DRNI优化配置介绍

H3C交换机技术专题篇

DRNI

经过之前几期DRNI技术专题的学习，小伙伴们应该都对DRNI的原理、故障处理机制和几种典型组网的配置有了一定的了解。我们知道设备级冗余是DRNI的一大优点，无论是单链路故障还是成员设备单独升级，对承载业务的影响都微乎其微。如果想让网络拓扑改变对业务的影响降到更小，我们还需要注意一些配置细节。

本期技术专题我们将从细节入手，一起看看在DRNI组网中，有哪些可以加快收敛时间的优化配置~

01

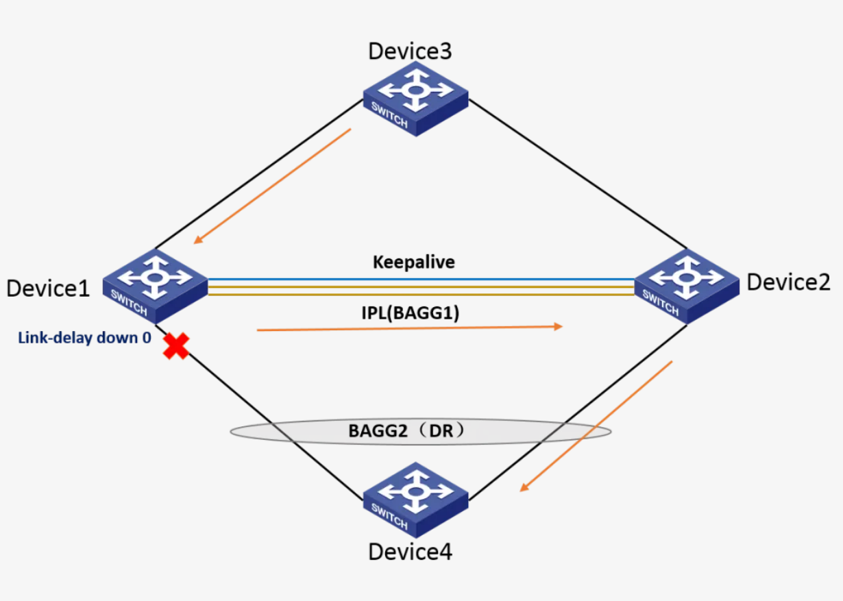
涉及流量转发的端口配置link-delay为0

link-delay命令用来配置接口物理连接状态抑制功能，如果不为0，接口物理状态变化时不会立刻上报CPU，而是会在配置的*delay-time*抑制时间后，再次检查接口状态，若还是变化的就上报CPU，若已经恢复则不上报。

```
【命令】
link-delay { down | up } [ msec ] delay-time
【参数】
delay-time：接口物理连接状态抑制时间值，0表示不抑制，即接口状态改变时立即上报CPU。
【举例】
# 设置以太网接口物理连接down状态抑制时间为8秒。
[Sysname] interface twenty-fivegige 1/0/1
[Sysname-Twenty-FiveGigE1/0/1]  link-delay  down 8
```

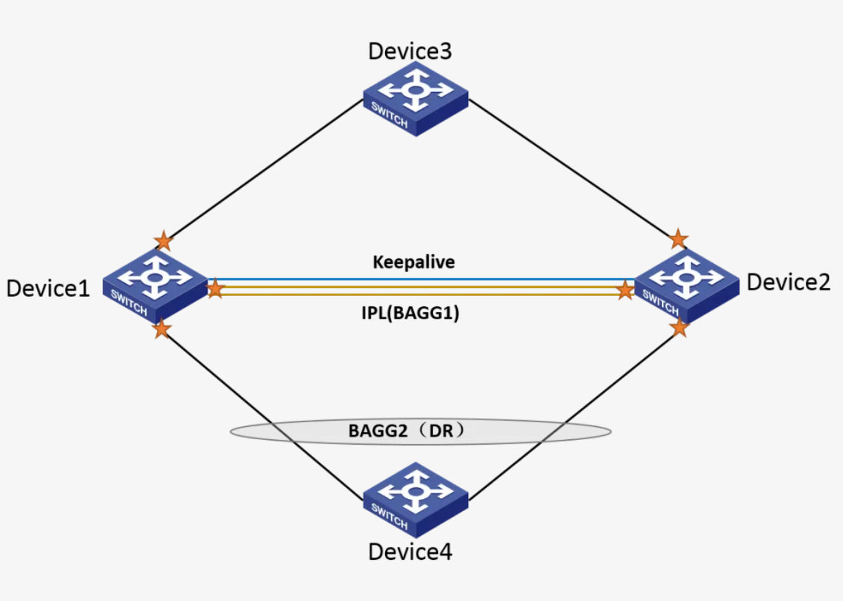
例如，我们在接口下配置link-delay down 1，这条命令表示该接口状态从up变成down时，不会立即上报CPU，而是在等待1s后，再次检查接口状态，如果状态仍然是down，就上报CPU；如果1s后端口状态为up，则不上报。

在DRNI组网配置中尤其需要关注汇聚层的交换机，部分汇聚交换机端口link-delay默认不是0，需要手工改成0加快收敛。这样当DRNI组网中涉及流量转发的端口状态发生变化后，设备CPU能第一时间感知并作出反应，比如下行DR物理口链路故障，CPU第一时间收到上报后，可以立刻将上行下来的流量切到IPL链路绕行。



但是在实际开局中，为了防止光传输或者光模块/光纤的抖动，很多客户会要求link-delay不能配置成0，这种情况下，考虑DRNI收敛时间合理性时，需要把link-delay时间考虑进去。例如配置link-delay 为1s时，设备需要1s后才能感知到端口故障，收敛时间也会相应增加*delay-time* 1s的时间。

那么在一个DRNI系统中，哪些端口需要修改link-delay为0呢？如下图所示，图中涉及流量转发的端口，如设备Device1和Device2的上下行口、横联的 IPP口都建议配置link-delay 为 0 。



02

配置drni restore-delay时间

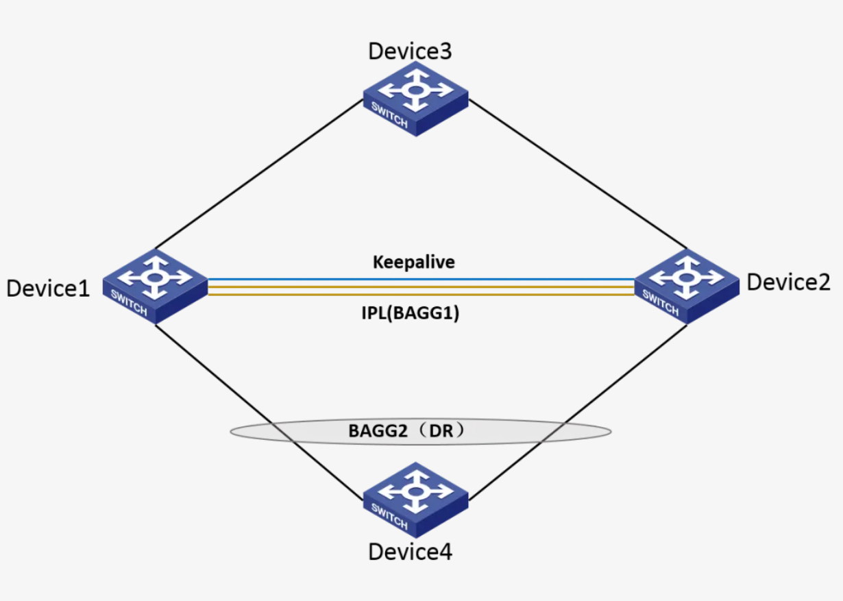
drni restore-delay命令用来配置延迟恢复时间。当设备作为从设备加入DRNI系统时，业务口（除DRNI保留接口以外的接口）状态为DRNI MAD DOWN，在配置延迟恢复时间之后，从设备业务口状态才会变为up。

```
【命令】
drni restore-delay value
【缺省情况】
定时器超时时间为30秒。
【参数】
value：定时器超时时间，取值范围1～3600，单位为秒。
【举例】
# 配置延迟恢复定时器超时时间为50秒。
[Sysname] drni restore-delay 50
```

配置延迟恢复时间有什么意义呢？当DR系统分裂或者成员设备重启时，从设备上除保留接口之外的接口都会被置为DRNI MAD DOWN状态，只有主设备进行业务转发，此时主备设备上的ARP\MAC表项也是不同步的。当故障恢复或者重启完成，IPP口重新up后，如果从设备的接口立刻up起来，由于ARP/MAC表项还未完全同步，可能出现转发异常。drni restore-delay时间就是在这种情况下预留给主从设备同步表项的时间，表项同步后，从设备接口再up起来，就可以根据表项正常转发业务流量了。

交换机drni restore-delay时间默认为30秒，定时器从IPL UP开始计时设备实际实现时，在1/2配置延迟恢复时间时会把DR逻辑聚合口放开，进行MAC和ARP同步，配置延迟恢复时间结束后，从设备的DR物理口开始up。如果不配置配置延迟恢复时间，或者配置过小，在IPP口从down到up的过程中丢包时间会增大。

例如下图中组网，假设Device1 设备是主设备，Device2 是从设备，drni restore-delay时间保持默认的30s，当device2加入DRNI系统时，IPP口up后，会预留30s的时间将Device1设备DR口的MAC和ARP表项同步给Device2，实现表项在底层的同步。



实际组网中，在表项较多或者进行ISSU升级时，可延长定时器时间，比如S125X设备ARP表项接近48K规格时，可以配置为drni restore-delay 900。另外在设备故障重启的场景中，drni restore-delay命令设置的延迟时间需要包括设备重启时间和表项恢复时间，请根据实际情况合理配置。

03

合理配置保留接口

IPL故障后，为了防止从设备继续转发流量，DRNI提供MAD（Multi-Active Detection，多Active检测）机制，即在DR系统分裂时将设备上部分接口置为DRNI MAD DOWN状态，不允许此类接口转发流量，避免流量错误转发，尽量减少对业务影响。DRNI MAD保留接口在DR系统分裂时不会被置于DRNI MAD DOWN状态，合理配置DRNI MAD保留接口可以加快收敛。

DRNI保留接口包括系统保留接口和用户配置的保留接口。系统保留接口不同设备可能不太一样，大家可以查阅具体设备的配置手册，大部分设备的系统保留接口包括：

- IPP口
- IPP口所对应的二层聚合接口的成员接口
- DR聚合逻辑口
- 管理以太网接口

用户配置的保留接口与具体组网和业务需求有关，**在不涉及EVPN的组网场景下：**

1、DR聚合口允许通过的vlan对应的vlan-interface要手工配置为保留接口。

这样是为了在drni restore-delay时间内能够同步该vlan内的ARP和MAC，否则设备只有在drni restore-delay时间后才能同步ARP、MAC表项。

2、Keep-alive链路需要手工配置成保留接口。

当Keepalive链路up，IPP口变为down时，会导致从设备上除IPP口、管理以太网口、保留接口以外的物理接口处于MAD DOWN状态。如果Keepalive链路的接口未配置为保留接口，该接口的状态将变为MAD DOWN，导致Keepalive链路down，认为邻居不存在，造成错误检测，所以Keepalive链路的接口需要配置为保留接口。

在涉及EVPN的场景下：

在EVPN+DRNI场景下，需要配置为保留接口的太多了（所有Vsi虚接口、loopback口等），为了减少配置工作量，可以先通过**drni mad default-action none**命令配置设备上的接口在DR系统分裂后保持原状态不变，然后再使用**drni mad include interface**命令配置DR系统分裂后需要处于DRNI MAD DOWN状态的接口。

那么哪些端口需要处于DRNI MAD DOWN状态呢？

1、DRNI系统上行接Spine/其他Leaf的物理接口需要MAD DOWN；

2、DRNI系统上的单挂AC口需要MAD DOWN；

3、DR成员口系统默认MAD DOWN。

推荐：

使用drni mad default-action none命令配置设备上的接口在DR系统分裂后保持原状态不变，然后使用drni mad include interface命令配置DR系统分裂后需要处于DRNI MAD DOWN状态的接口。

04

配置VRRP工作在非抢占方式

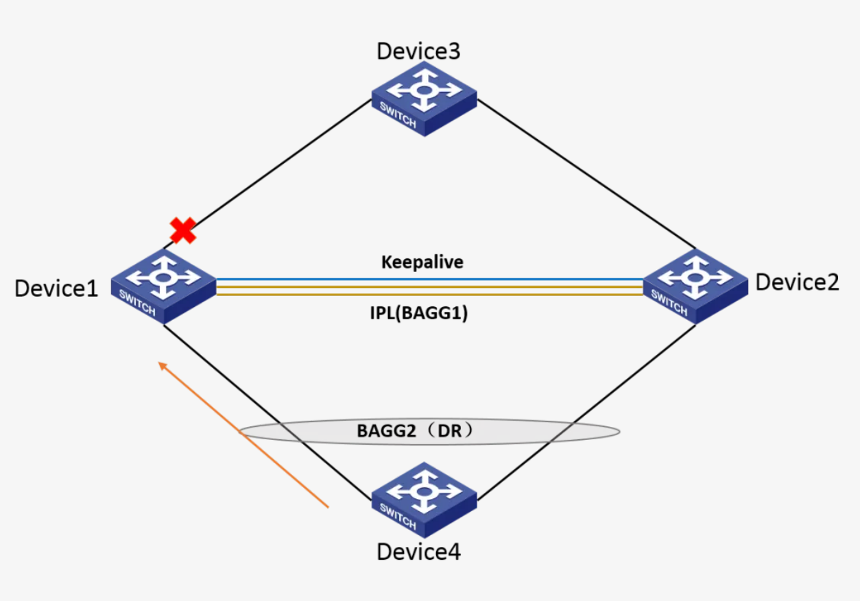
在DRNI+VRRP的组网中，考虑到DRNI备设备业务口被MAD DOWN时，如果VRRP主在DRNI备设备上，需要进行一次VRRP主备切换，可能影响业务，因此建议VRRP的主备与DRNI的主备能够保持一致。当DRNI主设备故障重启后变为备设备，那VRRP的主也会相应切换至另一台设备，如果VRRP工作在抢占模式，就会重新抢占为主，这样VRRP的主备与DRNI的主备就不一致了。为了防止这种情况，最好将VRRP配置为工作在非抢占方式下。

```
【命令】
vrrp vridvirtual-router-idpreempt-mode
undo vrrp vridvirtual-router-id preempt-mode
【缺省情况】
IPv4 VRRP备份组中的路由器工作在抢占方式下，抢占延迟时间为0厘秒。
【举例】
# 配置VRRP备份组1中的设备工作在非抢占方式
[Sysname] interface vlan-interface 2
[Sysname-Vlan-interface2]   undo   vrrp   vrid   1 preempt-mode
```

05

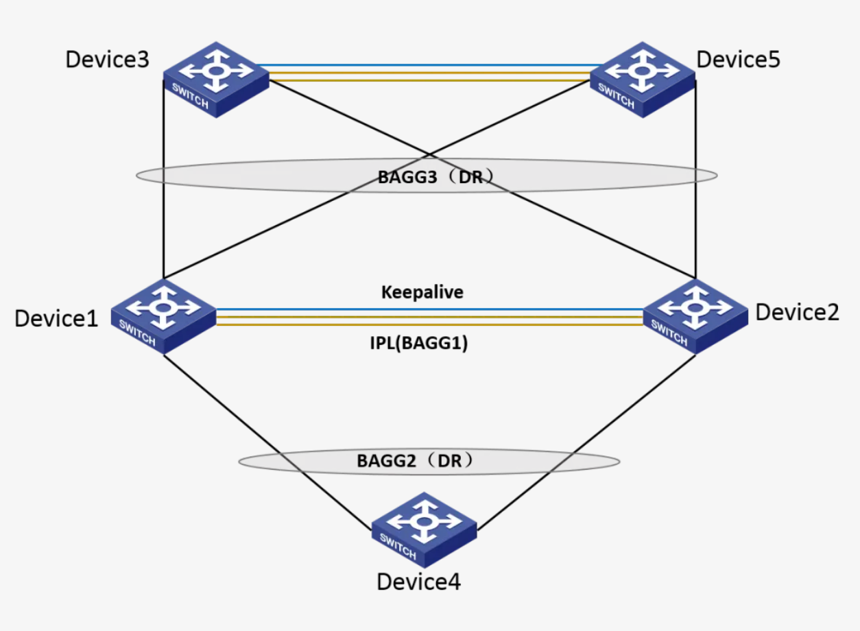
配置DR设备间三层路由互通

在之前几期中我们介绍过，如果DR系统两台设备间没有三层互通，当DR系统上行链路故障时，可以导致流量不通。以下图为例：



Device1和Device2组成DRNI系统，下行通过DR口对接Device4，上行通过三层路由对接Device3。如果Device1对接Device3的链路发生故障，Device4无感知，还是可能会把报文hash到Device1上，如果Device1和Device2间没有三层路由，那么报文就无法转发了。为了避免上行链路故障导致的业务不通，我们需要在DRNI系统两台设备之间起三层路由互通。

如果是多级DRNI，或者DRNI系统对上对下均是DR口对接，也可以不用配置DR设备间三层路由互通，以下图为例：



Device1和Device2的上下行口都配置为DR口，此时两台设备之间不需要三层互通，即使上下行链路故障，也可以直接依靠DRNI同步的表项进行跨IPL链路转发。

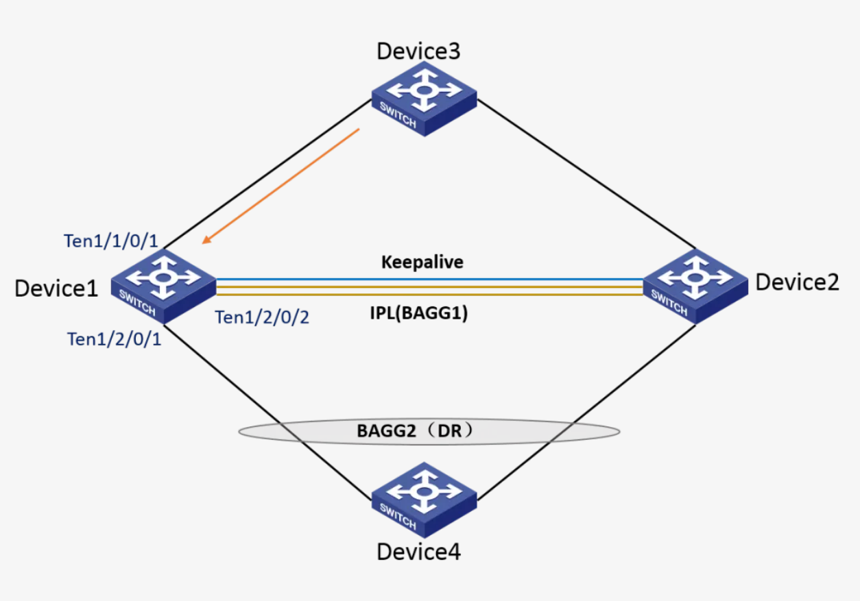
06

框式设备IPL跨板

对于框式设备配置DRNI的情况，建议保证每个业务板上都有端口加入IPP聚合口。这样不仅可以防止单块板卡故障导致DRNI分裂，还可以防止框式设备重启时单板normal时间不同导致的丢包。

设备重启时，由于不同业务板normal的时机不同，如果IPP口只在一块单板上，当IPP口所在单板normal时间晚于上行口所在单板时，就可能出现上行口已经up了，但IPP口和DR口还没有up的情况，此时上行设备如果将报文发给该设备，报文就会被丢弃。

以下图为例：



Device1和Device2都是框式设备，IPP口和DR口都在slot2上，上行口在slot1上。当Device1发生重启时，如果slot1先normal，slot2还在启动中，从上行Device3发到Device1上的报文，由于没有出接口up，就会被丢弃。

为了防止这种情况，最好保证每块单板上都有IPP聚合成员口，这样只要有单板normal，IPP口就可以up。

上面介绍的这种框式设备单板normal时间不同导致的业务不通，除了配置IPP口跨板聚合外，还可以通过monitor-link来规避，即当下行DR口down时，上行口联动保持为down，这样上行流量就不会发到重启过程中的设备上了。

Summary

本期介绍的六点是比较普适的DRNI收敛时间优化措施，在实际组网规划配置中，小伙伴们还是要具体组网具体分析，最好结合官网配置手册一起考虑。

之后我们会继续介绍DRNI的相关知识，有问题和建议欢迎留言哦~

—— end ——