某局点 S12508X-AF 2713设备做border 出向流量不均等

等价路由 **许家豪** 2022-05-07 发表

组网及说明

组网：S12508X-AF做Border，下行四条聚合链路到四台接入设备。

型号及版本：S12508X-AF 2713

问题描述

问题描述：boeder出方向流量ecmp到四台设备，但绝大部分流量只走了其中一条链路。

过程分析

查看诊断，发现设备路由表、fib表正常，端口状态正常。

1）缺省路由指向四个ecmp下一跳地址，路由表项正常

==========display ip routing-table==========

Destination/Mask Proto Pre Cost NextHop Interface

0.0.0.0/0 Static 60 0 22.247.129.185 RAGG10

22.247.130.185 RAGG11

22.247.131.185 RAGG12

22.247.132.185 RAGG13

2）fib表项存在4个下一跳，fib表项正常，border出方向到四台接入流量会从四个聚合口发出

==========display fib==========

Flag:

U:Useable G:Gateway H:Host B:Blackhole D:Dynamic S:Static

R:Relay F:FRR

Destination/Mask Nexthop Flag OutInterface/Token Label

0.0.0.0/0 22.247.129.185 USGR RAGG10 Null

0.0.0.0/0 22.247.130.185 USGR RAGG11 Null

0.0.0.0/0 22.247.131.185 USGR RAGG12 Null

0.0.0.0/0 22.247.132.185 USGR RAGG13 Null

3）查看4个聚合口，发现状态正常，物理口无错包，且RAGG10、12、13出方向存在流量，说明设备转发无异常，系出方向流量hash不均。

Route-Aggregation10

Current state: UP

Line protocol state: UP

Bandwidth: 20000000 kbps

…

Last clearing of counters: Never

Last 300 seconds input rate: 344196098 bytes/sec, 2753568784 bits/sec, 293320 packets/sec

Last 300 seconds output rate: 461 bytes/sec, 3688 bits/sec, 2 packets/sec

1229868796710 packets input, 1519504471426558 bytes, 0 drops

414772985 packets output, 117127975548 bytes, 0 drops

Route-Aggregation11

Current state: UP

Line protocol state: UP

Bandwidth: 20000000 kbps

…

Last clearing of counters: Never

Last 300 seconds input rate: 347567148 bytes/sec, 2780537184 bits/sec, 294616 packets/sec

Last 300 seconds output rate: 1185114531 bytes/sec, 9480916248 bits/sec, 985318 packets/sec

1305005829539 packets input, 1646712613329673 bytes, 0 drops

2303856336959 packets output, 1398568434514094 bytes, 0 drops

Route-Aggregation12

Current state: UP

Line protocol state: UP

Bandwidth: 20000000 kbps

…

Last clearing of counters: Never

Last 300 seconds input rate: 381647810 bytes/sec, 3053182480 bits/sec, 318532 packets/sec

Last 300 seconds output rate: 683 bytes/sec, 5464 bits/sec, 5 packets/sec

1203296471569 packets input, 1500229760508660 bytes, 0 drops

206007644 packets output, 32768301421 bytes, 0 drops

Route-Aggregation13

Current state: UP

Line protocol state: UP

Bandwidth: 20000000 kbps

…

Last clearing of counters: Never

Last 300 seconds input rate: 353941353 bytes/sec, 2831530824 bits/sec, 300559 packets/sec

Last 300 seconds output rate: 1126 bytes/sec, 9008 bits/sec, 9 packets/sec

1301731640239 packets input, 1634942752705068 bytes, 0 drops

2990604411 packets output, 39172508469 bytes, 0 drops

解决方法

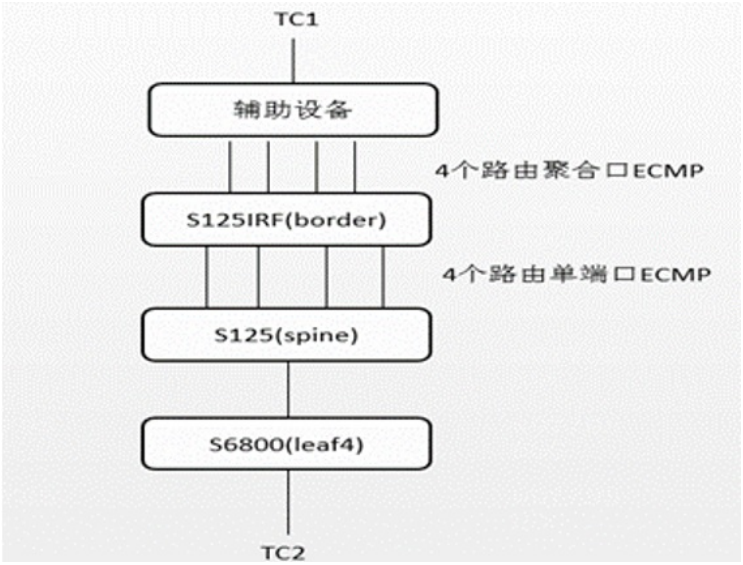
解决方发：可以通过配置ip load-sharing mode per-flow tunnel all global或者link-aggregation global load-sharing tunnel all命令，使设备同时根据内外层报文头部进行负载分担转发，两条命令效果相同，配置一条即可解决，后配置的命令生效。

4）经过确认，为保持软件实现方式的一致，该型号版本的设备默认情况下根据报文外层头部进行哈希，在VXLAN组网中，设备作为Border、leaf等角色对vxlan报文解封装后，报文无法根据VXLAN外层头部进行负载分担，因此出现上述流量哈希不均的现象，可修改配置为Tunnel all，使设备根据内外层报文头部进行hash，实现出方向流量的负载分担。

注意： 修改设备负载分担方式不会引发断流，只会使流量会重新计算负载出接口，哈希负载转发路径会发生变化。

实验室实测如下

1、 测试组网



2、 测试验证

流量目的地址是2.0.0.2，从路由表看，到改地址有4个下一跳：

```
[H3C]dis ip routing-table vpn-instance hs52rpmgr9nnb6k5sfvjvdhaj 2.0.0.0

Summary count : 5

Destination/Mask    Proto   Pre Cost        NextHop         Interface
0.0.0.0/0           Static  60  0           29.125.252.6    Vlan3
2.0.0.0/24          BGP     255 0           1.49.0.1        RAGG1
                                            1.50.0.1        RAGG2
                                            1.51.0.1        RAGG3
                                            1.52.0.1        RAGG4
```

测试仪打入变化的报文，缺省情况下，出方向流量只走了两个下一跳：

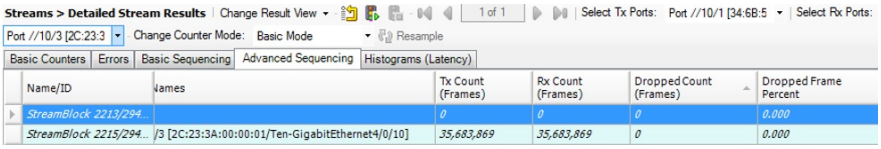
```
[H3C]dis counters rate outbound interface | in RAGG
RAGG1                       1         512577              --                --
RAGG2                       1         508598              --                --
RAGG3                       0              0              --                --
RAGG4                       0              0              --                --
```

全局配置ip load-sharing mode per-flow tunnel all global命令，流量可以从4个下一跳均HASH出去：

```
[H3C]ip load-sharing mode per-flow tunnel all global
```

```
[H3C]dis counters rate outbound interface | in RAGG
RAGG1                       0         258346              --                --
RAGG2                       0         254281              --                --
RAGG3                       0         254262              --                --
RAGG4                       0         254297              --                --
```

从测试仪观察流量，配置命令过程中无丢包：



3、 测试结论

配置ip load-sharing mode per-flow tunnel all global 命令后流量负载分担，且过程中无丢包。