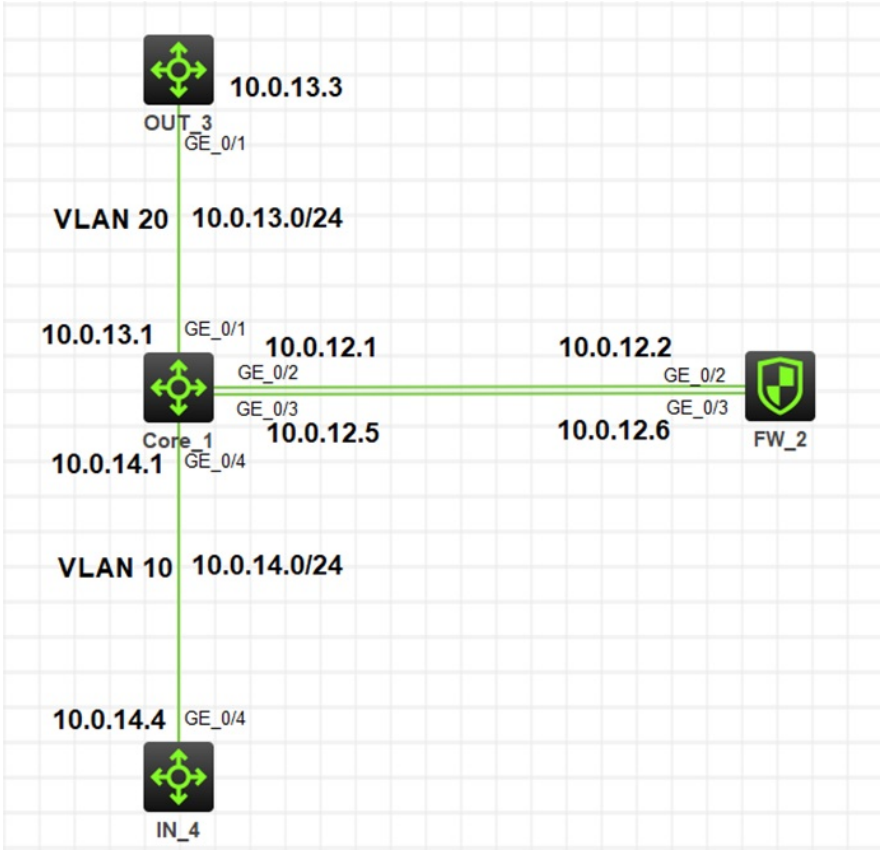
交换机快转负载分担导致防火墙策略路由方式旁路部署数据不通

域间策略/安全域 设备部署方式 **薛佳宇** 2022-07-19 发表

组网及说明

故障拓扑：



问题描述

故障现象：内网10.0.14.4无法ping通上行10.0.13.3

关键设备配置：

Core：

```
#
vlan 10
#
vlan 20
#
interface GigabitEthernet1/0/2
 port link-mode route
 ip address 10.0.12.1 255.255.255.252
#
interface GigabitEthernet1/0/3
 port link-mode route
 ip address 10.0.12.5 255.255.255.252
#
interface GigabitEthernet1/0/1
 port link-mode bridge
 port link-type trunk
 port trunk permit vlan 1 20
#
interface GigabitEthernet1/0/4
 port link-mode bridge
 port link-type trunk
 port trunk permit vlan 1 10
#
 ip route-static 0.0.0.0 0 10.0.13.3
#
#通过两个acl分别匹配vlan10 去vlan20的包和vlan20给vlan10的回包
acl advanced 3000
 description ToInternet
 rule 0 permit ip source 10.0.14.0 0.0.0.255
#
acl advanced 3005
 description InternetBack
 rule 0 permit ip destination 10.0.14.0 0.0.0.255
#
#配置策略路由，对vlan-interface10收到的vlan10访问vlan20的数据，更改下一跳地址为防火墙的1/0/3口，并在vlan-interface10 应用策略路由
policy-based-route ToInternet permit node 10
 if-match acl 3000
 apply next-hop 10.0.12.6
#
#配置策略路由，对vlan-interface20收到的vlan20回应vlan10的数据，更改下一跳地址为防火墙的1/0/2口，并在vlan-interface20 应用策略路由
policy-based-route InternetBack permit node 10
 if-match acl 3005
 apply next-hop 10.0.12.2
#
interface Vlan-interface10
 ip address 10.0.14.1 255.255.255.0
 ip policy-based-route ToInternet
#
interface Vlan-interface20
 ip address 10.0.13.1 255.255.255.0
 ip policy-based-route InternetBack
#
```

FW：

```
#
interface GigabitEthernet1/0/2
 port link-mode route
 combo enable copper
 ip address 10.0.12.2 255.255.255.252
#
interface GigabitEthernet1/0/3
 port link-mode route
 combo enable copper
 ip address 10.0.12.6 255.255.255.252
#
ip route-static 0.0.0.0 0 10.0.12.1
#
ip route-static 10.0.14.0 24 10.0.12.5
#
security-zone name Trust
 import interface GigabitEthernet1/0/3
#
security-zone name DMZ
#
security-zone name Untrust
 import interface GigabitEthernet1/0/2
#
security-policy ip
 rule 1 name pass
  action pass
  source-zone trust
  destination-zone untrust
```

过程分析

1、从核心交换机和防火墙的配置来看，现场是在交换机上下行接口配置策略路由将内网网段10.0.14.0/24的来回流量抛给防火墙处理

2、故障流量源地址10.0.14.4，目的地址10.0.13.3，故障期间查看防火墙会话信息：

```
<IN_4>ping 10.0.13.3                    //只ping了5个包
Ping 10.0.13.3 (10.0.13.3): 56 data bytes, press CTRL_C to break
Request time out
Request time out
Request time out
Request time out
Request time out
--- Ping statistics for 10.0.13.3 ---
5 packet(s) transmitted, 0 packet(s) received, 100.0% packet loss
<FW>dis session table ipv4 source-ip 10.0.14.4 destination-ip 10.0.13.3 protocol
icmp verbose
Slot 1:
Initiator:
#Source      IP/port: 10.0.14.4/228
 Destination IP/port: 10.0.13.3/2048
 DS-Lite tunnel peer:
#VPN instance/VLAN ID/Inline ID: -/-/-
 Protocol: ICMP(1)
#Inbound interface: GigabitEthernet1/0/3
 Source security zone: Trust
Responder:
 Source      IP/port: 10.0.13.3/228
 Destination IP/port: 10.0.14.4/0
 DS-Lite tunnel peer:
#VPN instance/VLAN ID/Inline ID: -/-/-
 Protocol: ICMP(1)
 Inbound interface: GigabitEthernet1/0/2
 Source security zone: Untrust
State: ICMP_REQUEST
Application: ICMP
Rule ID: 1
Rule name: pass
Start time: 2022-07-19 17:09:43  TTL: 53s
Initiator->Responder:          635 packets       53340 bytes
Responder->Initiator:            0 packets           0 bytes
Total sessions found: 1
<FW>
```

可以看到，明明终端只ping了4个包，但是防火墙会话统计中正向流量的报文数却有635个，明显有异常

3、流量沿途都是H3C设备，开启ip ttl-expires enable和ip unreachables enable后可在终端上tracert查看流量路径

```
<IN_4>tracert 10.0.13.3
traceroute to 10.0.13.3 (10.0.13.3), 30 hops at most, 40 bytes each packet, press CTRL_C to break
 1  10.0.14.1 (10.0.14.1)  0.000 ms  0.000 ms  1.000 ms
 2  10.0.12.6 (10.0.12.6)  0.000 ms  0.000 ms  1.000 ms
 3  10.0.12.1 (10.0.12.1)  1.000 ms  0.000 ms  0.000 ms
 4  10.0.12.6 (10.0.12.6)  1.000 ms  1.000 ms *
 5  * * *
 6  * * 10.0.12.6 (10.0.12.6)  2.000 ms
 7  10.0.12.1 (10.0.12.1)  2.000 ms  2.000 ms  1.000 ms
 8  10.0.12.6 (10.0.12.6)  1.000 ms  1.000 ms  1.000 ms
 9  10.0.12.1 (10.0.12.1)  1.000 ms
<IN_4>
```

可以看到故障流量在核心交换机上来回跑，所以导致了不通，为何会这样呢？

解决方法

4、交换机是硬件转发设备，而当一条数据流的第一个报文通过查找路由表转发后，在高速缓存中生成相应的转发表项，快速转发就可以通过直接查找快速转发表进行转发。而H3C交换机默认开启了快速转发负载分担，具体为：开启快速转发负载分担功能后，当一条数据流从不同入接口上来进行转发时，不再根据入接口不同区分数据流，根据五元组标识一条数据流。Ping测试时查看交换机快转表：

[Core_1]dis ip fast-forwarding cache

Total number of fast-forwarding entries: 4

SIP	SPort	DIP	DPort	Pro	Input_If	Output_If	Flg
10.0.13.3	230	10.0.14.4	0	1	GE1/0/3	N/A	1
10.0.14.4	230	10.0.13.3	2048	1	Vlan10	GE1/0/3	11
10.0.14.4	0	10.0.12.1	0	1	Vlan10	N/A	1
10.0.12.1	0	10.0.14.4	2816	1	InLoop0	Vlan10	1

5、根据上述原理分析：当10.0.14.4访问10.0.13.3时，因为是三层访问，数据先到网关核心交换机(vlan10)，命中vlan-int10下策略路由后经过G1/0/3发送给防火墙，此时交换机形成快转表：

SIP	SPort	DIP	DPort	Pro	Input_If	Output_If	Flg
10.0.14.4	230	10.0.13.3	2048	1	Vlan10	GE1/0/3	11

当防火墙处理完流量后，匹配到默认路由再将数据发给核心交换机的G1/0/2。默认情况下交换机开启快转负载分担，此时设备不根据入接口不同区分数据流，仅根据五元组(源目ip、源目端口和协议号)标识一条数据流，所以防火墙转发给交换机的数据会命中前面的快转表，就会再发给防火墙，防火墙处理完后再匹配默认路由发给核心交换机的G1/0/2，数据再匹配核心交换机快转表发给防火墙……这个分析跟tracert的结果也是一致的

6、尝试关闭快速转发负载分担，然后ping测试，再查看交换机快转表：

关快转负载分担：

[Core_1]undo ip fast-forwarding load-sharing

<IN_4>ping 10.0.13.3

Ping 10.0.13.3 (10.0.13.3): 56 data bytes, press CTRL_C to break

56 bytes from 10.0.13.3: icmp_seq=0 ttl=252 time=2.000 ms

56 bytes from 10.0.13.3: icmp_seq=1 ttl=252 time=1.000 ms

56 bytes from 10.0.13.3: icmp_seq=2 ttl=252 time=2.000 ms

56 bytes from 10.0.13.3: icmp_seq=3 ttl=252 time=2.000 ms

56 bytes from 10.0.13.3: icmp_seq=4 ttl=252 time=2.000 ms

--- Ping statistics for 10.0.13.3 ---

5 packet(s) transmitted, 5 packet(s) received, 0.0% packet loss

此时10.0.14.4已经可以ping通10.0.13.3了

[Core_1]dis ip fast-forwarding cache

Total number of fast-forwarding entries: 4

SIP	SPort	DIP	DPort	Pro	Input_If	Output_If	Flg
10.0.13.3	231	10.0.14.4	0	1	GE1/0/3	Vlan10	1
10.0.13.3	231	10.0.14.4	0	1	Vlan20	GE1/0/2	11
10.0.14.4	231	10.0.13.3	2048	1	GE1/0/2	Vlan20	1
10.0.14.4	231	10.0.13.3	2048	1	Vlan10	GE1/0/3	11

可以看到交换机上也对同一条流区分出入接口后生成了两条快转表。