HCRSE116-BGP EVPN 原理

EVPN (Ethernet VPN) 以太网 VPN
VPLS（Virtual Private Lan Service）虚拟专用局域网业务
HIS　(High Speed Internet) 高速互联网
BTV　(Broadband TV) 宽带电视
BRAS（Broadband Remote Access Server）宽带接入服务器
VoIP（Voice over Internet Protocol ）基于 IP 的语音传输
Multihoming 多宿主
NVE(Network Virtrualization Edge 网络虚拟边缘节点）
VNI（VXLAN Network Identifier）：VXLAN 网络标识
BUM（broadcast&unknown-unicast&multicast）广播&未知单播&组播
DF(Designated Forwarder)指定转发者

ES(Ethernet Segment，以太网段)：如果一个主机通过多条链路同时接入不同的 VTEP 设备，那么这多条链路就叫作 ES。
ESI(Ethernet Segment Identifier，以太网段标识符)：用来标识一个 ES 的值叫做 ESI。

EVPN 基本原理

基于 BGP 和 MPLS 的 L2 VPN，用于实现网络二层互通的 vpn 技术。通过扩展 BGP 协议的 NLRI，新增了几种类型的 BGP EVPN 路由类型，用于在不同站点之间通告主机的 mac 地址和 ip 地址信息。

EVPN（Ethernet Virtual Private Network）是一种用于二层网络互联的 VPN 技术。EVPN 技术采用类似于 BGP/MPLS IP

VPN 的机制，在 BGP 协议的基础上定义了一种新的 NLRI（Network Layer Reachability Information，网络层可达信息）即 EVPN NLRI，EVPN NLRI 定义了几种新的 BGP EVPN 路由类型，用于处在二层网络的不同站点之间的 MAC 地址学习和发布。原有的 VXLAN 实现方案没有控制平面，是通过数据平面的流量泛洪进行 VTEP 发现和主机信息（包括 IP 地址、MAC 地址、VNI、网关 VTEP IP 地址）学习的，这种方式导致数据中心网络存在很多泛洪流量。为了解决这一问题，VXLAN 引入了 EVPN 作为控制平面，通过在 VTEP 之间交换 BGP EVPN 路由实现 VTEP 的自动发现、主机信息相互通告等特性，从而避免了不必要的数据流量泛洪。
综上所述，EVPN 通过扩展 BGP 协议新定义了几种 BGP EVPN 路由，这些 BGP EVPN 路由可以用于传递 VTEP 地址和主机信息，因此 EVPN 应用于 VXLAN 网络中，可以使 VTEP 发现和主机信息学习从数据平面转移到控制平面。

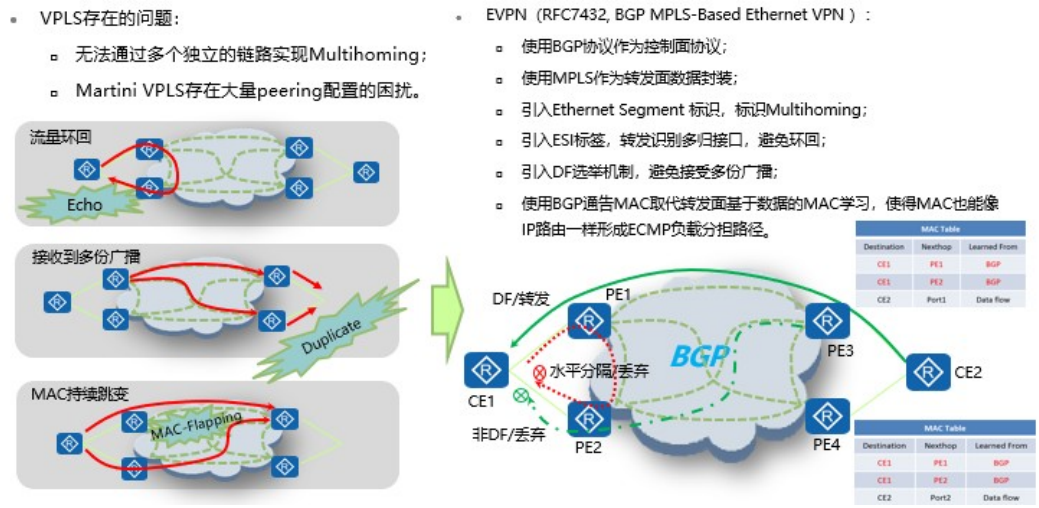
前言

- EVPN (Ethernet VPN) 被用于NVO (Network Virtualization Overlay)解决方案。EVPN可以结合VXLAN技术实现在IP层之上的隧道封装，作为路由的控制平面。

Ethenet L2VPN技术

- 现网大量采用Ethernet L2VPN服务，例如VoIP、HSI和BTV等业务网关部署在BRAS，企业机构之间通过城域网构建二层网络，数据中心互联。
- VPLS是现网广泛应用的技术，能够提供二层报文的透传，但依然存在很多限制和新的需求。
  - Multihoming
  - 组播优化
  - 配置复杂度高
  - 多租户的DCI互联
  - 快速收敛
  - 广播抑制
- EVPN被提出来解决这些问题。

- HIS(High Speed Internet)
- BTV(Broadband TV)
- Multihoming：当前 VPLS 只能支持 multihoming 的 single-active 冗余不是，不支持多路径多活转发。
- 组播优化：组播的 LSPs 可以结合 VPLS，但是只能用于 P2MP 的 LSPs。对于 MP2MP 的 LSPs 使用场景，VPLS 无法支持。
- 配置复杂度高：当前 VPLS 提供基于 BGP 的 auto-discovery 的 single-sided 接入，但是需要工程师在接入侧以太配置上再配置大量的网络参数。
- 多租户的 DCI 互联：DCI 链路之间不仅传统数据中心间二层，也需要扩展租户的的二层网络。



- 我们提到了 VXLAN 的不足之处需要引入新的控制面协议，那么这里我们先看看 EVPN 协议，EVPN 协议全称是 Ethernet VPN，RFC7432 中有定义，是用来解决 VPLS 的一些现存问题，比如无法通过多个独立的链路实现 multihoming。某些情况可能接收到多份广播报文、或者 MAC 持续漂移。而且 Martini VPLS 也存在大量 peering 配置的困扰。
- 那么 RFC7432 中定义的 EVPN 协议，则通过使用 BGP 协议作为控制面协议，MPLS 作为转发面数据封装，通过引入一些新的内容，来解决 VPLS 场景下比如容易产生环路、多份广播报文以及 MAC 地址学习的问题。

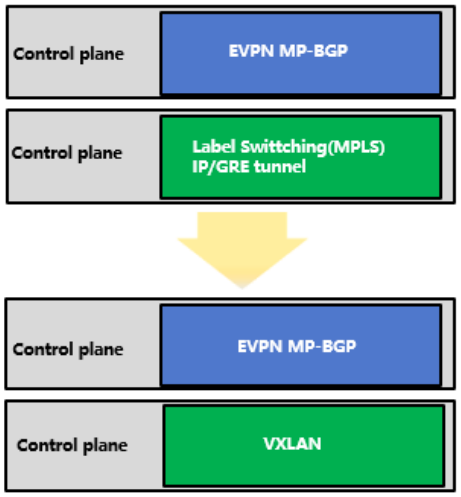
BGP-EVPN发展和应用

EVPN的发展

- Ethenet VPN 最初是在RFC-7432中定义。基于MPLS Based的VPN网络中满足高带宽、复杂QoS等需求而演进的，控制平面采用MP-BGP 定义了地址族。
- EVPN主要特点：控制平面与数据平面被抽象并隔离；MP-BGP控制平面承载了MAC/IP路由信息；数据平面的封装有若干种选择。

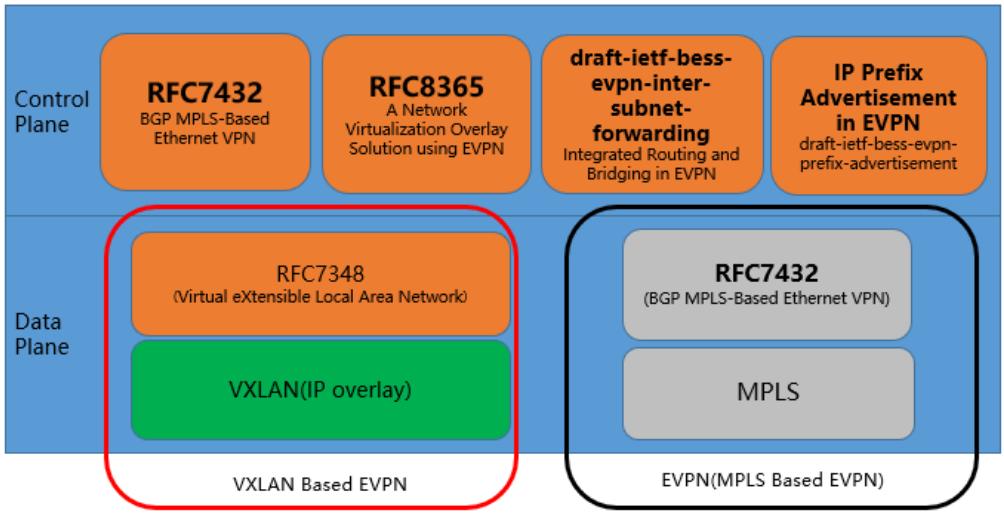
EVPN for VXLAN

- VXLAN没有控制面，而EVPN中的MP-BGP控制平面非常适合VXLAN。
- VXLAN对标准的EVPN 进行扩展。
- 如下三个草案：
  - draft-ietf-bess-evpn-overlay
  - draft-ietf-bess-evpn-inter-subnet-forwarding
  - draft-ietf-bess-evpn-prefix-advertisement



- EVPN（Ethernet Virtual Private Network）其实就是一种用于二层网络互联的 VPN 技术。EVPN 技术采用类似于 BGP/MPLS IP VPN 的机制，在 BGP 协议的基础上定义了一种新的 NLRI（Network Layer Reachability Information，网络层可达信息）即 EVPN NLRI，EVPN NLRI 定义了几种新的 BGP EVPN 路由类型，用于处在二层网络的不同站点之间的 MAC 地址学习和发布。
- 同时，VXLAN 实现方案没有控制平面，是通过数据平面的流量泛洪进行 VTEP 发现和主机信息（包括 IP 地址、MAC 地址、VNI、网关 VTEP IP 地址）学习的。这种方式导致数据中心网络存在很多泛洪流量。为了解决这一问题，VXLAN 引入了 EVPN 作为控制平面，通过在 VTEP 之间交换 BGP EVPN 路由实现 VTEP 的自动发现、主机信息相互通告等特性，从而避免了不必要的数据流量泛洪。
- 除了 RFC7432 以外，之前还有 3 个相关的草案，其中第一个 draft-ietf-bess-evpn-overlay 目前已经成为正式的 RFC，A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)RFC 8365，另外两个也正在努力成为标准的过程中。

EVPN的实现



- 使用 VXLAN 作为数据平面。

EVPN NLRI

- RFC 7432 EVPN定义了新的NLRI (Network Layer Reachability Information)，叫做EVPN NLRI，格式如下：

```
+---------------------------------------+
|      Route Type (1 octet)             |
+---------------------------------------+
|      Length (1 octet)                 |
+---------------------------------------+
| Route Type specific (variable)        |
+---------------------------------------+
```

- 根据路由类型字段，RFC7432中定义了4种路由类型：
  - 1 - Ethernet Auto-Discovery (A-D) route
  - 2 - MAC/IP advertisement route
  - 3 - Inclusive Multicast Ethernet Tag Route
  - 4 - Ethernet Segment Route

Type1：Ethernet Auto-Discovery Route

- EVPN NLRI格式：

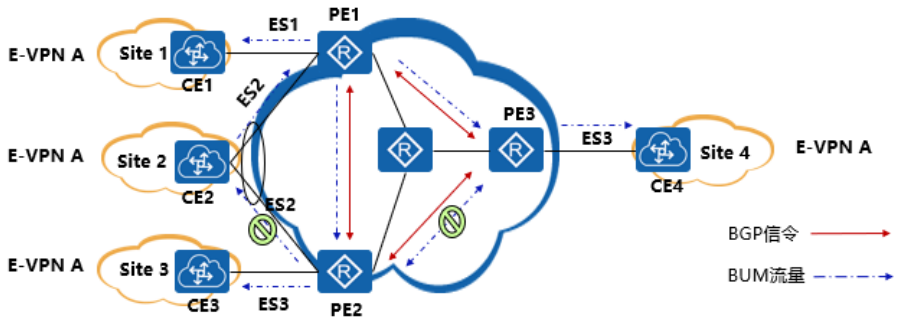
```
+---------------------------------------+
|  Route Distinguisher (RD) (8 octets)  |
+---------------------------------------+
| Ethernet Segment Identifier (10 octets)|
+---------------------------------------+
|      Ethernet Tag ID (4 octets)       |
+---------------------------------------+
|      MPLS Label (3 octets)            |
+---------------------------------------+
```

- 又称为Ethernet Auto-Discovery (A-D) route；
- 仅仅在通过ESI实现Multihoming接入时才需要。
- 用以实现：
  - 水平分隔
  - 快速收敛
  - 别名
- 前缀索引：
  - Ethernet Segment Identifier
  - Ethernet Tag ID

- 水平分隔（分发 ESI 标签）；
- 快速收敛（其他 PE 根据 RT1 路由实现端口下 MAC 等明细路由的批量快速切换）；
- 别名（任意的多归 PE 发布 MAC 等明细路由，其他 PE 可根据 RT1 路由形成到所有多归 PE 的 ECMP）；
- 使用 M-LAG 和堆叠技术实现 Multihoming 可以替代此类路由。

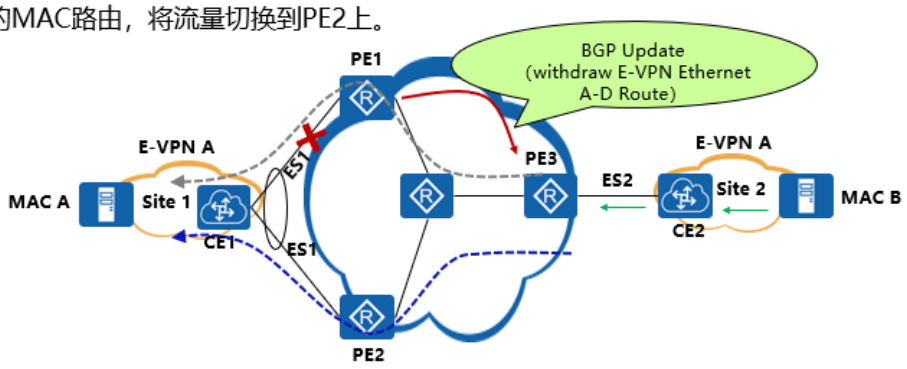
水平分割

- 如图， CE2双归到PE1和PE2，PE1收到来自CE2的BUM流量会复制转发给PE2，PE2收到报文后不应该将报文再发给CE2，避免在CE侧形成环路；同样，PE2从PE1收到的多播流量不会向PE侧转发，避免在公网侧形成环路。
- 水平分割是通过per ES AD路由中携带的ESI标签实现的；例如下图中，PE2会分配1个ESI标签，来标识ES2，并通过AD路由发布给PE1，PE1向PE2发BUM流量时需要打上这个标签，PE2收到后识别标签，即不会向ES2转发。



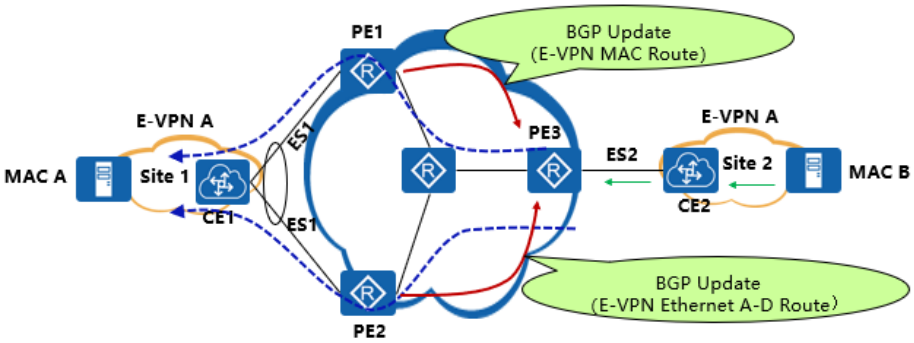
快速收敛

- 快速收敛：是指通过per ES AD路由，快速撤销MAC路由，达到快速收敛的目的。
- 如图， CE1双归到PE1和PE2，PE1和PE2学习到CE1的MAC，向PE3发布MAC路由；当CE1与PE1之间的链路故障时，PE1向PE3发送per ES A-D 路由撤销，可使PE3快速更新ES1的MAC路由，将流量切换到PE2上。



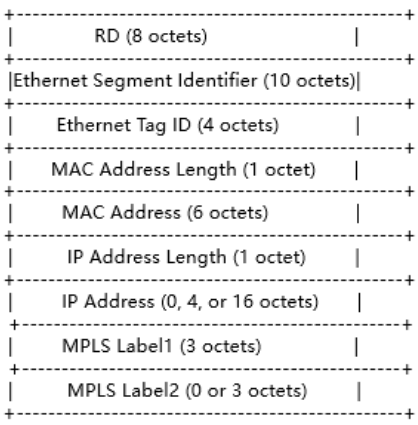
别名

- CE多归多活场景时，可能存在多归的PE中有PE没有学习到CE的MAC地址的情况，导致远端PE不能形成负载分担或备份；别名就是为了解决这个问题，别名通过per EVI AD路由实现；
- 如图，CE1双归到PE1和PE2，假设PE1学习到MAC A，向PE3发布MAC路由，PE2没有学习到MAC A，但是PE2可以向PE3发布per EVI AD路由，PE3就可以知道MAC A通过PE1和PE2均可达。



Type2：MAC/IP Advertisement Route

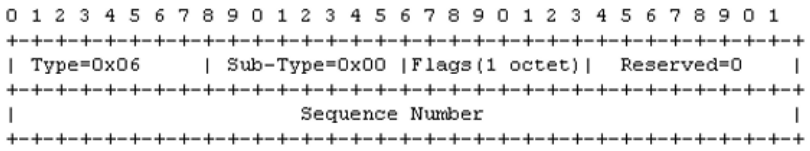
- MAC路由EVPN NLRI格式：



- 作用：
  - 主要用于指导单播流量转发；
  - 转发虚拟机MAC/IP地址。
- 前缀索引：
  - Ethernet Tag ID、MAC Address、IP Address作为前缀索引，ESI和MPLS Label作为路由属性。

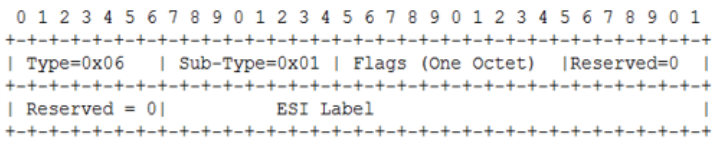
MAC迁移扩展团体属性

- MAC迁移扩展团体属性：
  - 是可传递扩展团体属性，类型是06，子类型是00，在MAC路由中携带，用于MAC迁移；
  - Flag：最低位，1表示是静态MAC不能迁移。
- MAC迁移扩展团体属性格式：



Ethernet Segment Identifier (ESI)

- ESI标签扩展团体属性格式：



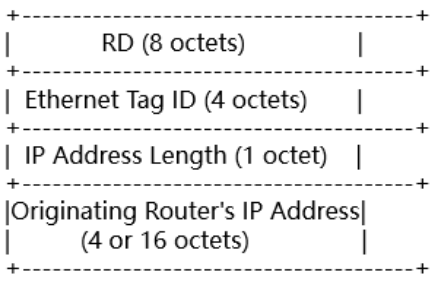
- ESI Label扩展团体属性：
  - 是可传递扩展团体属性，类型是06，子类型是01，在AD路由中携带；
  - Flags：最低位，0表示All-Active（多活模式），1表示Single-Active（单活模式）；
  - 标签值：3个字节的ESI标签，用作水平分割，对于Single-Active，标签值为0。

冗余模式

- 单活模式 (Single-Active Redundancy Mode)
  - 1个PE如果它收到的1组per ES AD路由中有1个通告的是单活模式，则对于该ES，按照单活模式处理；这时流量只能走主PE；
  - 如果主PE发生故障，它可以先撤销per ES AD路由；远端PE设备再切换到该组PE中的其他PE。
- 多活模式 (All-Active Redundancy Mode)
  - 1个PE如果它收到的1组per ES AD路由中全部通告的是多活模式，则对于该ES，按照多活模式处理；这时流量可以通过每个PE转发。

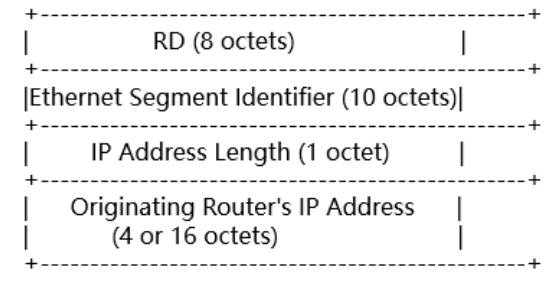
Type3：Inclusive Multicast Ethernet Tag Route

- 称为Inclusive Multicast Ethernet Tag route；
- 用于隧道自动建立、VNI广播成员的自动加入。



Type4: Ethernet Segment Route

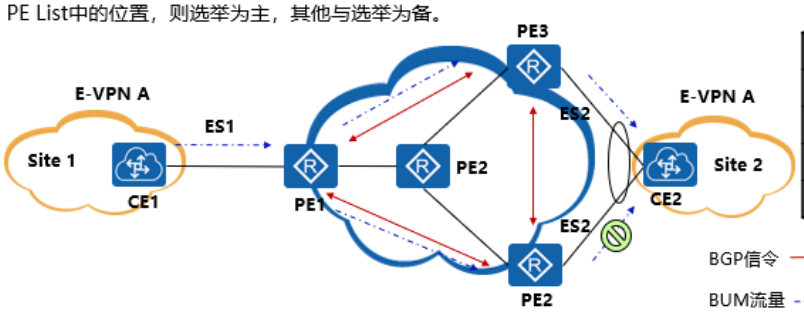
- NLRI格式如下：



- 在Multihoming才需要，用于Multihoming的链路间进行DF (Designated Forwarder)选举。

DF选举

- DF选举：当CE多归到多个PE时，只需要有1个PE向CE转发BUM流量，选出这个PE的过程就是DF选举；
- 如图，CE2双归到PE2和PE3，由CE1发出的多播报文会发送到PE2和PE3，在PE2和PE3之间进行DF选举，如果PE3为主，PE2为备，则PE2不会向CE2转发多播流量了。
- DF选举通过ES路由实现，当DF选举定时器（默认3s）超时后，各PE根据EVI的ES路由进行选举，选举算法如下：把相同ES按照Originator IP地址按升序排序，生成PE List，然后按照VLAN对PE数量进行取模，模值等于该PE在PE List中的位置，则选举为主，其他与选举为备。



VLAN	PE2	PE3
10	DF	
11		DF
12	DF	
13		DF

Type5路由的产生

- BGP EVPN最早定义了四种类型路由，提供了灵活的控制平面，实现同一网段能够跨三层网络的目标。在某些情况下，跨租户的系统需要直接接入网络，而不需要进行协议的交互。
- EVPN新定义了Type5路由来实现IP前缀的通告。
- IP前缀路由用于解决不同子网之间的连接和主机访问外部网络。

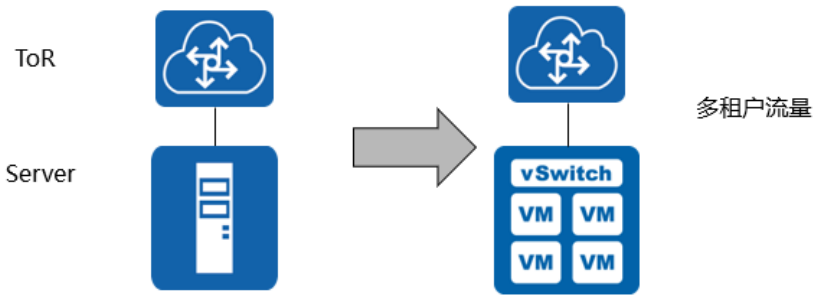
- 参考 RFC 草案，IP Prefix Advertisement in EVPN
- draft-ietf-bess-evpn-prefix-advertisement-11

华为EVPN技术的应用

- EVPN是一个新兴的技术。RFC冻结的协议标准，目前定义了Type1到Type4类路由，数据层面基于MPLS网络的LSP。
- 华为将EVPN技术应用于适配更广的IP网络，将EVPN技术和IP隧道技术结合。
- IP隧道采用VXLAN。BGP EVPN作为控制平面，VXLAN隧道作为数据转发平面。将BGP EVPN应用于NVO场景。

网络虚拟化Overlay (NVO)

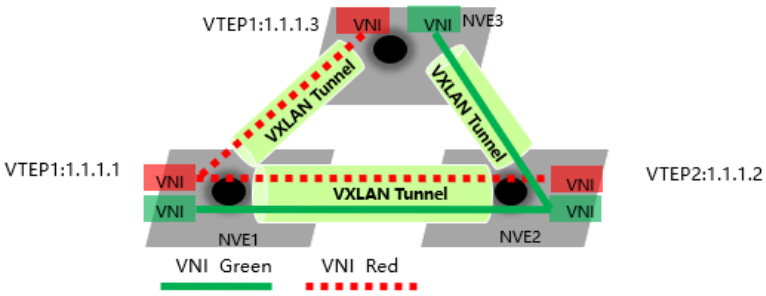
- 云数据中心中引入虚拟化技术，一个主机可以承载多个虚拟机，虚拟机属于不同租户。这对网络提出新的要求，NVO解决方案用来解决这个问题。



- NVO 实现每个租户的流量由独立的 Overlay 隧道承载。
- 在一个 underlay 网络之上可以承载多个 Overlay 隧道。

VXLAN封装

- 当前主流的NVO技术有VXLAN，NVGRE和MPLS Over GRE。
- VXLAN提供数据面的封装能力，用来在两个NVE设备之间，在普通IP网络上传输数据包。
- VXLAN封装基于UDP，使用8字节头部在UDP之后。24bit用于VNI，每个VNI标识一个租户。在进向的VTEP上封装好的数据帧不包含内部VLAN标签。出向的VTEP丢弃有inner VLAN的帧。



EVPN和网络虚拟化Overlay (1)

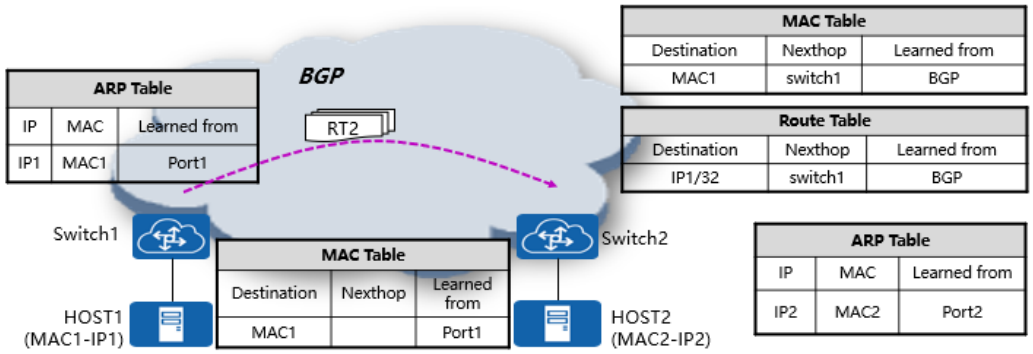
- NVO解决了多租户的问题，但是只有租户的数据平面。
- EVPN被用来作为NVO的控制平面。扩展BGP协议，定义基于EVPN的NLRI，用于传递数据转发之上控制信令。
- EVPN结合VXLAN可以满足：
  - 租户的网络隔离
  - 支持海量租户
  - 租户网络支持跨物理网络的大二层连接
  - 支持虚拟机网络跨物理网络的迁移

- 原有的 VXLAN 实现方案没有控制平面，是通过数据平面的流量泛洪进行 VTEP 发现和主机信息（包括 IP 地址、MAC 地址、VNI、网关 VTEP IP 地址）学习的，这种方式导致数据中心网络存在很多泛洪流量。为了解决这一问题，VXLAN 引入了 EVPN 作为控制平面，通过在 VTEP 之间交换 BGP EVPN 路由实现 VTEP 的自动发现、主机信息相互通告等特性，从而避免了不必要的数据流量泛洪。

EVPN和网络虚拟化Overlay (2)

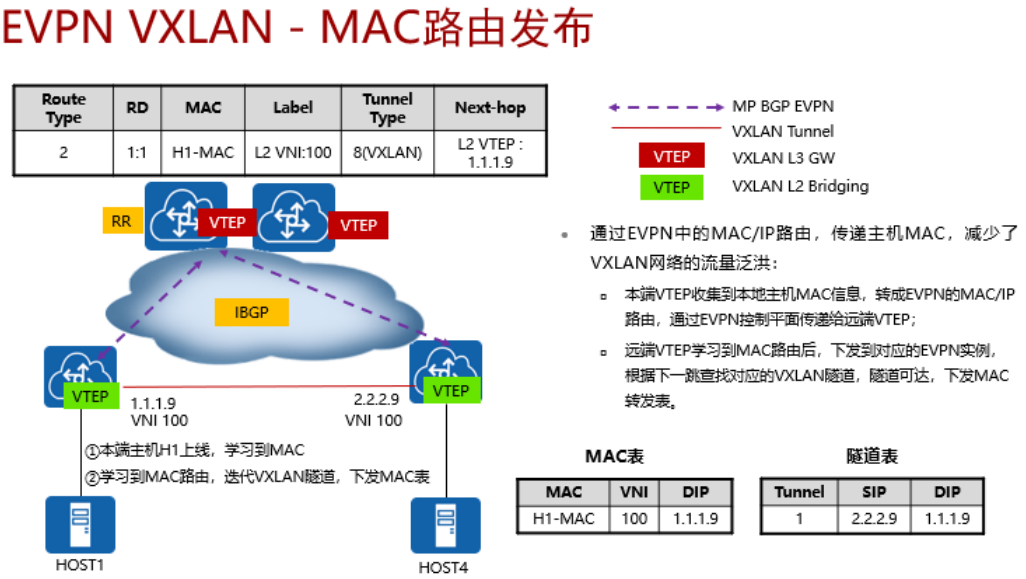
- EVPN最早是用于主要解决以下问题：
  - 控制平面由BGP分发，广播和多播由共享树或进向复制
  - 控制平面学习MAC
  - 路由反射替代全互联
- 因为NVO的需要解决百万实例在一个物理设施上，所以控制平面的扩展能力非常重要。EVPN及其扩展功能被设计满足这个需求。
- 华为将EVPN和VXLAN结合，主要使用Type2，Type3和Type5类路由。

EVPN路由 - Route Type 2

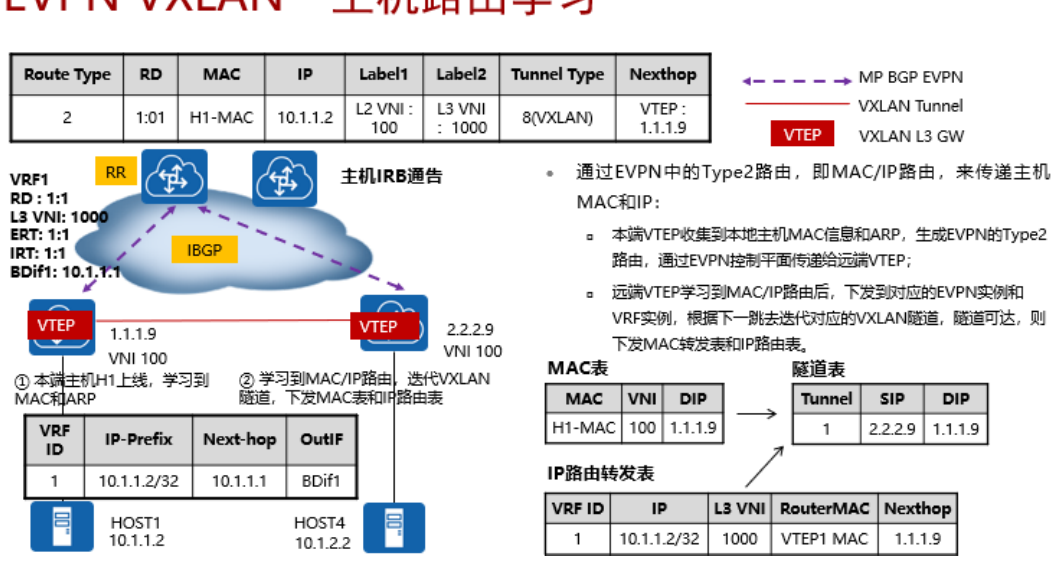


- 用于发布主机的MAC、MAC+IP。
- 将以太端口学习的MAC、ARP，转化为Type2路由发布给其他设备，其他设备接收后生成MAC转发表、主机路由转发表。
- 路由中IP字段为可选字段，可仅发布MAC，例如仅运行L2 VXLAN的交换机。

- EVPN 通过扩展 BGP 协议新定义了几种 BGP EVPN 路由，这些 BGP EVPN 路由可以用于传递 VTEP 地址和主机信息，因此 EVPN 应用于 VXLAN 网络中，可以使 VTEP 发现和主机信息学习从数据平面转移到控制平面。那么接下来，我们就开始，为大家一起揭开 BGP EVPN 路由的神秘面纱。
- 首先 VXLAN 需要使用 EVPN 协议规定的 Route Type2，又称为 MAC/IP Advertisement route，用于发布主机的 MAC 或 MAC+IP 的信息。BGP-EVPN 会通过 BGP 协议，将以太端口学习的 MAC、ARP，转化为 Route Type2 路由发布给其他设备，其他设备接收后生成 MAC 转发表、主机路由转发表。
- 这个功能可不简单，传统设备我们学习 MAC 地址都是通过报文触发，学习报文的源 MAC，而 BGP-EVPN 则通过 Route Type2 路由携带了 MAC 信息去发布 MAC，可以节省很多的 ARP 流量。



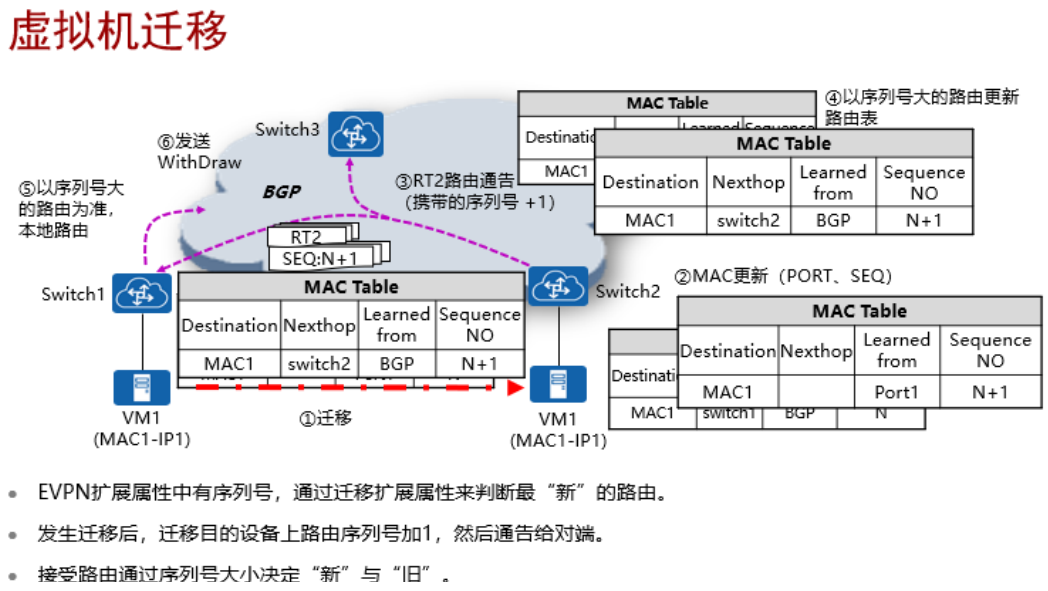
- 那么我们来看一下 Route Type2 的主要应用场景：
- 首先就是 MAC 路由发布，这里可以看到，当本端主机 H1 上线后，本端 NVE 学习到该主机的 MAC 地址，直接通过 BGP-EVPN 发送到远端设备；
- 远端 VTEP 学习到 MAC 路由后，下发到对应的 EVPN 实例，根据下一跳查找对应的 VXLAN 隧道，隧道可达，下发 MAC 转发表。



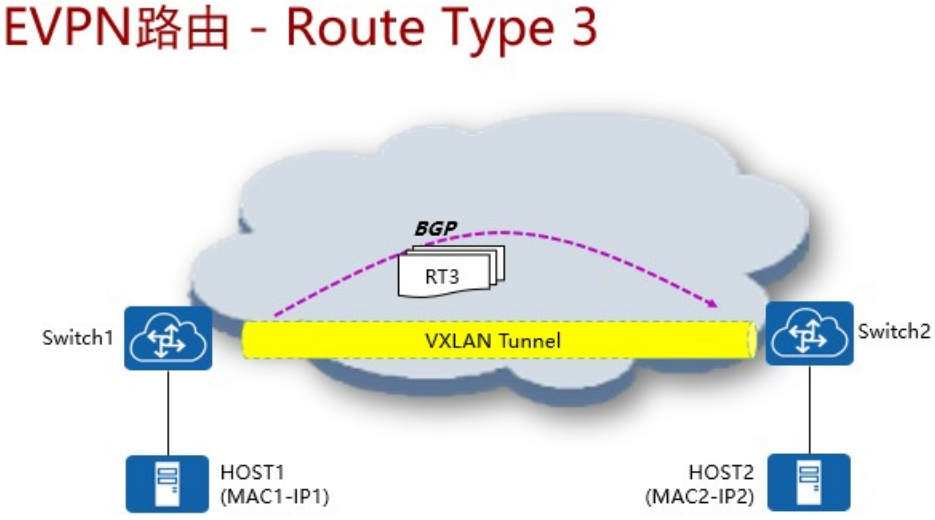
- 这里 Type2 路由发布的是 MAC/IP 路由，这里可以看到，当本端主机 H1 上线后，本端 VTEP 学习到该主机的 MAC 地

址和 ARP，生产 EVPN 的 Type2 路由，直接通过 BGP-EVPN 发送到远端设备。

- 远端 VTEP 学习到 MAC/IP 路由后，下发到对应的 EVPN 实例，根据下一跳查找对应的 VXLAN 隧道，隧道可达，下发 MAC 转发表和 IP 路由表。



- EVPN扩展属性中有序列号，通过迁移扩展属性来判断最“新”的路由。
- 发生迁移后，迁移目的设备上路由序列号加1，然后通告给对端。
- 接受路由通过序列号大小决定“新”与“旧”。

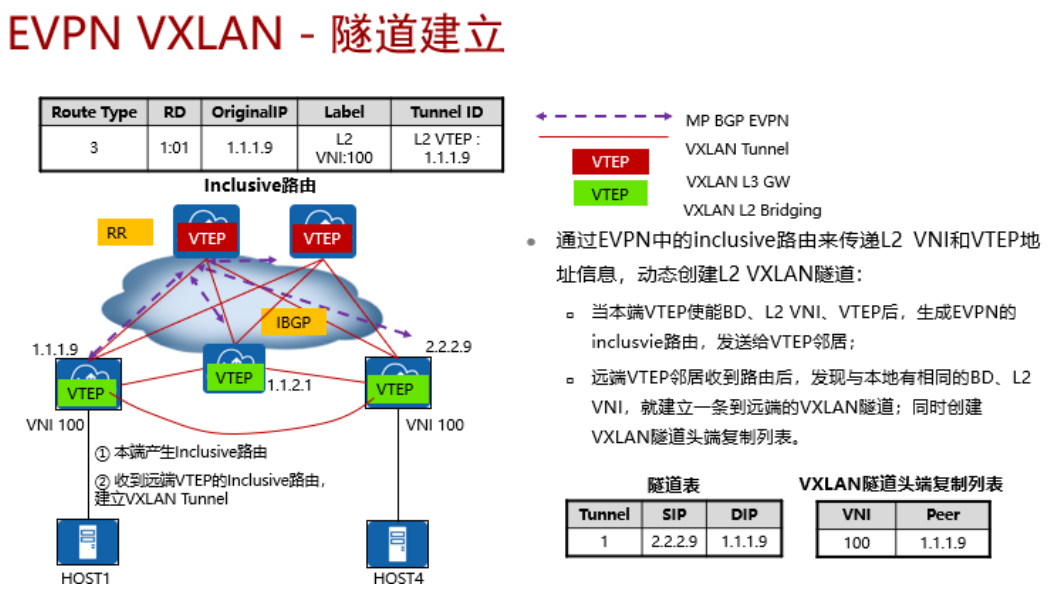


- Inclusive Multicast Ethernet Tag Route：
  - 用于隧道自动建立、VNI广播成员的自动加入。

- 接着我们看 EVPN 协议的 Type3 路由，又称为 Inclusive Multicast Ethernet Tag route，该类型路由是由前缀和 PMSI

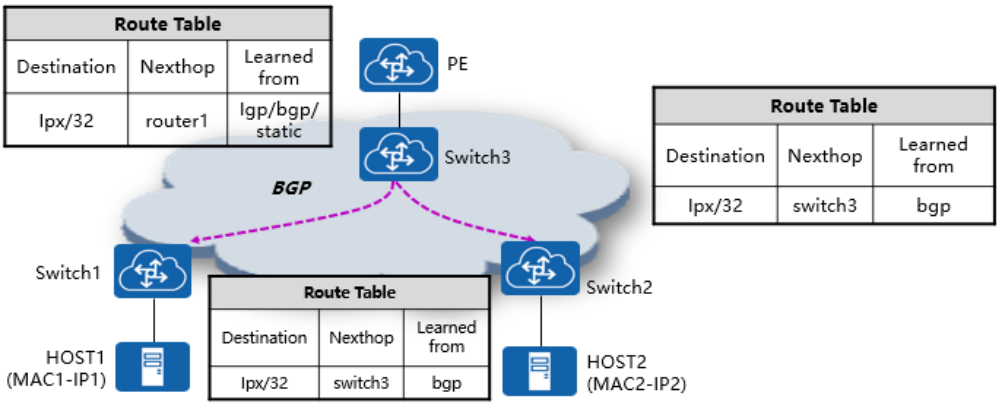
属性组成，用于隧道自动建立、VNI 广播成员的自动加入。

- 该类型路由在 VXLAN 控制平面中主要用于 VTEP 的自动发现和 VXLAN 隧道的动态建立。作为 BGP EVPN 对等体的 VTEP，通过 Inclusive Multicast 路由互相传递二层 VNI 和 VTEP IP 地址信息。
- 其中，Originating Router's IP Address 字段为本端 VTEP IP 地址，MPLS Label 字段为二层 VNI。如果对端 VTEP IP 地址是三层路由可达的，则建立一条到对端的 VXLAN 隧道。同时，如果对端 VNI 与本端相同，则创建一个头端复制表，用于后续 BUM 报文转发。



- VXLAN 隧道可以通过手动创建，通过指定两端的 VTEP 和 VNI 信息，静态配置创建隧道。而动态协议 BGP EVPN 中创建 VXLAN 隧道，则是通过 Type3 路由，可以把本端的 VTEP 地址、VNI 等信息发给远端，远端收到后会用来创建 VXLAN 隧道，并且创建 VXLAN 隧道的头端复制列表。

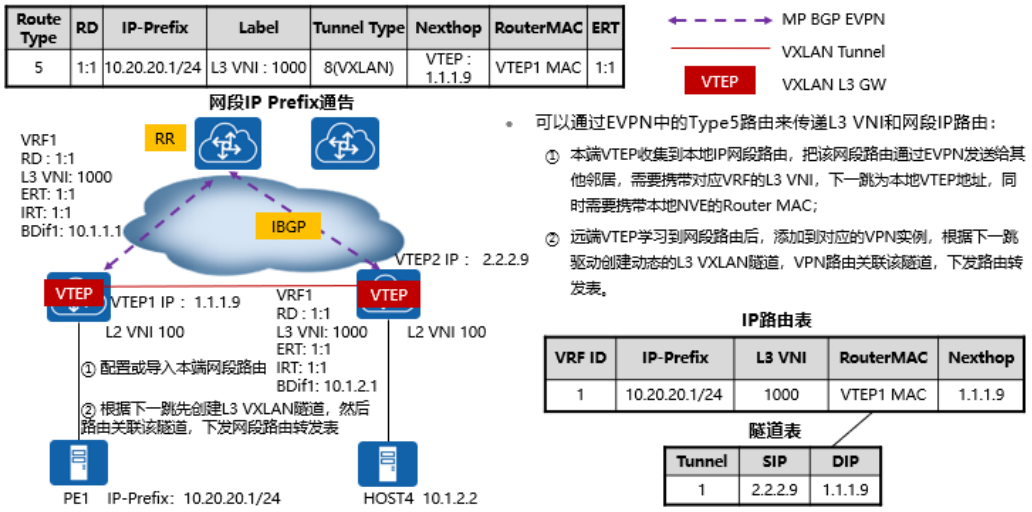
EVPN路由 - Route Type 5



- 又称为IP Prefix route：
  - 可用于将EVPN以外的子网网络引入到EVPN，也用于发布主机路由。

- 接下来我们再看一种路由，叫做 Type5 路由，又称 IP 前缀路由，可用于将 EVPN 以外的子网网络引入到 EVPN，当然可以掩码是 32 位，用于发布主机的 Host 路由。

EVPN VXLAN - 网段路由学习



- Type5 路由可以用来传递网段 IP 路由，同时可以携带对应 VRF 的 L3 VNI。
- 也可以用来传递代表 VRF 的 L3 VNI,这里什么是 L3 VNI 呢？

- 由于在分布式网关环境下，跨子网要通信，需要代表各自 VRF，而报文里面没有 VRF，因此我们这边通过一个特定的 VNI 映射成 VRF，这个 VNI 就是 L3 VNI。
- 这时，远端的 VTEP 在学习到网段路由后，可以添加到对应 VPN 实例中，并根据下一跳驱动创建 L3 VXLAN 隧道，并下发路由表。

思考题

1. BGP EVPN结合VXLAN使用了哪些类型路由？（ ）
   A. Type1
   B. Type2
   C. Type3
   D. Type4
   E. Type5
2. 简要描述Type2类路由作用。

- 参考答案：
- BCE
- MAC/IP 通告，分布式网关下虚拟机迁移，MAC 通告。