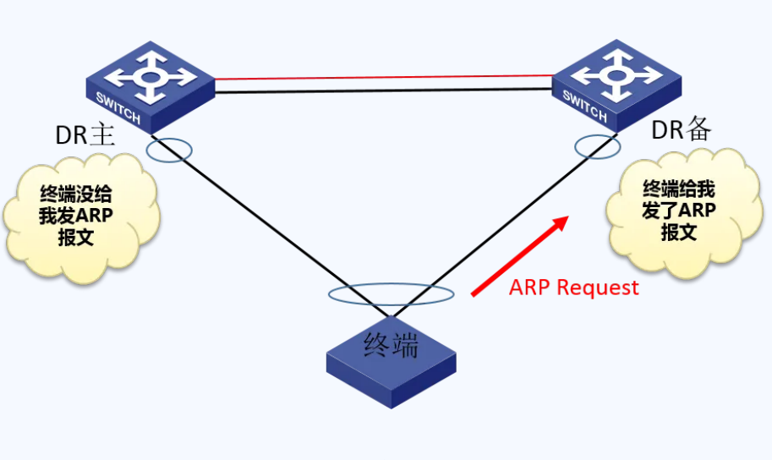
DRNI组网ARP及MAC表项同步机制介绍

H3C交换机技术专题篇

ARP和MAC表项是网络设备二三层转发过程中不可或缺的一部分。在之前几期DRNI分享中，我们了解到下联终端通过跨设备聚合链路和DR系统互联，终端上行报文在聚合链路上逐流hash，ARP报文和普通业务报文都可能hash到DR系统主设备，也可能hash到备设备，为了让DR主备设备都可以正常进行二三层转发，必须保证DR系统主备设备间表项同步。



那么，DR系统主备两台设备之间是怎样完成ARP和MAC表项同步的呢？

DRNI通过在IPL上运行DRCP（Distributed Relay Control Protocol，分布式聚合控制协议）来交互分布式聚合的相关信息，以确定两台设备是否可以组成DR系统。DRNI通过在IPL上运行Rlink通道来进行协议控制报文的交互和表项的同步等，eth协议号为8843，Rlink的主要作用如下：

- 同步协议控制信息（比如STP根信息）
- 同步配置一致性检查数据
- **同步协议报文（比如ARP/STP等报文）**
- **同步表项信息（比如MAC等）**

今天我们主要介绍一下DRNI组网中ARP和MAC表项的同步机制~

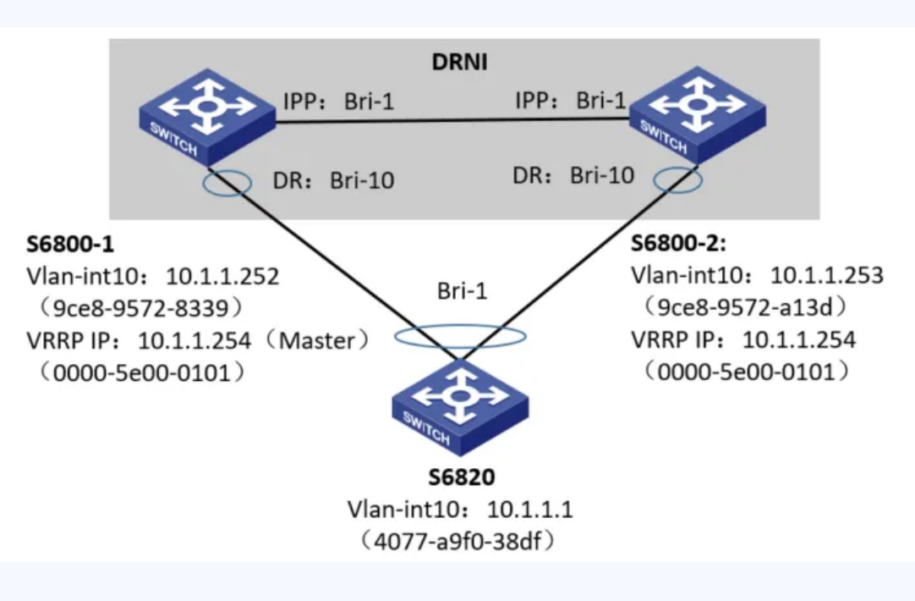
DRNI组网中的ARP同步机制

为了便于理解，我们结论先行，DRNI组网中的ARP同步有如下机制：

- DRNI设备之间同步ARP时，同步的是ARP报文，不是ARP表项；
- DR口收到接口ARP报文，原始报文按正常的转发逻辑正常从IPL和其他接口泛洪转发；同时会Copy到CPU，封装到RLink报文中同步给另一台DR设备；
- IPP口会收到两种ARP报文，一种是Rlink同步过来指向DR口的报文，一种是泛洪过来指向IPP口的报文，设备处理遵循DR口优先原则：指向DR口的报文可以覆盖IPP口的ARP表项，指向IPP口的ARP报文在一定条件下才能覆盖DR口的表项（查询此ARP表项的MAC对应的MAC表，若MAC对应的出接口也是IPP口，才能覆盖）
- Rlink过来的ARP报文会被打上标记，设备针对这个带标记的ARP报文，不会回复ARP应答；
- 单挂终端ARP报文不会被Rlink报文封装同步给另一台DR。

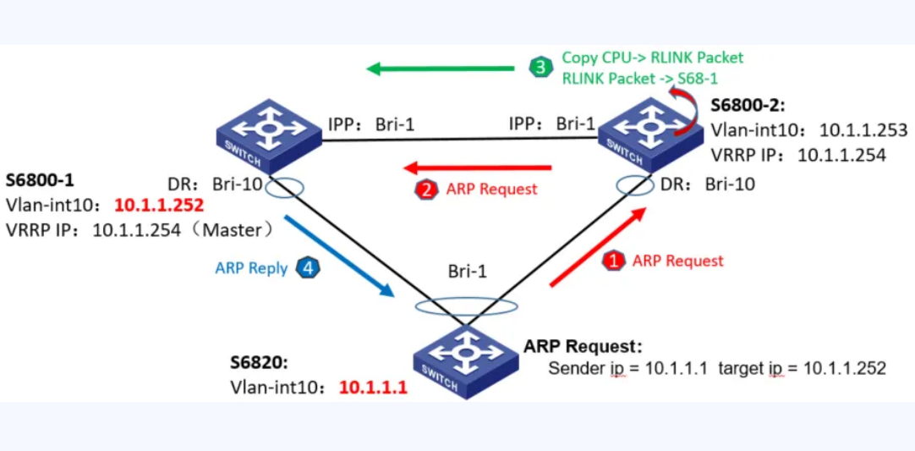
接下来我们来看几个测试案例，详细了解一下DRNI组网中的ARP同步机制。

测试组网如下，两台交换机（S6800-1/S6800-2）组成DRNI系统，下联终端设备（S6820），组网为DDRNI+VRRP组网，S6800-1是主设备，S6800-2是备设备。



测试一：终端PING 10.1.1.252 ARP 交互过程

从终端S6820去ping S68-1的实IP 10.1.1.252，ARP交互过程如下：



Step1：S6820发送ARP请求报文，Hash到S68-2设备，S68-2学习MAC表项到DR接口；
Step2：ARP请求报文为广播，正常泛洪到IPL链路，到达S68-1设备；
Step3：由于本地有Vlan虚接口，ARP报文会copy一份上CPU，由于是DR接口收到的ARP报文,会将该ARP报文封装到Rlink Packet报文中发给S68-1设备；
Step4：S68-1收到两份ARP报文；S68-1设备根据收到的Rlink Packet将ARP表项学习到DR接口，IPP口收到的泛洪ARP报文不学习表项，并从DR口回复ARP Reply报文给S6820。

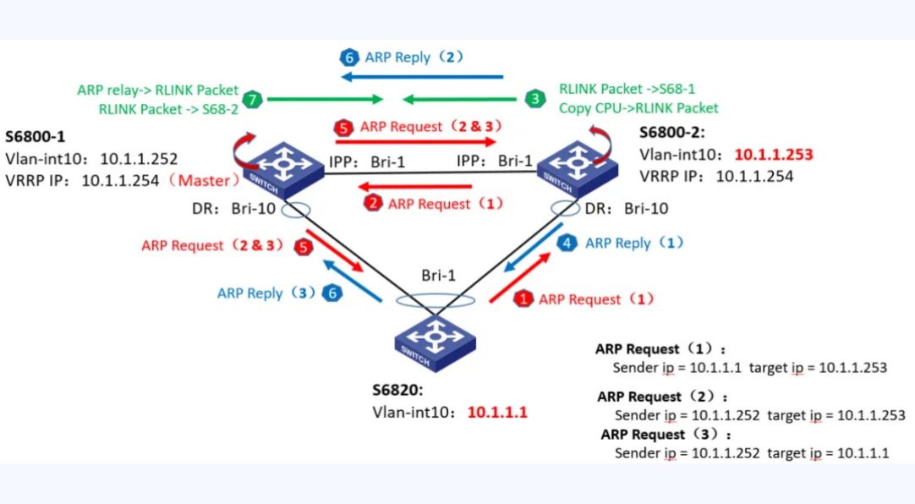
以上交互过程完成后，DR主备设备上的ARP表项情况分别如下：

```
S6800-1 ( ARP ) :
Type: S-Static   D-Dynamic   O-Openflow   R-Rule   M-Multiport  I-Invalid
IP address      MAC address     VLAN/VSI   Interface/Link ID      Aging  Type
10.1.1.1        4077-a9f0-38df    10       BAGG10                  243    D
S6800-2 ( ARP ) :
Type: S-Static   D-Dynamic   O-Openflow   R-Rule   M-Multiport  I-Invalid
IP address      MAC address     VLAN/VSI   Interface/Link ID      Aging Type
10.1.1.1        4077-a9f0-38df    10       BAGG10                  247    D
```

此时DRNI主备设备完成了ARP同步，两台设备均在DR聚合口BAGG10学到了终端ARP。

测试二：终端PING 10.1.1.253 ARP 交互过程

从终端S6820去ping S68-2的实地址10.1.1.253，ARP交互过程如下：



Step1：S6820发送ARP请求报文，Hash到S68-2设备，S68-2学习MAC表项到DR接口；
Step2：ARP请求报文为广播，正常泛洪到IPL链路，到达S68-1设备；
Step3：由于本地有Vlan虚接口，ARP报文会copy一份上CPU，由于是DR接口收到的ARP报文,会将该ARP报文封装到Rlink Packet报文中发给S68-1设备；
Step4：由于ARP请求的是S68-2的实地址，S68-2设备会本地回复ARP应答报文给S6820设备；
Step5： S68-1收到两份ARP报文， 根据收到的Rlink Packet将ARP表项学习到DR接口；由于S68-1收到10.1.1.1请求10.1.1.253的报文，此时会以10.1.1.252的地址为源，发起对10.1.1.1和10.1.1.253的ARP请求，都会从DR接口和IPL泛洪出去；
Step6：S68-2及S6820收到广播请求报文后，会各自回复ARP应答报文给S68-1设备；
Step7：由于S68-1收到10.1.1.1的应答报文是从DR接口收到的，也会封装该ARP应答报文为Rlink发给S68-2设备。

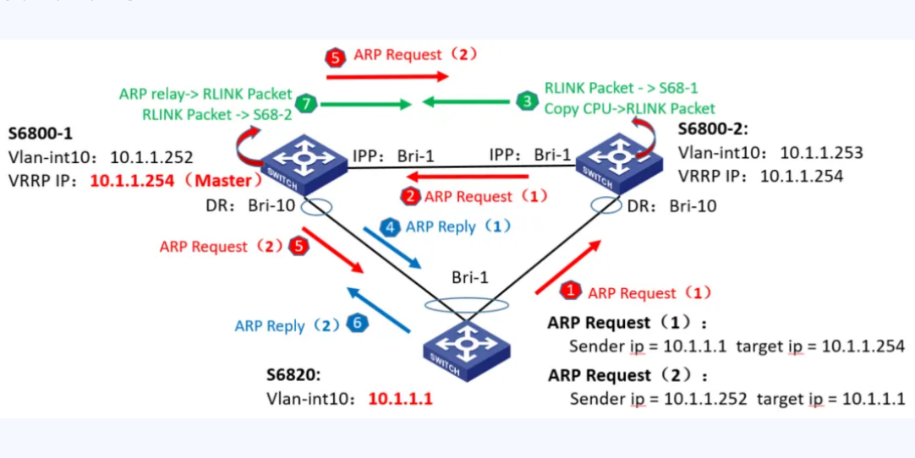
ARP交互完成后，终端及DR系统两台设备上的表项情况如下：

```
S6800-1 ( ARP ) :
 Type: S-Static   D-Dynamic   O-Openflow   R-Rule   M-Multiport  I-Invalid
IP address      MAC address     VLAN/VSI   Interface/Link ID      Aging Type
10.1.1.1       4077-a9f0-38e8      10      BAGG10                  277   D
10.1.1.253     9ce8-9572-a13d      10      BAGG1                   277   D
S6800-2 ( ARP ) :
Type: S-Static   D-Dynamic   O-Openflow   R-Rule   M-Multiport  I-Invalid
IP address      MAC address     VLAN/VSI   Interface/Link ID      Aging Type
10.1.1.1       4077-a9f0-38e8      10      BAGG10                  284   D
10.1.1.252     9ce8-9572-8339   10         BAGG1                   284   D
S6820 ( ARP ) :
 Type: S-Static   D-Dynamic   O-Openflow   R-Rule   M-Multiport  I-Invalid
IP address      MAC address     VLAN/VSI   Interface/Link ID      Aging Type
10.1.1.252     9ce8-9572-8339   10         BAGG1                   20   D
10.1.1.253     9ce8-9572-a13d   10         BAGG1                   20   D
```

可以看到，此时两台DR设备完成了ARP同步，且在IPP口学到了对端的实地址ARP。

测试三：终端PING 10.1.1.254 ARP 交互过程

我们再测试一下从终端ping VRRP虚地址10.1.1.254：



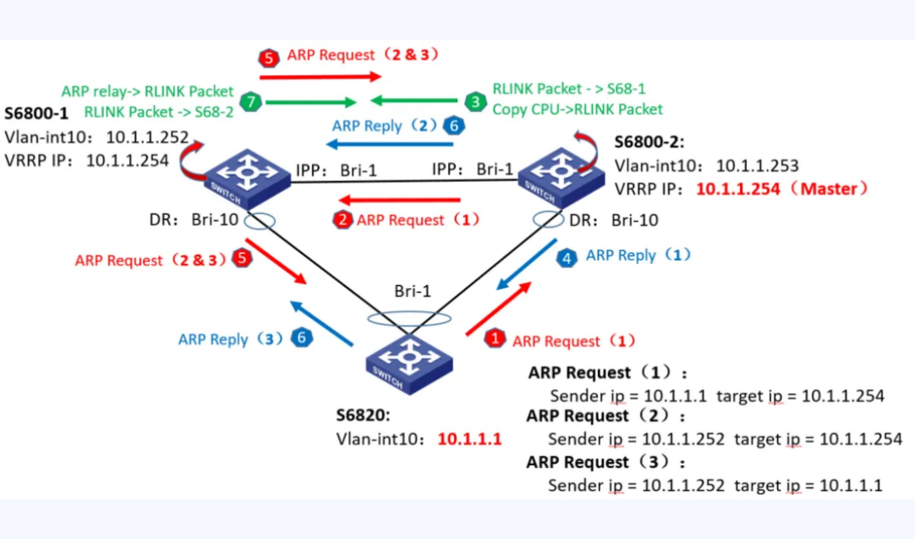
Step1：S6820发送ARP请求报文，Hash到S68-2设备，S68-2学习MAC表项到DR接口；
Step2：ARP请求报文为广播，正常泛洪到IPL链路，到达S68-1设备；
Step3：由于本地有Vlan虚接口，ARP报文会copy一份上CPU，由于是DR接口收到的ARP报文,会将该ARP报文封装到Rlink Packet报文中发给S68-1设备；
Step4：由于ARP请求的是S68-1的虚地址，S68-1设备会本地回复ARP应答报文给S6820设备；
Step5： S68-1收到两份ARP报文， 根据收到的Rlink Packet将ARP表项学习到DR接口；由于S68-1收到10.1.1.1请求10.1.1.254的ARP，会使用实地址发出ARP报文请求10.1.1.1，都会从DR接口和IPL泛洪出去；
Step6：S6820收到广播请求报文后，回复ARP应答报文给S68-1设备；
Step7：由于S68-1收到10.1.1.1的应答报文是从DR接口收到的，也会封装该ARP应答报文为Rlink发给S68-2设备。

交互完成后，各个设备上的表项如下：

```
S6800-1 ( ARP ) :
 Type: S-Static   D-Dynamic   O-Openflow   R-Rule   M-Multiport  I-Invalid
IP address      MAC address     VLAN/VSI   Interface/Link ID      Aging Type
10.1.1.1       4077-a9f0-38e8 10           BAGG10                  280   D
S6800-2 ( ARP ) :
Type: S-Static   D-Dynamic   O-Openflow   R-Rule   M-Multiport  I-Invalid
IP address      MAC address     VLAN/VSI   Interface/Link ID      Aging Type
10.1.1.1       4077-a9f0-38e8      10      BAGG10                  284   D
10.1.1.252     9ce8-9572-8339   10         BAGG1                   284   D
S6820 ( ARP ) :
 Type: S-Static   D-Dynamic   O-Openflow   R-Rule   M-Multiport  I-Invalid
IP address      MAC address     VID        Interface/Link ID      Aging Type
10.1.1.252     9ce8-9572-8339 10           BAGG1                   19   D
10.1.1.254     0000-5e00-0101 10           BAGG1                   20   D
```

测试四：VRRP主备倒换，PING 10.1.1.254 ARP交互过程

将S68-1的VRRP优先级调低，使S68-2成为VRRP主设备，此时从终端去ping虚地址10.1.1.254，ARP交互过程如下：



Step1：S6820发送ARP请求报文，Hash到S68-2设备，S68-2学习MAC表项到DR接口；
Step2：ARP请求报文为广播，正常泛洪到IPL链路，到达S68-1设备；
Step3：由于本地有Vlan虚接口，ARP报文会copy一份上CPU，由于是DR接口收到的ARP报文,会将该ARP报文封装到Rlink Packet报文中发给S68-1设备；
Step4：由于ARP请求的是虚地址，在S68-2上，S68-2设备会本地回复ARP应答报文给S6820设备；
Step5： S68-1收到两份ARP报文， 根据收到的Rlink Packet将ARP表项学习到DR接口；由于S68-1收到10.1.1.1请求10.1.1.254的ARP，此时会以10.1.1.252的地址为源，发起对10.1.1.1和10.1.1.254的ARP请求，都会从DR接口和IPL泛洪出去；
Step6： S68-2及S6820收到广播请求报文后，10.1.1.254在S68-2上，会各自回复ARP应答报文给S68-1设备；
Step7：由于S68-1收到10.1.1.1的应答报文是从DR接口收到的，也会封装该ARP应答报文为Rlink发给S68-2设备。

测试完成后，DR设备及终端上的ARP表项如下：

```
S6800-1 ( ARP ) :
  Type: S-Static   D-Dynamic   O-Openflow   R-Rule   M-Multiport  I-Invalid
IP address      MAC address     VLAN/VSI   Interface/Link ID      Aging Type
10.1.1.1       4077-a9f0-38e8 10           BAGG10                  278   D
10.1.1.254     0000-5e00-0101 10           BAGG1                   278   D
S6800-2 ( ARP ) :
Type: S-Static   D-Dynamic   O-Openflow   R-Rule   M-Multiport  I-Invalid
IP address      MAC address     VLAN/VSI   Interface/Link ID      Aging Type
10.1.1.1       4077-a9f0-38e8      10      BAGG10                  284   D
10.1.1.252     9ce8-9572-8339   10         BAGG1                   284   D
S6820 ( ARP ) :
Type: S-Static   D-Dynamic   O-Openflow   R-Rule   M-Multiport  I-Invalid
IP address      MAC address     VID        Interface/Link ID      Aging Type
10.1.1.252     9ce8-9572-8339 10           BAGG1                   19   D
10.1.1.254     0000-5e00-0101 10           BAGG1                   20   D
```

以上是整个DRNI组网中ARP同步机制的测试过程，简单粗暴总结就是，DR系统两台设备之间通过Rlink同步DR口收到的ARP报文，从而实现ARP表项同步，指导三层转发。

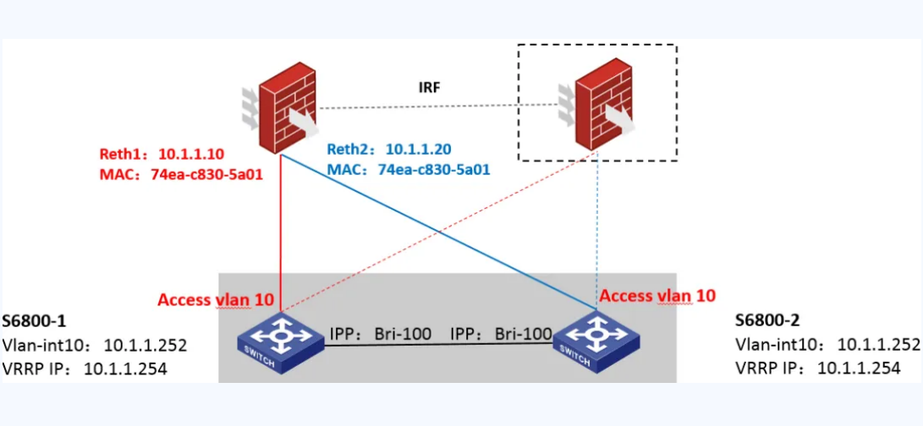
DRNI组网中的MAC表项同步机制

DRNI组网中的MAC表项同步机制如下：

- DRNI机制中IPP口不主动学习表项，包括MAC及ARP（单挂ARP学习到VLAN虚接口，出口是IPP口）；
- DRNI中MAC同步的是表项，与ARP同步报文实现不同；
- DRNI中DR接口学习的MAC表项同步到另一台设备的DR接口；
- DRNI中单挂口学习到的MAC表项会同步到另一台设备的IPP口；
- 单挂MAC同步原因：因为机制中IPP口不主动学MAC，单挂MAC不同步到对方IPP口的话，另一台DR到该单挂终端的流量就要泛洪转发

这里我们借由一个网上问题来说明DRNI组网中MAC表项的同步机制。

组网如图，防火墙做IRF，配置两个Reth口，分别单挂接入到DRNI设备上，正常情况备墙不工作，只有主墙的两个接口工作；两个Reth口的地址分别是10.1.1.10和10.1.1.20，MAC地址相同，两台DR设备与主墙互联使用相同的VLAN 10；防火墙接口为Route接口，不存在环路情况。



现场的故障现象是，一台DR设备会将直连的FW地址学习到自己的IPP接口，直连不通，另一台正常；当收到ARP更新消息后，两DR设备的ARP表项同时发生交替变化；本端直连ARP学习正常，另一台直连ARP学习到IPP口，另一台直连不通。

经过排查发现，其中一台DR设备上防火墙的MAC地址学习错误：

```
<S6800-2>display mac-address
74ea-c830-5a01   10   Learned    BAGG100  Y
// MAC地址错误学习到IPP口

<S6800-2>dis arp | include 10.1.1.
10.1.1.10       74ea-c830-5a01 10    BAGG100   283   D
//  10.1.1.10的表项学习正常，两设备到10.1.1.10都可以互通
10.1.1.20       74ea-c830-5a01 10    BAGG100   291   D
// 直连的10.1.1.20学习到IPP口，此时两DR只能与10.1.1.10互通
```

当ARP表项刷新时，原本MAC学习错误的DR设备直连ARP表项学习正常，另一台出现与上面表项一样的故障，直连不通。

出现这种故障现象的原因是什么呢？

现场的备防火墙不工作，主防火墙虽然是双链路分别连接DR系统两台设备，但实际上并不是链路聚合，对DR系统而言，主防火墙实际上是分别单挂在两台DR设备上的。在DRNI组网中，本端学习到单挂MAC后，会同步到对端的IPP口，但防火墙两个Reth口的MAC地址是相同的，两台DR设备同步MAC后，上行口和IPP口都会学习到相同的MAC地址，MAC地址发生迁移，会联动ARP表项迁移到新接口，由此导致了现场的故障现象。

那么这种情况应该如何解决呢？相信大家机智的小脑瓜已经想到了规避方法：

1. 调整FW的两个Reth口的MAC地址不一样（这样就不存在MAC迁移了）
2. DRNI设备上两个DR接口与FW互联的VLAN修改为不一样（相同MAC学习在不同的VLAN里面不是MAC迁移）

通过上面这个案例，大家对DRNI组网中MAC地址表的同步机制应该有了一定的了解，注意要和ARP同步机制区分开哦，MAC同步的是表项，ARP同步的是报文。

DRNI组网中ARP和MAC表项同步的机制比较复杂，最后再总结一下：

ARP同步机制

1. DRNI设备之间同步ARP时，同步的是ARP报文，不是ARP表项；
2. DR口收到接口ARP报文，原始报文按正常的转发逻辑正常从IPL和其他接口泛洪转发；同时会Copy到CPU，封装到RLink报文中同步给另一台DR设备；
3. IPP口会收到两种ARP报文，一种是RLink同步过来指向DR口的报文，一种是泛洪过来指向IPP口的报文，设备处理遵循DR口优先原则；
4. Rlink过来的ARP报文会被打上标记，设备针对这个带标记的ARP报文，不会回复ARP应答；
5. 单挂终端ARP报文不会被Rlink报文封装同步给另一台DR。

MAC同步机制

1. DRNI机制中IPP口不主动学习表项，包括MAC及ARP；
2. DRNI中MAC同步的是表项，与ARP同步报文实现不同；
3. DRNI中DR接口学习的MAC表项同步到另一台设备的DR接口；
4. DRNI中单挂口学习到的MAC表项会同步到另一台设备的IPP口；
5. 单挂MAC同步原因：机制中IPP口不主动学MAC，如果单挂MAC不同步到对方IPP口的话，另一台DR到该单挂终端的流量就要泛洪转发。

以上就是DRNI组网中ARP和MAC同步机制的全部内容啦，接下来我们还会继续介绍DRNI常见组网的具体配置，大家有什么疑问和需求可以留言告诉我们哦~

—— end ——