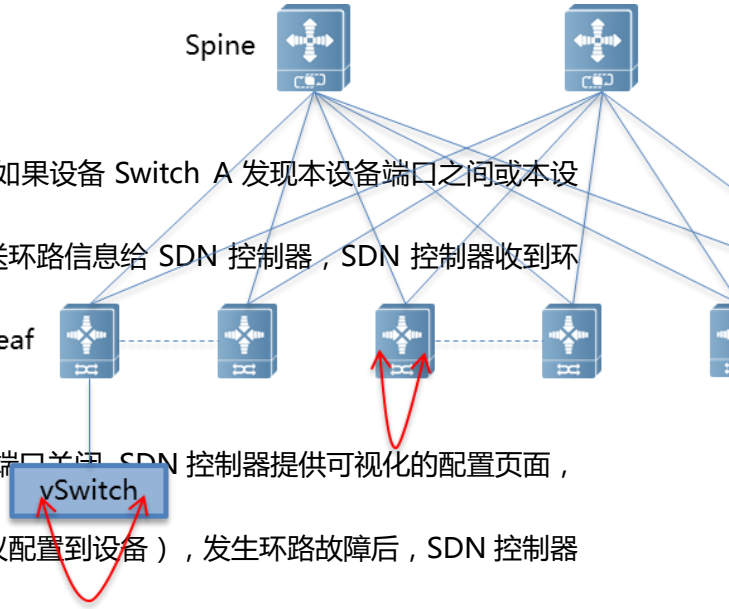
2）SDN 控制器通过 Openflow 在源节点下发探测报文(模拟源主机的 ARP 请求或 ICMP 请求）

3）目的主机收到该请求报文后，发出应答报文

SDN 控制器根据收到应答报文的情况来判断网络是否连通、是否健康

2.10.10 网络环路自动检测

网络设备基于 RLDP 协议自动发现网络环路。RLDP（Rapid Link Detection Protocol）协议是锐捷网络自主开发的一个用于快速检测以太网链路故障的链路协议。RLDP 定义了两种协议报文：探测报文（Probe）和探测响应报文（Echo）。RLDP 会在每个配置了 RLDP 并且是 linkup 的端口周期性地发送本端口的 Probe 报文，并期待邻居端口响应该探测报文，同时也期待邻居端口也发送自己的 Probe 报文。如果一条链路在物理和逻辑上都是正确的，那么一个端口应该能收到邻居端口的探测响应报文以及邻居端口的探测报文；否则链路将被认定是异常的。



SDN 控制器通过 Netconf 协议监听设备链路状态，如果设备 Switch A 发现本设备端口之间或本设备与其他设备之间发生了环路，则通过 Netconf 协议上送环路信息给 SDN 控制器，SDN 控制器收到环路信息后，记录环路信息并发布环路告警提醒。

发生网络环路故障后，可通过 shutdown 策略将环路端口关闭。SDN 控制器提供可视化的配置页面，通过页面配置 shutdown 策略并使能（通过 Netconf 协议配置到设备），发生环路故障后，SDN 控制器将自动将 shutdown 策略实施到环路端口

图 2-17 网络环路

2.10.11 交换机远程升级

交换机远程升级基于 TFTP 技术实现。SDN 控制器支持可视化的升级包管理，并通过 TFTP 将升级包发送到设备端，设备端以此升级包进行版本升级。

对于 M-LAG 架构部署的设备，M-LAG 成员对外表现为独立的设备，即如果是 2 台设备组成 M-LAG，则升级时按照 2 台设备分别升级。

对于 VSU 架构部署的设备，VSU 所有成员对外只表现为同一台设备， VSU 系统通常要求成员设备主程序版本号一致，线下通过手动方式单机逐一升级效率太低，且容易出错，锐捷交换机系统升级时支持 VSU 成员间自动进行版本同步：

- VSU 系统建立时：系统会自动匹配所有成员设备的主程序版本号，当发现主程序版本不一致时，会自动将主设备上的主程序版本同步到所有成员设备

- VSU 系统建立后：通过 TFTP 下载文件时自动同步给所有成员设备，即 VSU 成员设备同步升级

2.10.12 交换机重启

交换机发生故障或 CPU、内存使用率过高时，可能需要对交换机设备进行重启操作，传统的重启方式一般是通过 Telnet、SSH 登陆到交换机设备，使用命令行进行重启。SDN 控制器将基于 Telnet 协议的设备重启命令行封装成可视化按钮，实现单台、批量交换机设备一键重启。

2.10.13 Underlay 零配置上线

Underlay 零配置上线基于设备 ZAM(Zero-configure Automatic Manage ,零配置自动上线管理）机制实现，ZAM 是利用 Python、Go 等编程语言进行逻辑处理，实现网络设备上电后自动加载、上线功能的技术。Underlay 零配置上线流程如下：

1）在 SDN 控制器上配置设备的管理地址段、VTEP 地址段、互联地址段

2）在 SDN 控制器上配置 Underlay 部署参数并导入设备拓扑连接信息

- Underlay 的路由协议（OSPF、BGP 等路由参数配置）
- 设备接口信息及互联信息（设备上下联的邻居、上下联端口、M-LAG 聚合关系、AP 聚合关系）
- 设备角色信息（Border Leaf、Service Leaf、Leaf、Spine 等）

3）SDN 控制器给每台设备自动分配管理 IP、VTEP IP、互联 IP，自动生成路由协议配置和接口

配置。把管理 IP 和设备的映射作为租约写入内置的 DHCP 服务器，把设备开机脚本和自动生成的设备起始配置文件放在内置的 TFTP 服务器，把 TFTP 服务器地址和设备开机脚本名称作为 DHCP 选项设置到 DHCP 服务器。

4）设备空配置启动，自动运行 DHCP Client，从 DHCP 服务器获取到 SDN 控制器分配的管理 IP，并获得 TFTP 服务器地址和开机脚本名称。

5）设备获取开机脚本并运行。在脚本运行中，设备开启 LLDP、SNMP、SSH 服务。

6）SDN 控制器通过 SNMP、LLDP 获取网络拓扑，并校验设备连接关系是否和导入设备邻居关系匹配，匹配则通过 ssh 控制设备获取控制器生成的配置文件并重启设备。

7）设备重启后运行获取到的配置文件，完成 Underlay 组网。

2.10.14 交换机故障零替换

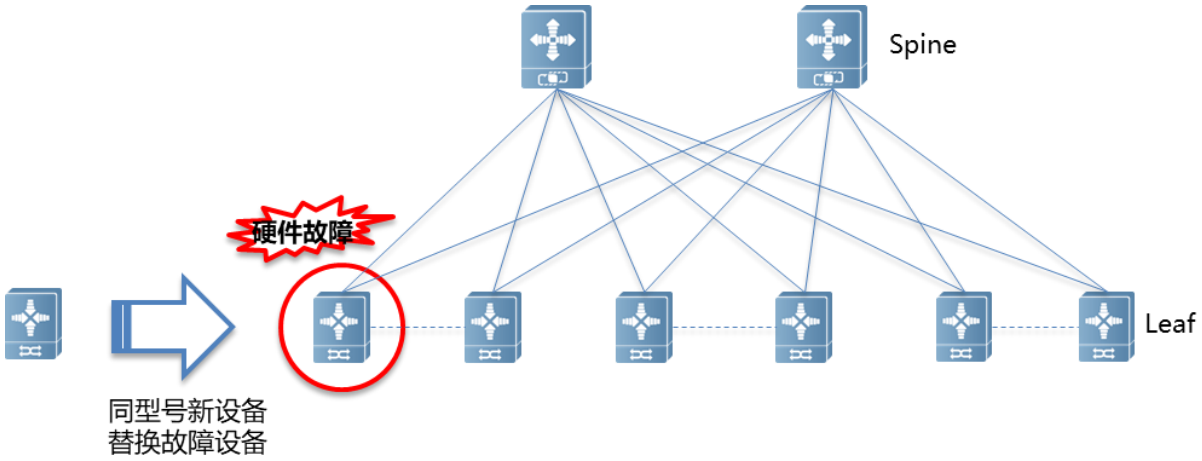


图 2-18 设备故障零替换

设备替换时，SDN 控制器首先要识别出发生了替换事件，其次还要标识出故障设备与新入网设备。SDN 控制器本地保存一份拓扑信息，通过拓扑关系标识设备位置，通过设备 IP 标识某一台设备，结合这两个信息就可识别出发生了设备替换事件。SDN 控制器引入临时 IP 的机制，新设备在故障设备的端口以临时 IP 入网，以标识发生了替换事件；新设备上线后，SDN 控制器通过对比新、旧拓扑信息识别替换后与替换前的设备 IP。确定上述两个要素后，SDN 控制器将本地备份的故障设备配置同步到新设备，同步完成后再将新设备以故障设备 IP 上线，释放临时 IP，拓扑信息恢复到故障前的状态，设备替换完成。临时 IP 由 SDN 控制器自动分配，用户无需感知。

由于替换后与替换前的设备需要下发相同的配置，为保证替换成功，设备替换除了满足设备零配置入网的条件外，还需要保证替换的新设备与旧设备型号一致。

设备运行过程中会有各种突发故障，设备零替换上线后需要将旧设备的所有配置同步到新设备，此时就需要定期或按照计划进行备份。SDN 控制器提供基于设备的自动备份计划，备份计划包括执行周期、设备列表，每到执行时间，SDN 控制器触发这些设备的配置备份，设备配置备份文件存储于 SDN 控制器内置的 TFTP 服务器。

若设备配置已备份，在设备配置出现异常时（如误删、误配置），可在 Web 页面手动选定设备后恢复到该设备最新一次备份的配置状态。SDN 控制器提供设备配置恢复的历史记录，包括配置恢复时间、恢复的配置文件名、恢复执行结果等，并允许通过 Web 页面查看。

2.11 高可用性

2.11.1 控制器集群

由 2*N+1 台控制器组成分布式集群，提供高可用的服务。当小于一半（不超过 N+1）的控制器异常时，剩余的控制器正常提供服务。当超过一半的控制器异常时，所有控制器都不提供服务。这种方式相比于传统的主备方式，有效的防止了控制器间失联引起的脑裂问题。

控制器集群自身提供虚 IP 服务，集群内的主控制器持有虚 IP。当主控制器异常时，虚 IP 也自动切换到新选出的主控制器。在同一个二层网络内部署集群，网关通过主控制器发出的 ARP 更新虚 IP 的位置。在不同二层网络内部署集群时，控制器通过内置的 Qugga 组件发布虚 IP 的路由。

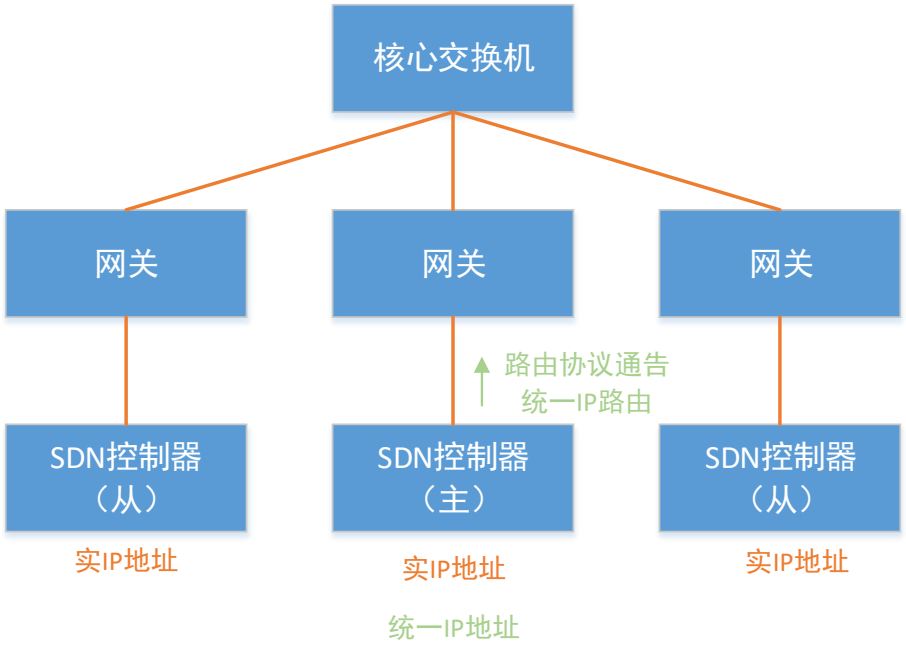


图 2-19 SDN 控制器跨二层网络部署

2.11.2 业务负载均衡（北向）

多台控制器组成集群时，所有控制器都能独立提供服务，可以通过外置负载均衡把业务服务均衡调度到不同的控制器。对于虚 IP 场景，控制器通过内置的 LVS 提供负载均衡。

2.11.3 南向负载均衡

SDN 控制器为集群部署 ,设备一般会同时连接到 SDN 控制器集群的多个成员 ,设备连接包括 Netconf 协议连接、Openflow 协议连接、SNMP 协议连接等，不同类型的协议连接之间互不影响，所以这里讨论的负载均衡是针对某种协议类型而言的。上述某种协议类型的多个连接对于设备而言是主备关系，主连接负责 SDN 控制器与设备的信息交互，备连接作为冗余备份。对于设备协议连接数量而言，每个集群成员连接的设备数量应尽可能均衡分布；对于连接的主备关系而言，每个集群成员对应的设备主连接应该尽可能均衡分布。即南向负载均衡包括连接负载均衡、主备负载均衡两个概念，并且主备负载均衡优先级高于连接负载均衡。

- SDN 控制器集群包括 N 个成员（假设 N=3），SDN 控制器监听集群成员状态及设备的协议连接状态。如果设备与控制器集群存在 N 个连接，当集群成员发生异常时，SDN 控制器将该集群成员的设备协议连接自动均衡到剩下的（N-1）个健康成员节点，并保证这（N-1）个控制器集群成员之间主备均衡、连接数均衡。此时该异常成员不再接受新的设备协议连接；当设备协议连接从异常恢复后，SDN 控制器检查集群成员已有连接数，并自动将恢复后的设备协议连接分配到当前连接数最少的集群成员节点。

- 新增设备时，SDN 控制器检查集群成员已有连接情况，并自动将新增的设备主连接分配到当前主

连接最少的集群成员节点，自动将新增的备连接分配到连接数最少的集群成员节点。

3 极简云数据中心技术特点

极简云数据中心本阶段结合物理网络设备、防火墙、负载均衡、SDN 控制器、对接云平台提供完整的数据中心网络解决方案，实现计算、存储、网路的统一管理和业务网络的自动化。

4 结束语

极简云数据中心解决方案，通过 VXLAN 实现业务弹性扩展、灵活迁移、解决业务规模受限问题，通过 SDN 控制器和云平台对接实现计算/存储/网络统一管理，通过 SDN 控制器的业务网络自动化实现业务快速上线/变更，通过可视化/自动化使运维更加简单。