H3C SecPath M9000系列NAT地址池中地址数目与业务板数目限制的原理说明

业务板 NAT 胡伟 2021-11-23 发表

组网及说明

组网无

问题描述

在H3C SecPath M9000系列的版本说明书里面关于NAT地址池有这样的限制说明：

> SecBladeIII 和单逻辑芯片的SecBladeIV NAT地址池中的地址数必须大于或等于备份组个数，双逻辑芯片的SecBladeIV NAT地址池中的地址数必须大于或者等于备份组个数*2。这里的备份组特指自动备份组和手动备份组，且整机环境中所有业务板使用的备份组类型必须一致，不允许自动备份组和手动备份组混用。支持双逻辑改成单逻辑命令，重启业务板生效，命令行：hardware fast-forwarding standalone

至于为什么有这样的限制，是由于H3C SecPath M9000系列作为分布式设备以及业务板插卡的独有特性决定的。请见如下分析：

过程分析

正常来说 H3C SecPath M9000系列由机框、主控板、业务板、接口板、 网板（M9000-S系列网板集成在主控上）。

所有的业务板对内呈现是一个Blade聚合口的形式，所有的流量以缺省的源目地址HASH的方式分散到不同的Blade成员接口，进而上送到对应业务板。

这样可以保证不做NAT的同一个业务的来回流量可以上到同一个业务板进行处理，避免了流量来回路径不一致场景的产生。

如下图所示，设备单机存在四块业务板，型号为NSQM1FWDFG0，这是一款带有2颗FPGA芯片（FG标识）的四代（D标识）防火墙(FW标识)业务板。

```
RBM_P<M9010_1>display device
Slot No. Brd Type         Brd Status   Subslot Sft Ver                Patch Ver
 0       NSQ1CGC2SE0      Normal       0       M9010-9153P3001        None
 1       NSQ1TGS32SF0     Normal       0       M9010-9153P3001        None
 2       NSQM1TGS32QSSG0  Normal       0       M9010-9153P3001        None
 3       NONE             Absent       0       NONE                   None
 4       NSQ1SUPB0        Master       0       M9010-9153P3001        None
 5       NSQ1SUPB0        Standby      0       M9010-9153P3001        None
 6       NSQM1FWDFG0      Normal       0       M9010-9153P3001        None
         CPU 1            Normal       0       M9010-9153P3001
 7       NSQM1FWDFG0      Normal       0       M9010-9153P3001        None
         CPU 1            Normal       0       M9010-9153P3001
 8       NSQM1FWDFG0      Normal       0       M9010-9153P3001        None
         CPU 1            Normal       0       M9010-9153P3001
 9       NSQM1FWDFG0      Normal       0       M9010-9153P3001        None
         CPU 1            Normal       0       M9010-9153P3001
 10      NSQ1FAB08D0      Normal       0       M9010-9153P3001        None
 11      NSQ1FAB08D0      Normal       0       M9010-9153P3001        None
 12      NSQ1FAB08D0      Normal       0       M9010-9153P3001        None
 13      NSQ1FAB08D0      Normal       0       M9010-9153P3001        None
```

```
RBM_P[M9010_1]display blade-controller-team Default
ID: 1    Name: Default
  Slot    CPU    Status    LBGroupID
  6       1      Normal        1
  7       1      Normal        1
  8       1      Normal        1
* 9       1      Normal        1

* : Primary blade controller of the team.

Load balancing group information for the blade controller team:
  LBGroupID   Name                BLAGG
  1           Blade4fw1           Blade-Aggregation257
```

这四块业务板在同一个缺省引擎组中，对内集成为一个Blade-Aggregation1来处理业务（其实还有一个Blade-Aggregation257，这个是形式上的接口，不处理业务，可以忽略不记）。如下图所示，每一块业务板都有两个Blade接口，序号为1口和3口(Blade-Aggregation1 成员接口使用奇数号标注），分别代表Blade板的两个通道channel0和channel1。当业务板为单FPGA和无FPGA的纯CPU转发时，则只有一个序号为1的Blade接口。

```
RBM_P[M9010_1]display link-aggregation verbose Blade-Aggregation 1
Loadsharing Type: Shar -- Loadsharing, NonS -- Non-Loadsharing
Port Status: S -- Selected, U -- Unselected, I -- Individual
Port: A -- Auto port
Flags:  A -- LACP_Activity, B -- LACP_Timeout, C -- Aggregation,
        D -- Synchronization, E -- Collecting, F -- Distributing,
        G -- Defaulted, H -- Expired

Aggregate Interface: Blade-Aggregation1
Aggregation Mode: Static
Loadsharing Type: Shar
  Port             Status  Priority Oper-Key
--------------------------------------------------------------------------------
  Blade6/0/1       S       32768    2
  Blade6/0/3       S       32768    2
  Blade7/0/1       S       32768    2
  Blade7/0/3       S       32768    2
  Blade8/0/1       S       32768    2
  Blade8/0/3       S       32768    2
  Blade9/0/1       S       32768    2
  Blade9/0/3       S       32768    2
```

重点来了，在开启NAT Outbound的场景下，则NAT地址池中 地址数目至少要8个，否则无法满足业务板的需求。

如果是NAT地址池中地址数目少于8个，则部分业务板的channel将无法获取地址（一个channel至少要满足一个NAT地址），导致在这个channel中处理的流量将无法NAT地址转换。

如下所示，当NAT地址池中只有3个地址时，只有slot6cpu1的channel0、channel1和slot7cpu1的channel0能获取到NAT地址，其他channel都无法获取地址，会导致上到这些channel的流量无法进行NAT转换。

```
RBM_P[M9010_1-address-group-0]display this
#
nat address-group 0
 address 111.0.0.1 111.0.0.3
#
return
```

```
RBM_P[M9010_1-probe]display system internal nat slot 6 cpu 1
NAT address group information:
  Totally 2 NAT address groups.
  Address group 0
    Port range: 1-65535
    Address information(address model is master-standby):
      Start address          End address
      111.0.0.1              111.0.0.1         (channel 0, ofpid 4026531889, active 1)
  Address group [illegible]
    Port range: 1-65535
    Address [illegible]
      Start address          End address
      11[illegible]

NAT easy-ip port-range information:
    Totally 2 NAT easy-ip port-range rules.
    Port-range : 1025 to 9088  Channel ID: 0
    Port-range : 9089 to 17152  Channel ID: 1

NAT global-policy information:
  Totally 4 NAT global-policy rules.
  Accelerated status: Y
  Rule name: test
RBM_P[M9010_1-probe]display system internal nat slot 7 c
RBM_P[M9010_1-probe]display system internal nat slot 7 cpu 1
NAT address group information:
  Totally 1 NAT address groups.
  Address group 0
    Port range: 1-65535
    Address information(address model is master-standby):
      Start address          End address
      111.0.0.3              111.0.0.3         (channel 0, ofpid 4026531897, active 1)

NAT easy-ip port-range information:
    Totally 2 NAT easy-ip port-range rules.
    Port-range : 17153 to 25216  Channel ID: 0
    Port-range : 25217 to 33280  Channel ID: 1
```

SecBladeIII 和单逻辑芯片的SecBladeIV NAT地址池中的地址数必须大于或等于备份组个数，双逻辑芯片的SecBladeIV NAT地址池中的地址数必须大于或者等于备份组个数*2。这里的备份组特指自动备份组和手动备份组，且整机环境中所有业务板使用的备份组类型必须一致，不允许自动备份组和手动备份组混用。支持双逻辑改成单逻辑命令，重启业务板生效，命令行：hardware fast-forwarding standalone

根据ofpip可以通过以下命令查看对应的流表group entry信息(可以理解为流表的action)

【probe视图下】display system internal openflow instance inner-redirect group

```
RBM_P[M9010_1-probe]display system internal openflow instance inner-redirect gro
up 4026531889
Group entry 4026531889:
 Type: All, byte count: --, packet count: --
 Bucket 1 information:
  Action count 1, watch port: Blade6/0/1, watch group: any
  Byte count --, packet count --
  Output interface: Blade6/0/1
 Group DrvContext: [7][4294967295][4294967295][4294967295]
 Referenced information:
  Count: 1
  Flow table: 200
  Flow entry: 43

RBM_P[M9010_1-probe]display system internal openflow instance inner-redirect gro
up 4026597425
Group entry 4026597425:
 Type: All, byte count: --, packet count: --
 Bucket 1 information:
  Action count 1, watch port: Blade6/0/3, watch group: any
  Byte count --, packet count --
  Output interface: Blade6/0/3
 Group DrvContext: [8][4294967295][4294967295][4294967295]
 Referenced information:
  Count: 1
  Flow table: 200
  Flow entry: 44

RBM_P[M9010_1-probe]display system internal openflow instance inner-redirect gro
up 4026531897
Group entry 4026531897:
 Type: All, byte count: --, packet count: --
 Bucket 1 information:
  Action count 1, watch port: Blade7/0/1, watch group: any
  Byte count --, packet count --
  Output interface: Blade7/0/1
 Group DrvContext: [11][4294967295][4294967295][4294967295]
 Referenced information:
  Count: 1
  Flow table: 200
  Flow entry: 45
```

根据流表group entry 中显示的flow table id来进一步查看详细flow-table。

【probe视图下】display system internal openflow instance inner-redirect flow-table

```
Flow entry 43 information:
 cookie: 0x0, priority: 5300, hard time: 0, idle time: 0, flags: check_overlap
 |reset_counts|no_pkt_counts|no_byte_counts, byte count: --, packet count: --
Match information:
 IP Range: IPv4 destination address from 111.0.0.1 to 111.0.0.1
 Input interfaces:  XGE2/0/16
Instruction information:
 Write actions:
  Group: 4026531889

Flow entry 44 information:
 cookie: 0x0, priority: 5300, hard time: 0, idle time: 0, flags: check_overlap
 |reset_counts|no_pkt_counts|no_byte_counts, byte count: --, packet count: --
Match information:
 IP Range: IPv4 destination address from 111.0.0.2 to 111.0.0.2
 Input interfaces:  XGE2/0/16
Instruction information:
 Write actions:
  Group: 4026597425

Flow entry 45 information:
 cookie: 0x0, priority: 5300, hard time: 0, idle time: 0, flags: check_overlap
 |reset_counts|no_pkt_counts|no_byte_counts, byte count: --, packet count: --
Match information:
 IP Range: IPv4 destination address from 111.0.0.3 to 111.0.0.3
 Input interfaces:  XGE2/0/16
Instruction information:
 Write actions:
  Group: 4026531897
```

由于Blade聚合口缺省以源目地址HASH方式分担流量到不同业务板的两个channel上，当设备配置NAT Outbound转换时，设备会针对NAT地址池中的每一个IP地址作为目的地址Match information下发一条o

penflow流表，这样来保证NAT转换之后的回程流量能正确上到同一块业务板上处理，保证来回流量业务一致性。