DRNI+EVPN组网进阶版介绍

H3C交换机技术专题篇

DRNI+EVPN

上一期交换那些事儿我们主要介绍了DRNI+EVPN组网的部署方案和简单配置，小伙伴们都掌握了吗?

本期我们将进入DRNI+EVPN组网的进阶教程，主要介绍一些特殊组网的配置和特殊命令的使用，包括动态AC、Border特殊配置、单挂AC特殊配置等，不要眨眼哦~

动态AC

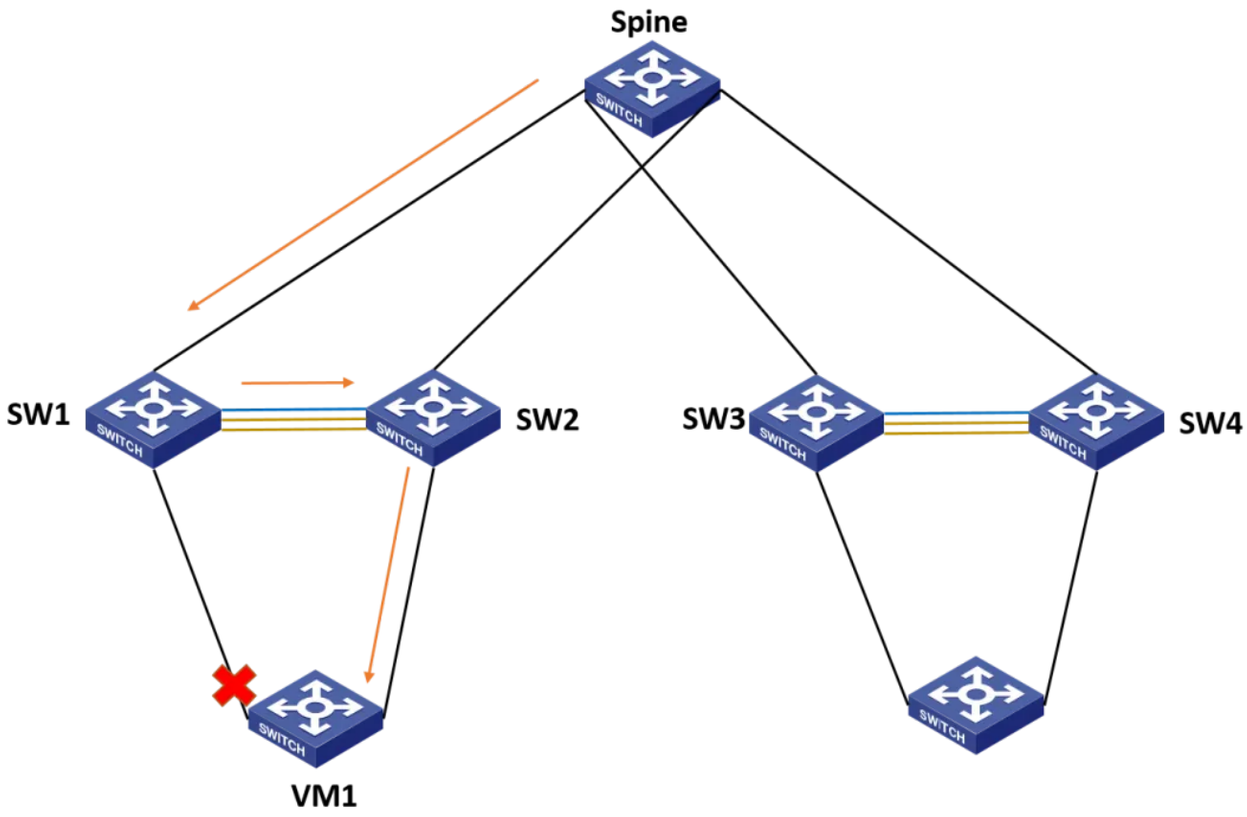
在上一期我们提到，作为Leaf的两台DR设备上会自动在IPL上创建AC，并将其与VSI关联，比如下面这个Type为DRNI的AC：

```
[SWA]display l2vpn vsi verbose
VSI Name: vpna
  VSI Index                : 0
  VSI State                : Up
...
Gateway Interface        : VSI-interface 1
  VXLAN ID                 : 10
  Tunnels:
    Tunnel Name    Link ID  State    Type  Flood proxy
    Tunnel0       0x5000000  UP      Auto   Disabled
  ACs:
    AC              Link ID State  Type
    BAGG4 srv1000     0       Up    Manual
    BAGG3 srv2        1       Up   Dynamic (DRNI)
```

在EVPN组网中，终端通过AC口连接VTEP设备

时，会在相应的接口根据其匹配的服务实例建立AC；在DRNI加入后，用户侧接口上创建以太网服务实例、配置报文匹配规则并关联VSI后，VTEP会在IPL上自动创建具有相同报文匹配规则、关联相同VSI的动态AC，以实现用户侧链路的备份。

如下图所示，如果SW1连接VM1的链路故障，远端访问VM1的报文hash到SW1上，由于SW1在IPL链路上有VM1的动态AC，报文就可以通过IPL绕行到SW2，同时正常进行AC口的解封装打tag进程。



这里有一个小问题：

如果不同用户侧接口上AC的报文匹配规则相同，但关联了不同的VSI，则会导致IPL上动态创建的AC之间发生冲突，可能导致业务不通。为了避免这种情况，我们可以配置**l2vpn drni peer-link ac-match-rule**

vxlan-mapping命令来规避。

配置这条命令后，IPL上的AC报文匹配规则将由VXLAN ID映射生成，这样和本身的匹配规则不一样，也就避免了冲突。这个映射规则看起来比较复杂，有兴趣的同学可以研究一下~

报文匹配规则为：

- 外层VLAN标签（s-vid）为VXLAN ID除以4094，取整后加1，即VXLAN ID/4094+1。
- 内层VLAN标签（c-vid）为VXLAN ID除以4094，取余后加1，即VXLAN ID%4094+1。
- 如果计算出来的外层VLAN标签为IPP口的PVID，则s-vid变更为VXLAN ID%4094+1、无c-vid。

注：执行本命令后，VTEP上创建的VXLAN的ID不能大于16000000。

这个问题在某些交换机上还可以通过另一条命令来解决，这条命令是**l2vpn drni peer-link tunnel source** *source-address* **destination** *destination-address*，配置这条命令后，IPL上不会动态创建AC，作为DRNI成员设备的两台设备之间会自动建立VXLAN隧道，并将该VXLAN隧道自动与所有VXLAN关联。此时，两台DR设备之间存在两条逻辑链路（IPL链路和自

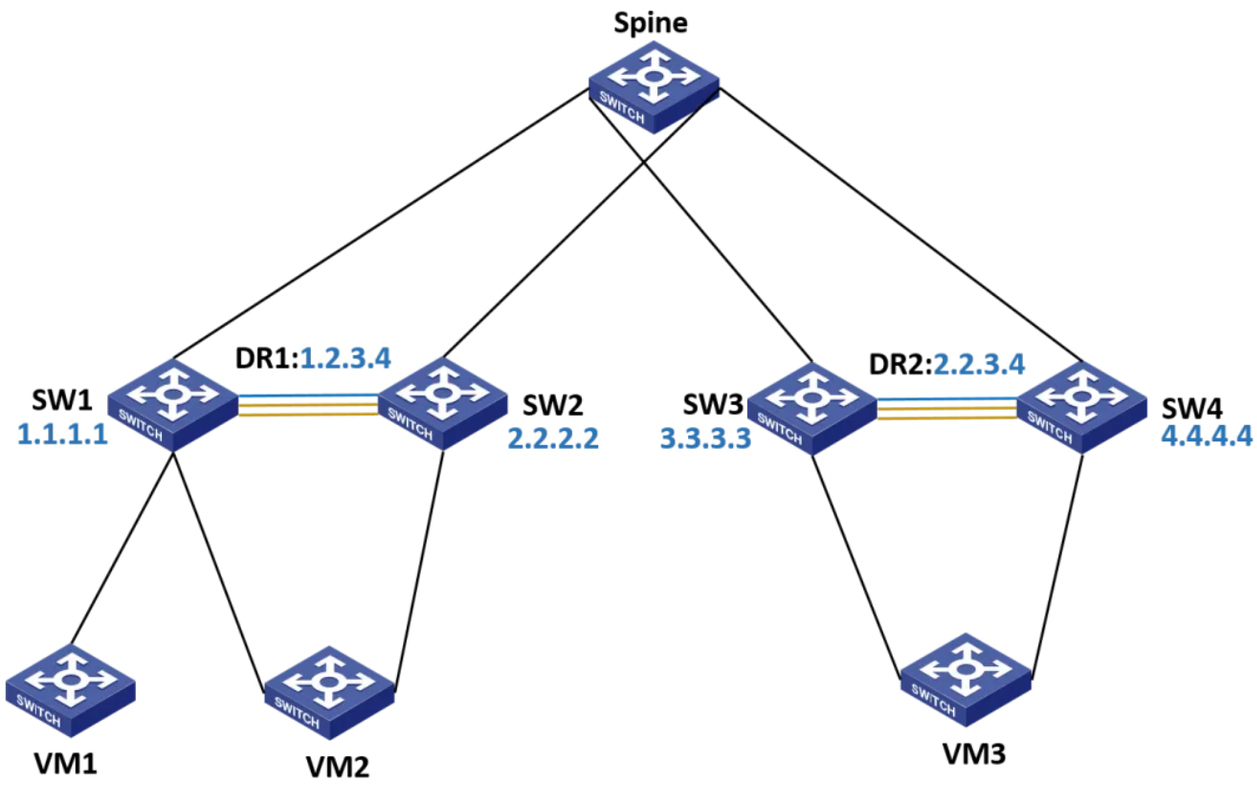
动创建的VXLAN隧道），二者负责转发不同的流量，以实现负载分担。由于不会创建动态AC，这条命令还可以解决动态AC资源不足的问题，目前这个功能还在持续开发中，使用前一定要记得看相应设备的配置指导哦~

单挂AC

在上一期的实验中，VM都是双上行通过DR聚合口和DR系统互联的。整个实验组网就是简单的双挂DR系统做Leaf，使用group ip和远端Leaf建立VXLAN隧道。

但在实际组网中，部分终端设备也可能并不支持双上行对接，如下图所示，SW1/2和SW3/4组成两组DRNI系统作为VTEP，上行对接Spine，下行通过DR

聚合口对接终端VM2和VM3。但SW1仅通过单链路对接VM1，不通过DR口，对SW1/2这个DR系统组成的VTEP而言，VM1就是一个单挂AC。



在DRNI+EVPN组网中，我们会通过evpn drni group *ip-address*命令开启EVPN分布式聚合功能，这条命令也同时配置group ip作为虚拟VTEP地址，配置后DRNI系统在整个EVPN组网中虚拟为一台VTEP，通过配置的group ip发布2、3、5类路由，远端VTEP收到这些路由消息后，由于下一跳是group ip，就会根据其下一跳与DRNI系统建立虚地址隧道。

因此在上图这个组网中，我们按照之前介绍的配置对各台设备进行配置后，由于SW1/2和SW3/4各自是一个DRNI系统，都会使用自己的group ip发布2、3类路由，因此SW1/2和SW3/4之间会通过group ip建立一条tunnel：

SW1/2：1.2.3.4>>2.2.3.4

SW3/4：2.2.3.4>>1.2.3.4

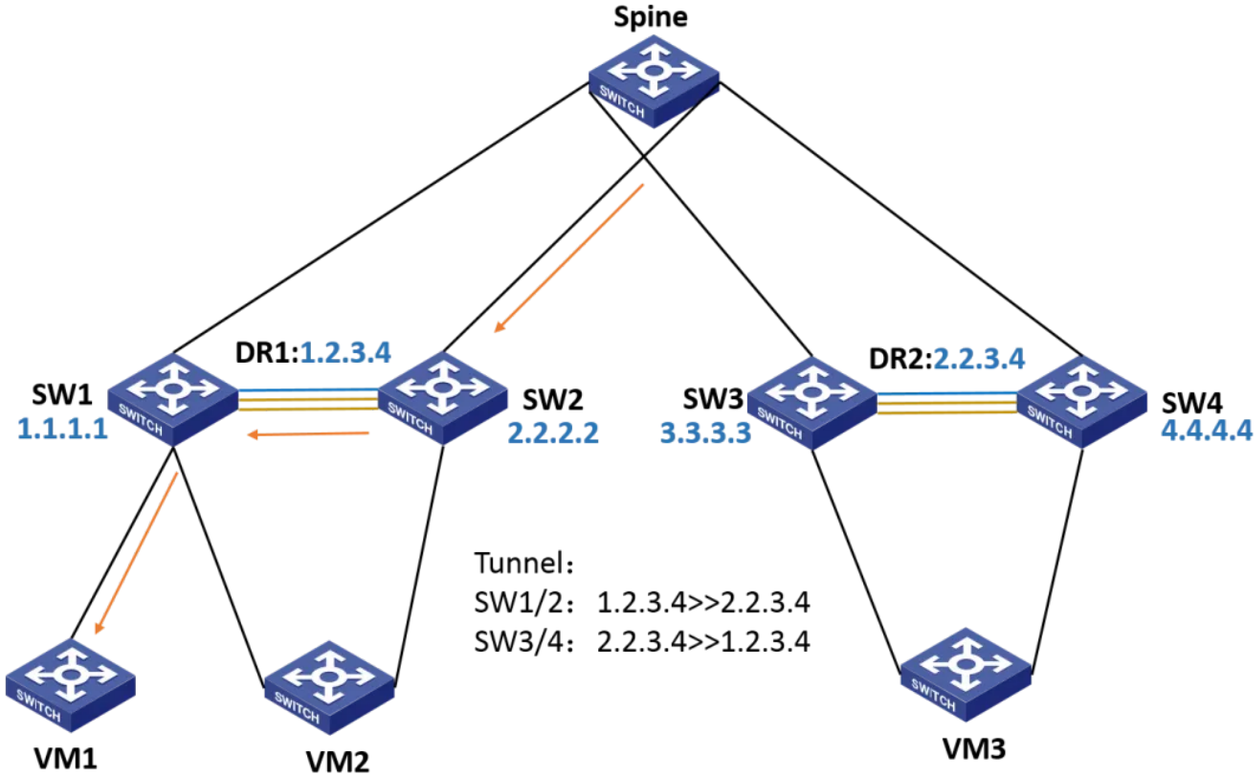
我们在这个组网中讨论如下情况：

SW1/2/3/4都是Leaf设备，VM1单挂在SW1下。

此时由于配置了group ip，SW1/2和SW3/4间只会通过虚地址建立隧道，每台设备上都只有一条虚地址隧道。VM3访问VM1和VM2的报文可能hash到SW1上，也可能hash到SW2上。

在DR系统内部，我们以SW1/2为例，SW1上生成VM1和VM2的2类路由，通过group ip：1.2.3.4对外发布；SW2上生成VM2的2类路由，同样由group ip：1.2.3.4对外发布。由于Spine是反射器，收到SW1发布的路由后会反射给SW2和远端的SW3/4，SW2收到反射器反射回来的SW1发布的下一跳是group ip为1.2.3.4的路由，检查发现下一跳就在本机上，不会加表。

这种情况下对双挂的VM2没有影响，无论从远端过来的报文hash到哪一台设备上，都可以转发到VM2上；对VM1而言，SW2上不会有VM1的主机路由，但是SW2会通过IPL链路备份表项，出接口下发在IPP口上，远端访问VM1的流量如果hash到SW2上，就会通过IPL绕行到SW1。如下图所示：



这种流量绕行增加了中间IPL链路的负担，那有什么办法能让访问VM1的流量只走SW1吗？有的，方法就是在SW1/2上都配置**evpn drni local***local-ip* **remote** *remote-ip*命令（其中local-ip是本地实地址，remote-ip是对端另一台DR设备的实地址）。我们在SW1/2上分别增加如下配置：

SW1上：evpn drni local 1.1.1.1 remote 2.2.2.2

SW2上：evpn drni local 2.2.2.2 remote 1.1.1.1

配置完成后，SW1/2除了以group ip发布路由外，还会以命令行中配置的local ip发布实地址下一跳的路由，这样在远端SW3/4上就会收到2+2四条路由，由于SW1/2发布的虚地址路由都是一样的，不会重复建立隧道，因此SW3/4上会与SW1/2建立3条隧道，分别是：

SW3/4：

A 2.2.3.4>>1.2.3.4

B 2.2.3.4>>1.1.1.1

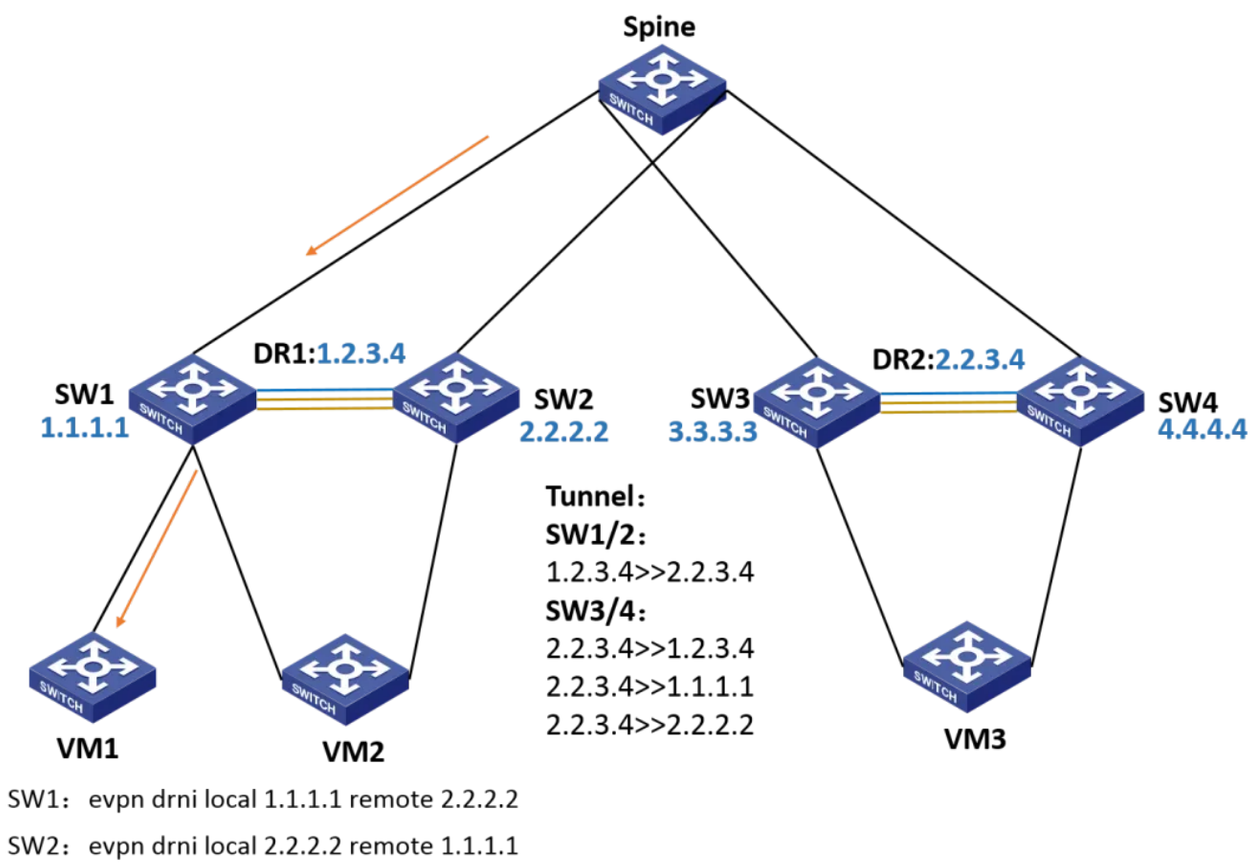
C 2.2.3.4>>2.2.2.2

在SW1/2上，由于SW3/4未做其他配置，只会发布虚地址下一跳的路由，因此SW1/2上只有一条到SW3/4的虚地址隧道：

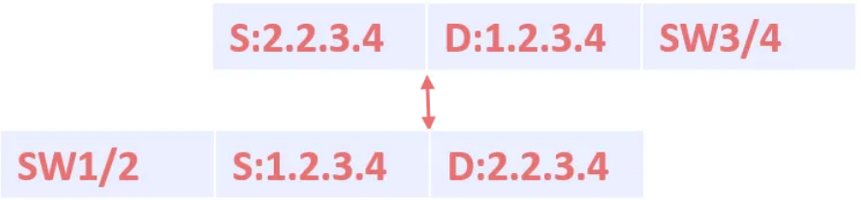
SW1/2：

D 1.2.3.4>>2.2.3.4

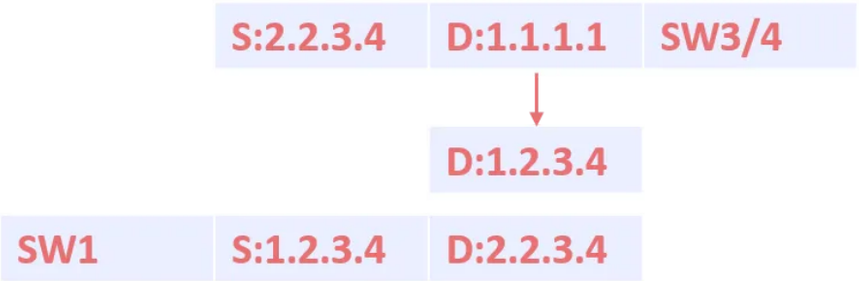
此时，当VM3去访问VM1时，可以直接走隧道B 2.2.3.4>>1.1.1.1到SW1，不会再绕行。



这里我们再补充一个隧道对称的概念，设备收到带封装的报文后，如果其封装的源目IP在设备上可以找到对称隧道，则可以解封装。比如VM3访问VM2的报文，走隧道A封装源2.2.3.4目的1.2.3.4到达SW1上，SW1上有其对称隧道D 1.2.3.4>>2.2.3.4，就可以正常解封装转发。



当VM3访问VM1时，报文走隧道B，带着源2.2.3.4目的1.1.1.1封装到达SW1，这条隧道的对称隧道应该是1.1.1.1>>2.2.3.4，SW1上没有，但是由于配置了evpn drni local 1.1.1.1 remote 2.2.2.2命令，SW1驱动可以将报文隧道目的地址1.1.1.1（配置的local-ip）映射为虚地址1.2.3.4，这样报文的封装对SW1而言就变成了源2.2.3.4目的1.2.3.4，SW1上有其对称隧道D 1.2.3.4>>2.2.3.4，就可以正常解封装报文再转发给VM1了。



除此以外，由于SW2上配置了evpn drni local 2.2.2.2 remote 1.1.1.1，当SW1发布的实地址下一跳

路由经Spine（RR）反射到SW2后，SW2查看其下一跳为自己配置的remote-ip 1.1.1.1，会将这个路由过滤掉，不会根据这条路由建立隧道。

根据上述说明，**我们总结一下evpn drni local** *local-ip* **remote** *remote-ip***这条命令的原理和应用：**

local-ip：DR设备本地实地址

remote-ip：对端另一台DR设备的本地实地址

配置命令后，除了原有的虚地址下一跳路由，设备会以实地址发布一条路由，远端VTEP设备可以与其建立实地址目的地址隧道，防止访问单挂AC的流量绕行；

配置命令后，设备驱动可以根据local-ip字段进行虚实地址映射，当收到远端发过来的封装目的IP是local-ip的报文时，可以将其映射为虚地址group ip，解决了隧道不对称的问题；

配置命令后，两台DR设备都会向外发布实地址下一跳路由，该路由可以经RR反射到对端另一台DR设备，对端DR设备会根据自己配置的remote-ip过滤这些路由，不会根据这条路由建立隧道。

注意事项：

evpn drni local *local-ip* remote *remote-ip*命令需要一组DR设备对称配置，不能只单独配置一台；

该命令只需要在有单挂AC的DR Leaf设备上配置，没有单挂的不需要配；

命令配置后，无论该DR设备是否下连单挂AC，都会发布实地址下一跳路由。

Border

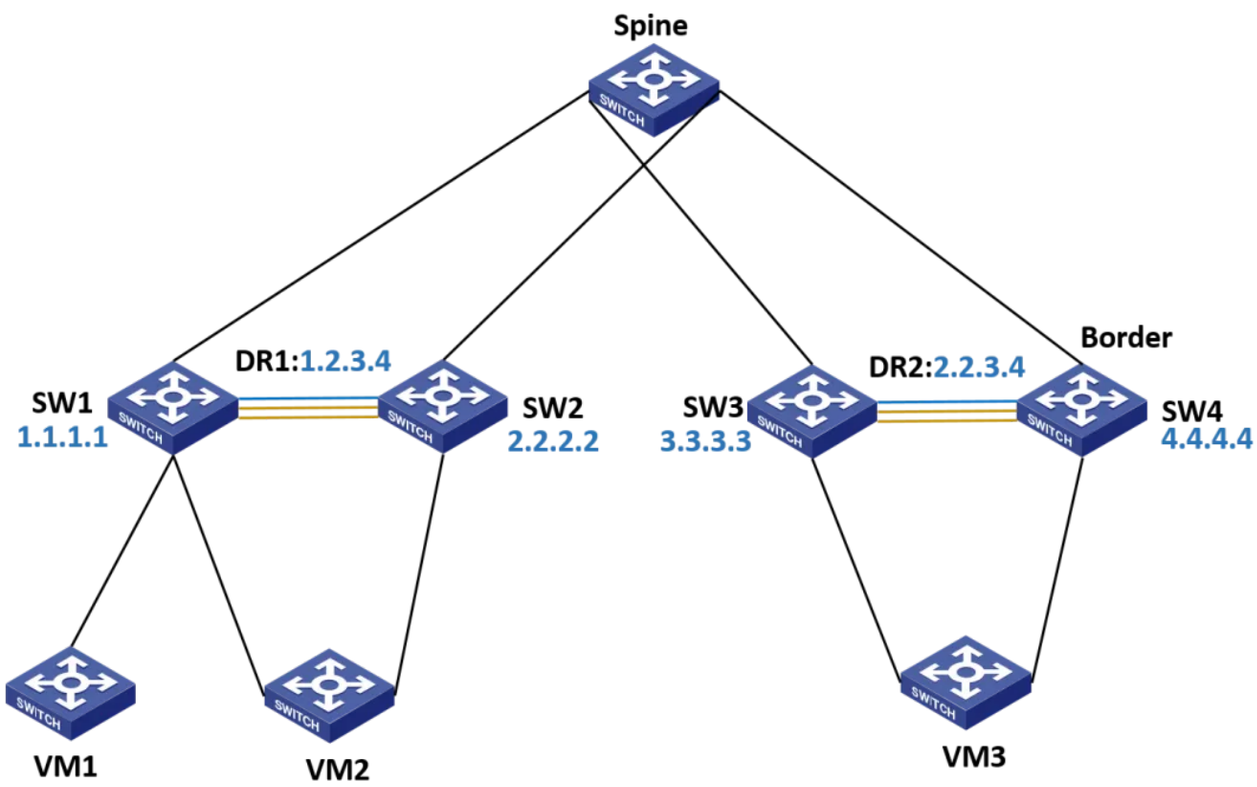
Border是边界网关，在EVPN组网中的作用是与外网互通，相比Leaf设备，Border在发布EVPN路由时有如下特点：

1、和普通的2、3、5类路由不一样，在Border上通过bgp vpn ipv4/6引入的5类路由在发布时不会以group ip发布，只能携带实地址下一跳；

2、当Border 上没有VXLAN或者DR口的配置时，无法发布虚地址下一跳的2、3类路由。

基于这些特点，我们一起来看下DRNI+EVPN组网中Border的配置有哪些特别之处。

我们回到之前的组网，SW3/4是border设备，双挂接入外网，引入5类路由；SW1/2还是Leaf设备，下联VM1和VM2。



这里我们分两种情况来讨论，**第一种情况，SW3/4上没有VXLAN或者DR口的配置，只是在bgp vpn ipv4/6视图引入外网路由。**

在这种情况下，Border引入的5类路由只会通过实地址发布，且不会发布虚地址下一跳的路由。那么在远端Leaf SW1/2上只会收到从SW3/4上发布过来的两台实地址下一跳的5类路由，建立两条隧道：

```
Leaf SW1/2：
A 1.2.3.4>>3.3.3.3
B 1.2.3.4>>4.4.4.4
```

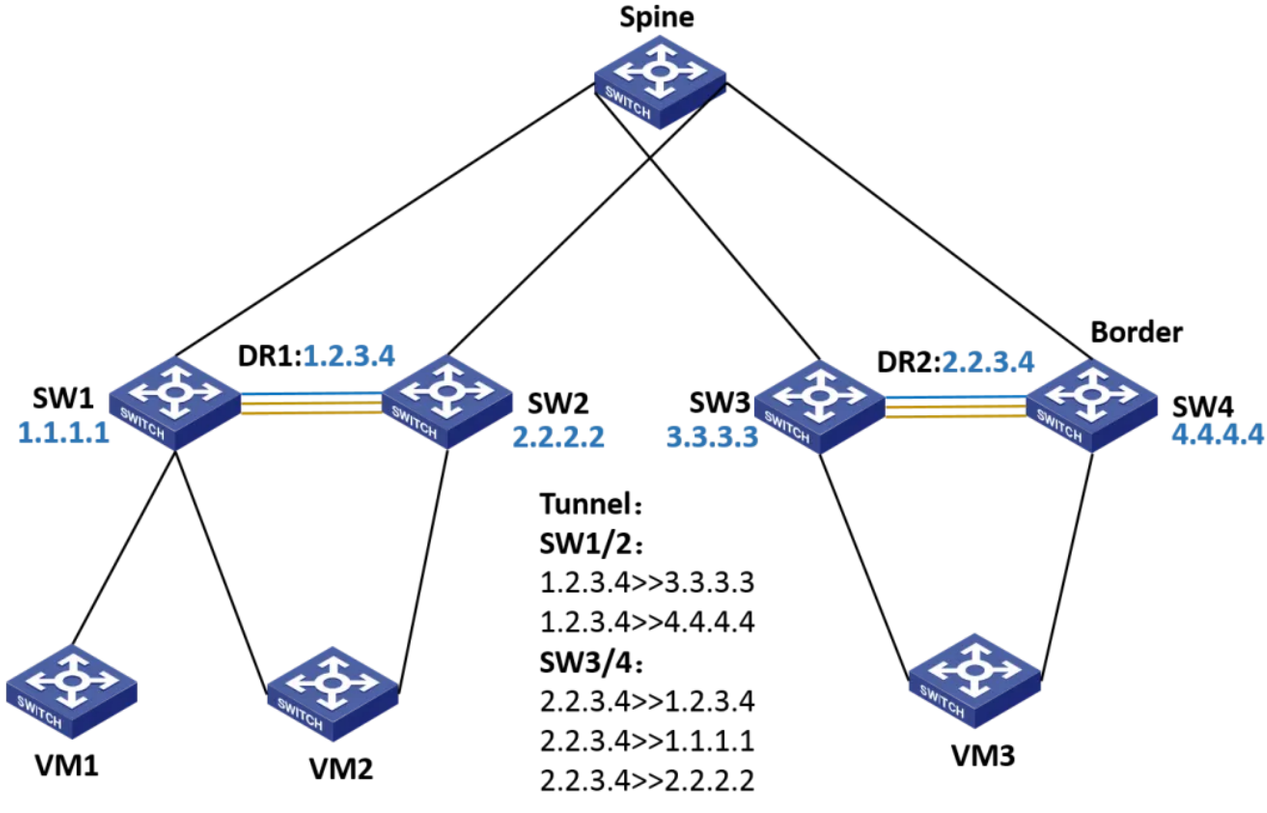
SW3/4上正常收到SW1/2发布的虚地址下一跳路由，由于SW1/2下有单挂AC，正常会配置evpn drni local *local-ip* remote *remote-ip*命令来防止绕行，所以SW3/4上还会收到两条实地址下一跳路由，一共建立3条隧道：

Border SW3/4：

C 2.2.3.4>>1.2.3.4

D 2.2.3.4>>1.1.1.1

E 2.2.3.4>>2.2.2.2



此时从VM2访问VM3的报文，如果走隧道A 1.2.3.4>>3.3.3.3到SW3，SW3上没有对称的隧道，也没有配置evpn drni local *local-ip* remote *remote-ip*命令，无法映射到虚地址，因此无法正常解封装报文；

从VM3到VM2的报文，从隧道C 2.2.3.4>>1.2.3.4到SW1/2上，此时SW1/2上都没有隧道C的对称隧道，也不能解封装；

从VM3到VM1的报文，从隧道D 2.2.3.4>>1.1.1.1到SW1上，此时SW1上没有隧道D的对称隧道，即使将local-ip 1.1.1.1映射到group ip 1.2.3.4，也没有对称的虚地址隧道，不能解封装。

在这种隧道不对称的情况下，我们可以在无法解封装的设备上配置 **vxlan default-decapsulation source interface** *interface-type interface-number* 命令来对报文强制解封装。配置这条命令后，设备收到带封装的报文，不论是否存在对称隧道，都可以强制解封装再查表转发。这样VM1/2和VM3就可以互访了。

第二种情况，SW3/4上配置了VXLAN或者DR口，在bgp vpn ipv4视图引入外网路由。

这种情况下，SW3/4会发布下一跳是虚地址group ip的2/3类路由，也会发布引入的下一跳是实地址的5类路由。此时在SW1/2上会建立3条隧道：

```
Leaf SW1/2：
A 1.2.3.4>>3.3.3.3
B 1.2.3.4>>4.4.4.4
C 1.2.3.4>>2.2.3.4
```

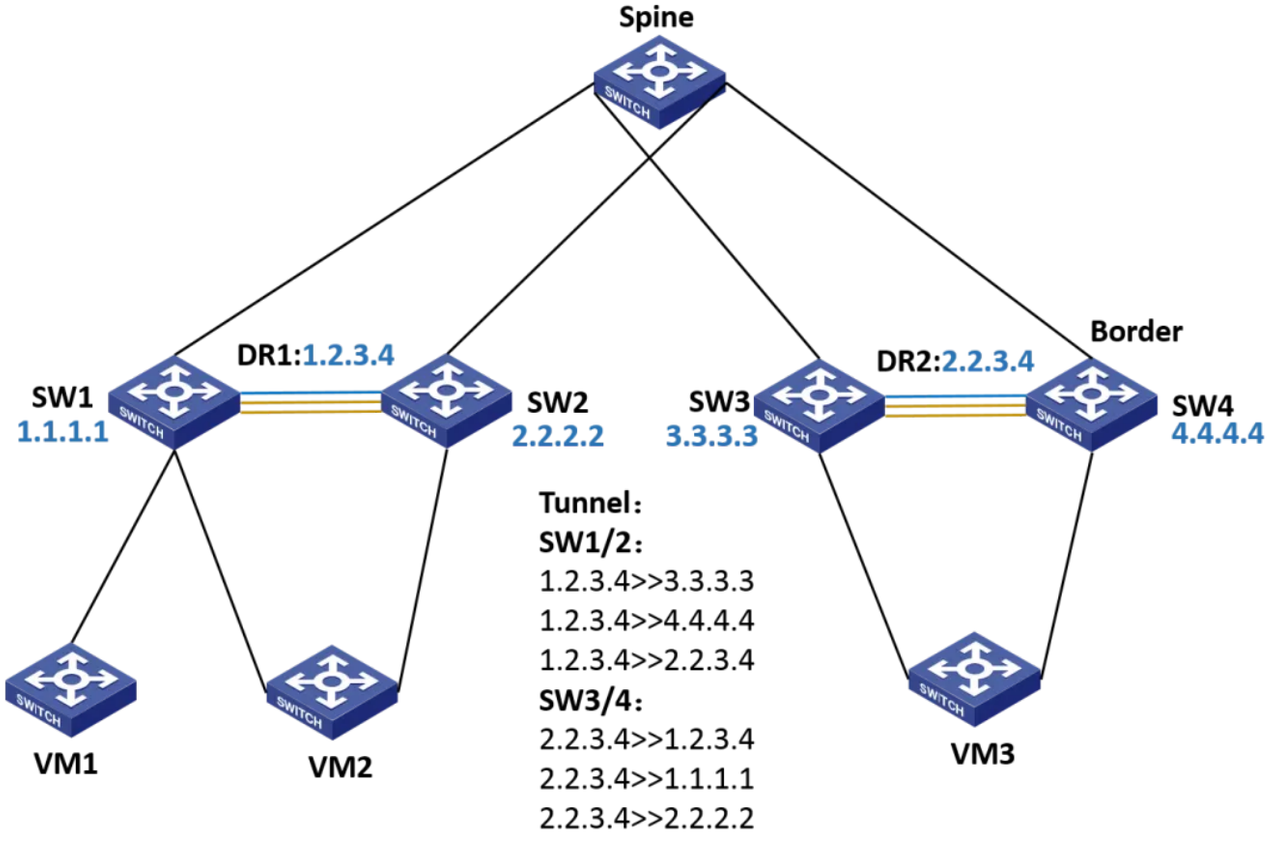
由于SW1/2下有单挂AC，会配置evpn drni local *local-ip* remote *remote-ip*命令来防止绕行，除了虚地址下一跳路由，SW3/4上还会收到两条实地址下一跳路由，一共建立3条隧道：

Border SW3/4：

A 2.2.3.4>>1.2.3.4

B 2.2.3.4>>1.1.1.1

C 2.2.3.4>>2.2.2.2



此时从VM2访问VM3的报文，如果走的是隧道A 1.2.3.4>>3.3.3.3或者隧道B 1.2.3.4>>4.4.4.4，由于SW3/4没有配置evpn drni local *local-ip* remote *remote-ip*命令，无法映射，还是无法在SW3/4上正常

解封装，需要配置强制解封装命令；

但从VM3访问VM2的报文，通过隧道D 2.2.3.4>>1.2.3.4到SW1/2，SW1/2上有对称的隧道C 1.2.3.4>>2.2.3.4，因此可以正常处理；

VM3访问VM1的报文，从隧道E 2.2.3.4>>1.1.1.1到SW1后，SW1根据evpn drni local 1.1.1.1 remote 2.2.2.2命令将隧道目的1.1.1.1映射到1.2.3.4，映射后也有对称隧道C 1.2.3.4>>2.2.3.4，因此可以正常处理。

因此在这种情况下，Leaf SW1/2上无需配置vxlan default-decapsulation source interface *interface-type interface-number*强制解封装命令，只需要在Border SW3/4上配置即可。

那这里是否可以在SW3/4上配置evpn drni local *local-ip* remote *remote-ip*命令，让设备将A 1.2.3.4>>3.3.3.3或者B 1.2.3.4>>4.4.4.4隧道过来的报文隧道目的地址映射到虚地址呢？答案是不可以的，Border配置这条命令有些问题，且我们找到了更好的办法来解决这个问题，下文会说到。

综合以上两种情况，由于Border从bgp vpn视图下引入5类路由时缺省填充实地址下一跳，不能配置evpn drni local *local-ip* remote *remote-ip*命令，无法映射，

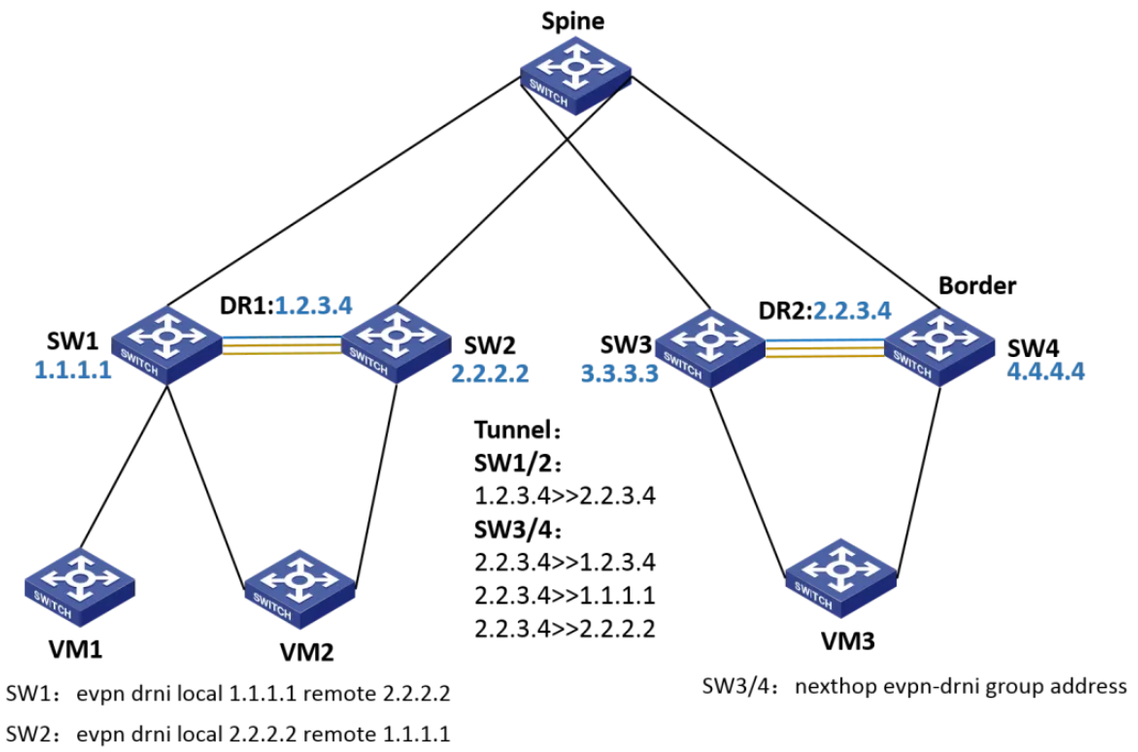
导致和远端Leaf建立的隧道不对称，需要配置强制解封装 vxlan default-decapsulation source interface *interface-type interface-number*命令；

如果Border上没有VXLAN或者DR口配置（即border无法发布虚地址下一跳的路由）时，Leaf侧没有到border的虚地址隧道，也需要在Leaf上配置强制解封装。

> vxlan default-decapsulation source interface *interface-type interface-number* 命令中 source interface一般配置BGP邻居地址或者loopback地址，没有特殊要求，配置后无论隧道是否对称，都可以直接解封装报文再进行转发。现网实际组网中，Border上大部分情况也会有VXLAN配置或者DR口，一般不存在无法发布虚地址下一跳路由的情况，所以一般Leaf侧无需配置强制解封装。

从上面第二种情况的讨论中可以发现，其实Border上配置这么麻烦，最大的问题就是从bgp vpn视图引入5类路由时只能填充实地址。不过我们已经有办法解决这个问题啦，目前很多设备都支持配置**nexthop evpn-drni group address**命令，配置这条命令后，Border引入5类路由在发布时也会直接填充group ip，不会再发布实地址下一跳的路由。

以下图为例，SW1/2上有单挂AC，配置了evpn drni local *local-ip* remote *remote-ip*命令；SW3/4是border，引入了5类路由，配置了nexthop evpn-drni group address命令。这种情况下，SW1/2上只会收到SW3/4发布的虚地址下一跳路由，建立虚地址隧道；SW3/4上会收到SW1/2发布的虚/实地址下一跳路由。



此时终端互访的报文都可以正常封装、解封装、查表转发，这种情况下就不需要配置强制解封装命令了；且由于Border不会发布实地址下一跳路由，Leaf侧不管何种情况，都不需要配置强制解封装。

nexthop evpn-drni group address命令有以下几点需要注意：

nexthop evpn-drni group address命令优先级没有evpn drni local *local-ip* remote *remote-ip*

命令高，所以不能同时配置；

nexthop evpn-drni group address仅border使用。

总结

基于上文的讨论，我们可以对DRNI+EVPN组网中Border和单挂AC的情况做一个总结：

1、 DR系统作为VTEP都需要配置evpn drni group *ip-address*命令；

2、 DR系统做Leaf时，如果有单挂AC，就需要在两台DR设备上对称配置evpn drni local *local-ip* remote *remote-ip*命令；

3、 DR系统做Border时，引入5类路由，建议配置nexthop evpn-drni group address命令；如果当前设备不支持该命令，则配置vxlan

default-decapsulation source interface *interface-type interface-number*命令。

以上就是本期全部内容啦，单挂和Border的配置理解起来比较复杂，但实际配置并不复杂的，有问题或者建议欢迎留言哦~

—— end ——

扫码关注我们哦