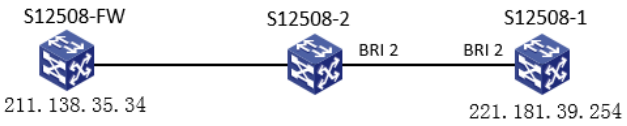
S12508由于配置URPF导致设备丢包案例分析

uRPF 丢包 程飞 2017-04-26 发表

如下拓扑图：S12508-1和S12508-2做VRRP，现场发现从S12508-FW这台设备跨S12508 - 02去ping S12508-01有大量丢包，丢包很规律，每五个包只会通一个。S12508-FW直连ping S12508-2不会丢包，S12508-2与S12508-1直连互ping也不丢包。并且业务一直也不受影响，就如下两个地址互ping有丢包:
从S12508-FW的本地地址（211.138.35.34）到S12508-1（221.181.39.254）



[12508-FW]ping -c 12 -a 211.138.35.34 221.181.39.254
Ping 221.181.39.254 (221.181.39.254): 56 data bytes, press CTRL_C to break
Request time out
Request time out
Request time out
Request time out
Request time out
56 bytes from 221.181.39.254: icmp_seq=0 ttl=255 time=8.305 ms
Request time out
Request time out
Request time out
Request time out
Request time out
56 bytes from 221.181.39.2549.1.1.2: icmp_seq=4 ttl=255 time=1.651 ms

--- Ping statistics for 221.181.39.254 ---
12 packet(s) transmitted, 2 packet(s) received, 83.3% packet loss

1、分别在S12508-2与S12508-1进行流量统计，可以确定S12508-1已经把icmp reply报文发回去了，但是在直连的S12508-2的入口流量统计发现报文变少了。如下流量统计：现场ping 20个包，只通了4个，在S12508-1的出口可以看到已经回复了20个报文，但是在S12508-2入口只统计到了4个。
acl number 3999
rule 0 permit icmp source 221.181.39.254 0 destination 211.138.35.34 0
rule 10 permit icmp source 211.138.35.34 0 destination 221.181.39.254 0

S12508-1的出口流量统计：
Interface: Ten-GigabitEthernet8/0/9
  Direction: Outboun
  Policy: liutong
   Classifier: liutong
     Operator: AND
     Rule(s) : If-match acl 3999
     Behavior: liutong
      Accounting Enable:
        20 (Packets)

S12508-2的入口流量统计：
  Interface: Ten-GigabitEthernet8/0/9
  Direction: Inbound
  Policy: liutong
   Classifier: liutong
     Operator: AND
     Rule(s) : If-match acl 3000
     Behavior: liutong
      Accounting Enable:

4 (Packets)

2、由于两台设备互连接口是聚合接口2，有四个物理接口，最开始怀疑是链路问题导致丢弃，因此把Ten8/0/9接口shutdown来排除该链路问题，但是shutdown该接口后，报文从Ten8/0/11转发，问题依旧，S12508-2的入口的报文总是比S12508-1出口少。

3、进行路由排查，S12508-02到S12508-01走的是直连路由，出口是interface vlan140，该VLAN虚接口配置了严格URPF，严格URPF会检查从这个接口发出去的报文是否还从这个接口进来。如果不从这个接口回来，那么报文将会被丢弃。

[H3C12508-02]dis ip routing-table 221.181.39.254
Routing Table : Public
Summary Count : 8
Destination/Mask Proto Pre Cost NextHop Interface
0.0.0.0/0 O_ASE 150 201 211.138.35.33 Vlan11
221.181.39.0/24 O_ASE 150 202 221.181.39.62 Vlan120
O_ASE 150 202 221.181.39.190 Vlan130
O_ASE 150 202 221.181.39.254 Vlan140
O_ASE 150 202 221.181.39.30 Vlan150
O_ASE 150 202 221.181.46.253 Vlan160
O_ASE 150 202 221.181.33.125 Vlan170
221.181.39.224/24 Direct 0 0 221.181.39.248 Vlan140

该互连vlan虚接口配置了严格URPF：
interface Vlan-interface140
ip address 221.181.39.248 255.255.255.224
vrrp vrid 140 virtual-ip 221.181.39.225
vrrp vrid 140 priority 110
vrrp vrid 140 track interface Vlan-interface11 reduced 20
ip urpf strict

右边设备S12508-1上回应报文走的是6条等价路由，而icmp reply报文是CPU发出的，CPU发出的报文是逐包转发，因此icmp reply报文是逐条路由进行回包，如果该报文正好走vlan-interface 140的路由，左边设备S12508-02可以URPF检查通过，ping则是正常的，如果走其他vlan-interface的路由，左边设备S12508-02的会URPF检查不通过，那么就会出现ping不通，所以表现就是每通1个包丢5个包：

[H3C12508-01]dis ip routing-table 211.138.35.34
Routing Table : Public
Summary Count : 7
Destination/Mask Proto Pre Cost NextHop Interface
0.0.0.0/0 O_ASE 150 201 211.138.35.41 Vlan11
211.138.35.32/29 OSPF 10 201 221.181.39.61 Vlan120
OSPF 10 201 221.181.39.189 Vlan130
OSPF 10 201 221.181.39.248 Vlan140
OSPF 10 201 221.181.39.29 Vlan150
OSPF 10 201 221.181.46.252 Vlan160
OSPF 10 201 221.181.33.124 Vlan170

4、对于这样的组网进行实验室复现。

如下实验室复现拓扑图，S12508与S95E-2通过四个VLAN接口互连，VLAN 30 40 50 60对应的虚接口开启URPF：

通过OSPF方式在S12508上发布4条等价路由到S95E-2，让S95E-2访问S95E-1的时候有4条等价路由，这个时候是每发送4个包，通一个包，丢三个包：

[S9500E-1]ping -c 10 30.0.0.2
PING 30.0.0.2: 56 data bytes, press CTRL_C to break
Request time out
Request time out
Request time out
Reply from 30.0.0.2: bytes=56 Sequence=3 ttl=254 time=1 ms
Request time out
Request time out

```
  Request time out
  Reply from 30.0.0.2: bytes=56 Sequence=7 ttl=254 time=1 ms
  Request time out
  Request time out
 --- 30.0.0.2 ping statistics ---
10 packet(s) transmitted,2 packet(s) received, 80.00% packet loss
 round-trip min/avg/max = 1/1/1 ms
```

从S12508直连ping S95E-2是不丢包的：

```
[S12508-V5] ping 30.0.0.2
  PING 30.0.0.2: 56  data bytes, press CTRL_C to break
    Reply from 30.0.0.2: bytes=56 Sequence=0 ttl=255 time=1 ms
    Reply from 30.0.0.2: bytes=56 Sequence=1 ttl=255 time=1 ms
    Reply from 30.0.0.2: bytes=56 Sequence=2 ttl=255 time=1 ms
    Reply from 30.0.0.2: bytes=56 Sequence=3 ttl=255 time=1 ms
    Reply from 30.0.0.2: bytes=56 Sequence=4 ttl=255 time=1 ms

 --- 30.0.0.2 ping statistics ---
    5 packet(s) transmitted ,5 packet(s) received ,0.00% packet loss
round-trip min/avg/max = 1/1/1 ms

[S12508-V5]dis ip routing-table
Routing Tables: Public
        Destinations : 28       Routes : 31
Destination/Mask    Proto  Pre  Cost        NextHop         Interface
1.1.1.1/32          Direct 0    0           127.0.0.1       InLoop0
10.0.0.0/24         Direct 0    0           10.0.0.2        Vlan10
10.0.0.2/32         Direct 0    0           127.0.0.1       InLoop0
30.0.0.0/24         Direct 0    0           30.0.0.1        Vlan30
……
```

互连接口进行URPF配置：

```
interface Vlan-interface30
 ip address 30.0.0.1 255.255.255.0
 ip urpf strict
interface Vlan-interface40
 ip address 40.0.0.1 255.255.255.0
 ip urpf strict
interface Vlan-interface50
 ip address 50.0.0.1 255.255.255.0
 ip urpf strict
interface Vlan-interface60
 ip address 60.0.0.1 255.255.255.0
 ip urpf strict
#
```

S95E-2到S95E-1的10.0.0.1地址走的ospf四条等价路由：

```
[S9500E-2]dis ip routing-table
Routing Tables: Public
        Destinations : 16       Routes : 22
Destination/Mask    Proto  Pre  Cost        NextHop         Interface
1.1.1.1/32          OSPF   10   100         50.0.0.1        Vlan50
10.0.0.0/24         OSPF   10   101         50.0.0.1        Vlan50
                    OSPF   10   101         60.0.0.1        Vlan60
                    OSPF   10   101         30.0.0.1        Vlan30
                    OSPF   10   101         40.0.0.1        Vlan40
```

在S95E-2上调整ospf cost值，不形成等价：

```
[S9500E-2-Vlan-interface30]ospf cost 10
[S9500E-2-Vlan-interface30]dis ip rou
[S9500E-2-Vlan-interface30]dis ip routing-table
Routing Tables: Public
        Destinations : 16       Routes : 16
Destination/Mask    Proto  Pre  Cost        NextHop         Interface
1.1.1.1/32          OSPF   10   10          30.0.0.1        Vlan30
10.0.0.0/24         OSPF   10   11          30.0.0.1        Vlan30
```

30.0.0.0/24 Direct 0 0 30.0.0.2 Vlan30
30.0.0.2/32 Direct 0 0 27.0.0.1 InLoop0
40.0.0.0/24 Direct 0 0 40.0.0.2 Vlan40

在S95E-1上测试，恢复正常：
[S9500E-1]ping -c 10 30.0.0.2
PING 30.0.0.2: 56 data bytes, press CTRL_C to break
Reply from 30.0.0.2: bytes=56 Sequence=0 ttl=254 time=1 ms
Reply from 30.0.0.2: bytes=56 Sequence=1 ttl=254 time=1 ms
Reply from 30.0.0.2: bytes=56 Sequence=2 ttl=254 time=2 ms
Reply from 30.0.0.2: bytes=56 Sequence=3 ttl=254 time=1 ms
Reply from 30.0.0.2: bytes=56 Sequence=4 ttl=254 time=1 ms
Reply from 30.0.0.2: bytes=56 Sequence=5 ttl=254 time=1 ms
Reply from 30.0.0.2: bytes=56 Sequence=6 ttl=254 time=1 ms
Reply from 30.0.0.2: bytes=56 Sequence=7 ttl=254 time=1 ms
Reply from 30.0.0.2: bytes=56 Sequence=8 ttl=254 time=1 ms
Reply from 30.0.0.2: bytes=56 Sequence=9 ttl=254 time=2 ms
--- 30.0.0.2 ping statistics ---
10 packet(s) transmitted, 10 packet(s) received, 0.00% packet loss
round-trip min/avg/max = 1/1/2 ms

由于现场互连接口配置了严格URPF，并且报文出入不在同一个接口，导致报文丢弃。因此关闭URPF检查可以解决。但是由于现场要防止有异常流量攻击，因此不能去掉URPF检查，因此现场调整HJ-H3C12508-01设备上的等价路由，把出口为vlan140的OSPF路由的cost值调小，到211.138.35.34的路由不走等价，而走这条vlan140出口的最优路由，这样问题也可以解决。
[H3C12508-01]dis ip routing-table 211.138.35.34
Routing Table : Public
Summary Count : 7
Destination/Mask Proto Pre Cost NextHop Interface
0.0.0.0/0 O_ASE 150 201 211.138.35.41 Vlan11
OSPF 10 100 221.181.39.248 Vlan140//将该路由cost值调小，让现场不再是等价路由

以后遇到两台互连设备之间进行流量统计,发现对端流量统计发出了报文，本地却没有流量统计到、没有镜像到，或者统计计数减少、镜像到的报文减少了，那么就要看看接口是否配置了URPF,被URPF丢弃的报文是不会被流量统计到的。
注：对于S12500/S9500E系列交换机，对于上送CPU的报文是流量统计不到的。如果现在没有配置URPF的情况，那么真的需要检查一下链路是否正常，接口是否有错包，链路问题也会导致本端入口流量统计接口计数比对端出口流量统计计数计数少的情况。