瞭解PIM斷言機制

目錄

簡介

本文描述了協定無關組播(PIM)斷言機制，集中於PIM斷言優勝標準，並深入到某些角的情況。

必要條件

需求

思科建議您瞭解PIM斷言機制。

採用元件

本檔案中的資訊是根據Cisco CSR1000V版本16.4.1

本文中的資訊是根據特定實驗室環境內的裝置所建立。文中使用到的所有裝置皆從已清除（預設）的組態來啟動。如果您的網路運作中，請確保您瞭解任何指令可能造成的影響。

什麼是PIM斷言機制？

當共用網段上有多個啟用PIM的路由器時，這些路由器可能會遇到重複的組播流量。這種情況可能是因為，同一共用網段上的兩台或多台路由器可能有一個有效的(S，G)或(*,G)條目，該條目會向同一源IP/目標組的共用網段填充傳出介面。

PIM斷言機制用於檢測和消除共用網段上的組播流量重複。必須注意的是，此機制並不防止重複發生，相反，它使用組播流量的重複作為觸發器，以便啟用此機制，此機製為此流選擇單個轉發器。

當共用網段上存在組播流量重複時，可以假設有多台路由器在一個共用網段上傳送相同的(S，G)或(*,G)。如果您選擇一台路由器來有效轉發此資料流，就可以消除重複。

PIM利用當您在傳出介面清單(OIL)上收到組播資料包時觸發的PIM斷言消息。 這些斷言消息包含用於計算哪些人將成為斷言獲勝者的度量。LAN上的下游路由器也會收到PIM斷言訊息。下游裝置接

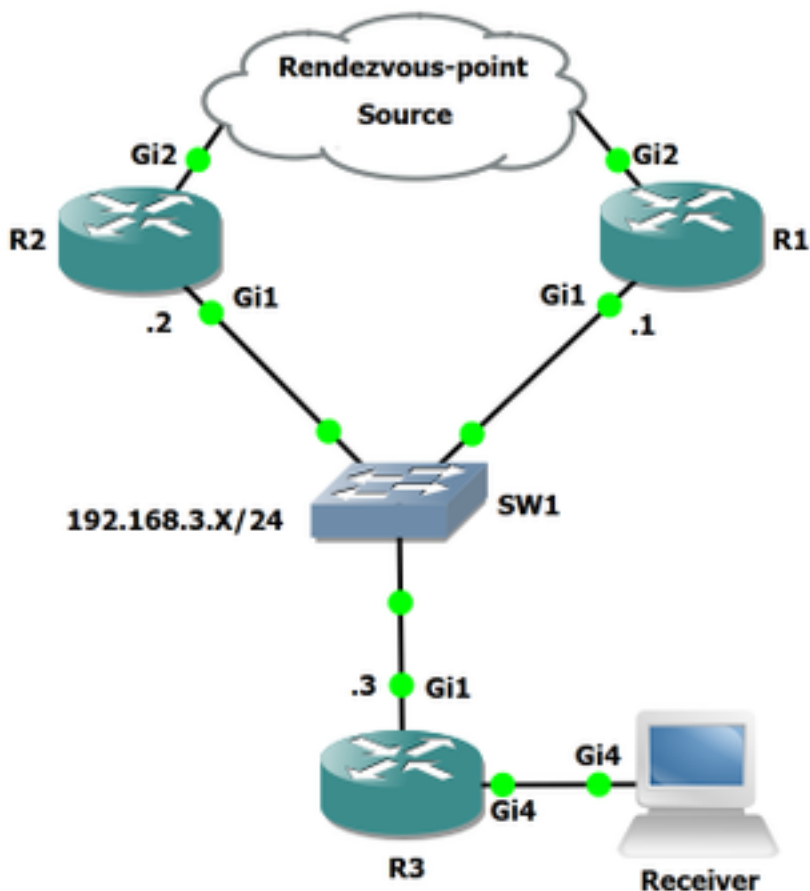著使用這些訊息將適當的加入/修整訊息傳送至贏得斷言選擇的上游路由器。

案例1. LHR原理

圖1.

在網路圖中，R3是最後一跳路由器(LHR),R3通過共用網段連線到R2和R1。

當您從接收方收到網際網路組管理協定(IGMP)報告時，R3會檢查指向RP的RPF鄰居是誰。在拓撲中，R1是指向RP的RPF鄰居，因此R3向R1傳送一個(*,G)連線。一旦R1拉下流（假定組處於活動狀態），R3向源傳送一個(S，G)連線，並將源樹拉下。R2是通向源樹的RPF鄰居，這意味著R3將向R2傳送(S，G)連線。R3對RP和源具有相同的RPF介面。在此處可以看到組239.1.1.1的R3路由表。

```
R3#show ip mroute
IP Multicast Routing Table
Outgoing interface flags: H - Hardware switched, A - Assert winner, p - PIM Join
 Timers: Uptime/Expires
 Interface state: Interface, Next-Hop or VCD, State/Mode
(*, 239.1.1.1), 00:00:55/stopped, RP 192.168.0.100, flags: SJC
  Incoming interface: GigabitEthernet1, RPF nbr 192.168.3.1
  Outgoing interface list:
    GigabitEthernet4, Forward/Sparse, 00:00:55/00:02:04

(10.0.0.2, 239.1.1.1), 00:00:52/00:02:07, flags: JT
  Incoming interface: GigabitEthernet1, RPF nbr 192.168.3.2, Mroute
  Outgoing interface list:
    GigabitEthernet4, Forward/Sparse, 00:00:52/00:02:07
```

```
(*, 224.0.1.40), 00:01:22/00:02:09, RP 192.168.0.100, flags: SJPCL
  Incoming interface: GigabitEthernet1, RPF nbr 192.168.3.1
```

您可以看到，在R3上，(*,G)RPF鄰居是192.168.3.1，而指向(S，G)的RPF鄰居是192.168.3.2。現在，這應該會導致R1和R2都有一個指向R3的有效OIL。讓我們看一下這些條目：

```
R1#show ip mroute
Outgoing interface flags: H - Hardware switched, A - Assert winner, p - PIM Join
 Timers: Uptime/Expires
 Interface state: Interface, Next-Hop or VCD, State/Mode

(*, 239.1.1.1), 00:15:02/00:02:33, RP 192.168.0.100, flags: S
  Incoming interface: GigabitEthernet2, RPF nbr 192.168.5.2
  Outgoing interface list:
    GigabitEthernet1, Forward/Sparse, 00:15:02/00:02:33

(10.0.0.2, 239.1.1.1), 00:13:24/00:02:33, flags: PR
  Incoming interface: GigabitEthernet2, RPF nbr 192.168.5.2
  Outgoing interface list: Null

(*, 224.0.1.40), 00:29:17/00:02:51, RP 192.168.0.100, flags: SJCL
  Incoming interface: GigabitEthernet2, RPF nbr 192.168.5.2
  Outgoing interface list:
    GigabitEthernet1, Forward/Sparse, 00:16:06/00:02:51
  Outgoing interface list: Null
```

```
R2#show ip mroute
IP Multicast Routing Table
Outgoing interface flags: H - Hardware switched, A - Assert winner, p - PIM Join
 Timers: Uptime/Expires
 Interface state: Interface, Next-Hop or VCD, State/Mode

(*, 239.1.1.1), 00:08:00/stopped, RP 192.168.0.100, flags: SP
  Incoming interface: GigabitEthernet2, RPF nbr 192.168.4.1
  Outgoing interface list: Null

(10.0.0.2, 239.1.1.1), 00:00:03/00:02:56, flags: T
  Incoming interface: GigabitEthernet2, RPF nbr 192.168.4.1
  Outgoing interface list:
    GigabitEthernet1, Forward/Sparse, 00:00:03/00:03:26

(*, 224.0.1.40), 01:37:30/00:02:22, RP 192.168.0.100, flags: SJPL
  Incoming interface: GigabitEthernet2, RPF nbr 192.168.4.1
```

如前所述，當共用網段上填充了有效OIL的兩台上游路由器時，可以觸發斷言。由於R1和R2都具有有效的OIL，因此請檢查資料包捕獲中是否存在斷言機制。

此資料包捕獲是在R3介面Gi1上捕獲的，指向SW1。

Apply a display filter ... <Ctrl-/>   Expression... +

No.	Time	Source	Destination	Protocol	Length	Info
1	0.000000	192.168.3.1	224.0.0.5	OSPF	98	Hello Packet
2	0.705389	192.168.3.2	224.0.0.5	OSPF	98	Hello Packet
3	3.124776	192.168.3.3	224.0.0.5	OSPF	98	Hello Packet
4	7.733948	192.168.3.2	224.0.0.13	PIMv2	72	Hello
5	9.480827	192.168.3.1	224.0.0.5	OSPF	98	Hello Packet
6	10.256987	192.168.3.1	224.0.0.5	OSPF	98	Hello Packet
7	11.954130	192.168.3.1	224.0.0.13	PIMv2	72	Hello
8	12.621371	192.168.3.3	224.0.0.13	PIMv2	72	Hello
9	13.015136	192.168.3.3	224.0.0.5	OSPF	98	Hello Packet
10	19.046520	192.168.3.1	224.0.0.5	OSPF	98	Hello Packet
11	19.670571	192.168.3.2	224.0.0.5	OSPF	98	Hello Packet
12	22.114741	10.0.0.2	239.1.1.1	ICMP	114	Echo (ping) request id=0x000d, seq=0/0, ttl=253 (multicast)
13	22.137371	192.168.3.3	224.0.0.13	PIMv2	68	Join/Prune
14	22.137597	192.168.3.1	224.0.0.13	PIMv2	68	Join/Prune
15	22.972394	192.168.3.3	224.0.0.5	OSPF	98	Hello Packet
16	23.085520	10.0.0.2	239.1.1.1	ICMP	114	Echo (ping) request id=0x000d, seq=1/256, ttl=253 (multicast)
17	24.087827	10.0.0.2	239.1.1.1	ICMP	114	Echo (ping) request id=0x000d, seq=2/512, ttl=253 (multicast)
18	24.723777	192.168.3.3	224.0.0.13	PIMv2	96	Join/Prune
19	25.088340	10.0.0.2	239.1.1.1	ICMP	114	Echo (ping) request id=0x000d, seq=3/768, ttl=253 (multicast)
20	26.091246	10.0.0.2	239.1.1.1	ICMP	114	Echo (ping) request id=0x000d, seq=4/1024, ttl=253 (multicast)
21	27.091219	10.0.0.2	239.1.1.1	ICMP	114	Echo (ping) request id=0x000d, seq=5/1280, ttl=253 (multicast)
22	28.109058	10.0.0.2	239.1.1.1	ICMP	114	Echo (ping) request id=0x000d, seq=6/1536, ttl=253 (multicast)
23	29.000065	192.168.3.1	224.0.0.5	OSPF	98	Hello Packet
24	29.118436	10.0.0.2	239.1.1.1	ICMP	114	Echo (ping) request id=0x000d, seq=7/1792, ttl=253 (multicast)
25	29.225379	192.168.3.2	224.0.0.5	OSPF	98	Hello Packet

```
0000  01 00 5e 00 00 05 00 15  e5 9c 3a 00 08 00 45 c0   ..^..... ..:...E.
```

○ ⚡ wireshark_-_20171228095051_a07600   Packets: 25 · Displayed: 25 (100.0%) · Dropped: 0 (0.0%) Profile: Default

在此資料包捕獲中，您看不到任何斷言資料包，即使在R1、R2和R3之間的共用網段上建立複製的所有先決條件也是如此。啟用(S、G)流時，為什麼看不到任何PIM斷言資料包？

RFC 7761似乎可以找到這些問題的答案。

摘自RFC 7761第4.2.2節。

```
4.2.2.  Setting and Clearing the (S,G) SPTbit

   Basically, Update_SPTbit(S,G,iif) will set the SPTbit if we have the
   appropriate (S,G) join state, and if the packet arrived on the
   correct upstream interface for S, and if one or more of the following
   conditions apply:

   1.  The source is directly connected, in which case the switch to the
       SPT is a no-op.

   2.  The RPF interface to S is different from the RPF interface to the
       RP.  The packet arrived on RPF_interface(S), and so the SPT must
       have been completed.

   3.  No one wants the packet on the RP tree.

   4.  RPF'(S,G) == RPF'(*,G).  In this case, the router will never be
       able to tell if the SPT has been completed, so it should just
       switch immediately.  The RPF'(S,G) != NULL check ensures that the
       SPTbit is set only if the RPF neighbor towards S is valid.

   In the case where the RPF interface is the same for the RP and for S,
   but RPF'(S,G) and RPF'(*,G) differ, we wait for an Assert(S,G), which
   indicates that the upstream router with (S,G) state believes the SPT
   has been completed.
```

(S，G)SPTbit用於區分是按(*,G)狀態轉發還是按(S，G)狀態轉發。當您從RP樹切換到源樹時，有一個過渡期，當資料由於上游(*,G)狀態而到達時，上游(S，G)狀態被建立，此時，路由器應只繼續在(*,G)狀態上轉發。這防止在上游(S，G)狀態完成建立之前傳送修剪(S，G，rpt)引起的臨時黑洞。

雖然這個方案似乎可以和上面提到的最後一點有關。如果RP和S的RPF介面相同，但是RPF'(S，G)和RPF'(*,G)不同，我們等待一個Assert(S，G)，這表示處於(S，G)狀態的上游路由器認為SPT已經完成。

要觸發斷言，路由器必須在其已填充的OIL上接收該網段上同一源IP/目標組的重複資料包。R3也是LHR，這意味著當從(*,G)接收資料包時，它被指定從(*,G)切換到SPT(S，G)。

在資料包捕獲中，我們發現未觸發斷言。雖然我們確實會看到收到第一個ICMP回應後立即傳送的修剪。

您可以看到，在R3介面G1上收到第一個網際網路控制訊息通訊協定(ICMP)要求封包後，系統會向上游鄰居192.168.3.1傳送(*,G)SR位元修剪。這會修剪所定義的特定來源(*,G)。

您還可以看到以下標誌已設定：(SR):

The S flag: indicates that the multicast group is a sparse mode group.
The R flag: The R flag is the RP-bit flag and indicates that the information in the (S, G) entry is applicable to the shared tree.
在第二個PIM資料包14中，您可以看到R3嘗試加入(S，G)樹。

可以觀察到，一旦接收到第一資料平面，資料包R3修剪(*,G)並構建(S，G)。 這就是看不到PIM斷言資料包的原因。如果您的LHR的(S，G)和(*,G)具有相同的RPF介面，則此特定場景將生效。 雖然此行為可能與RFC 7761略有不同，但它不會導致任何問題。

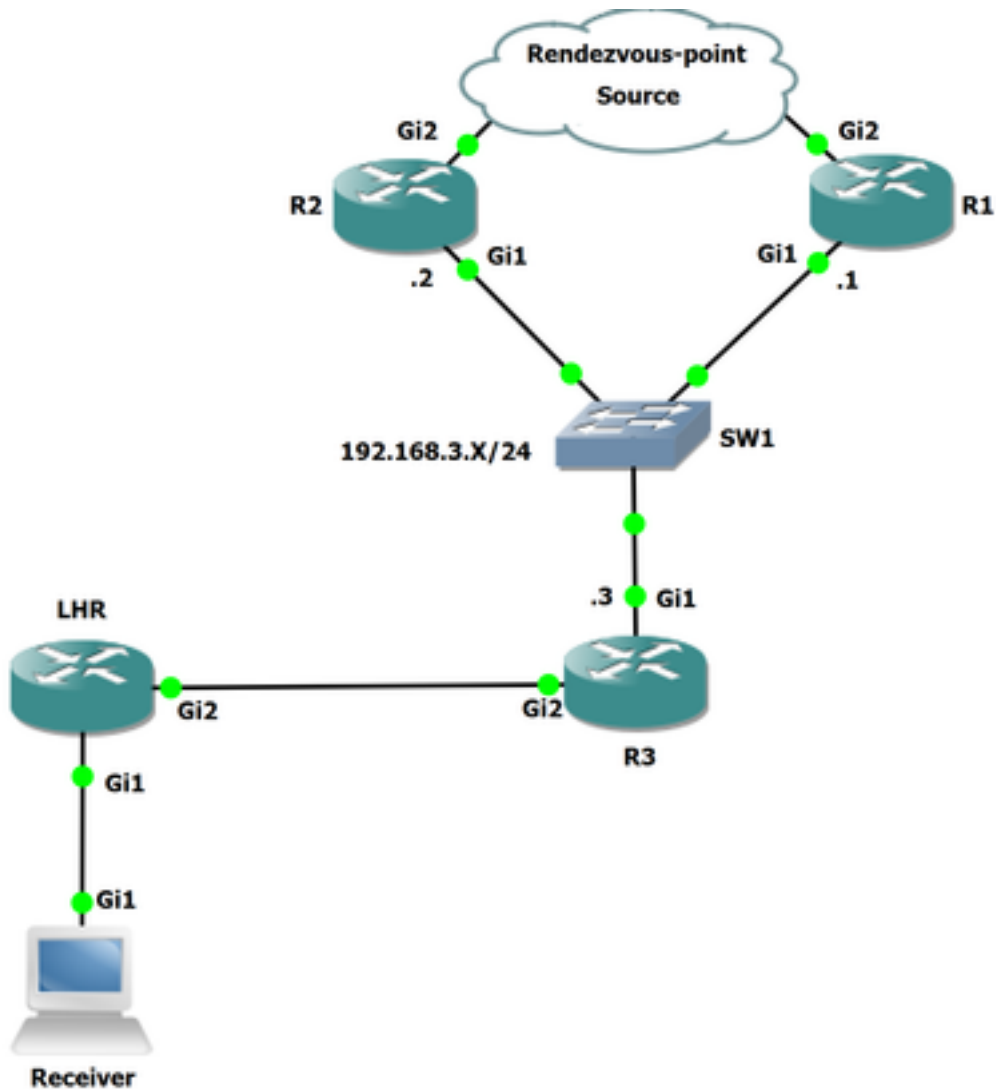現在讓我們繼續講案例2。您可以在此處看到這個案例的圖表：

案例2.斷言路徑選取

圖2.

在此拓撲中，R3上連線的另一台路由器是LHR。LHR直接連線到接收器。源和RP都高於R2和R1。R3上指向RP的RPF鄰居是R1，指向源的RPF鄰居是R2。

讓我們檢查源和RP的RPF鄰居。

在此您可以看到指向RP的RPF鄰居：192.168.0.100是192.168.3.1。

```
R3#show ip rpf 192.168.0.100
RPF information for ? (192.168.0.100)
  RPF interface: GigabitEthernet1
  RPF neighbor: ? (192.168.3.1)
  RPF route/mask: 192.168.0.100/32
  RPF type: unicast (ospf 1)
  Doing distance-preferred lookups across tables
  RPF topology: ipv4 multicast base, originated from ipv4 unicast base
```

這裡您可以看到來源的RPF鄰居：10.0.0.2是192.168.3.2。

```
R3#show ip rpf 10.0.0.2
RPF information for ? (10.0.0.2)
  RPF interface: GigabitEthernet1
  RPF neighbor: ? (192.168.3.2)
```

```
RPF route/mask: 10.0.0.0/24
RPF type: unicast (ospf 1)
Doing distance-preferred lookups across tables
RPF topology: ipv4 multicast base, originated from ipv4 unicast base
```
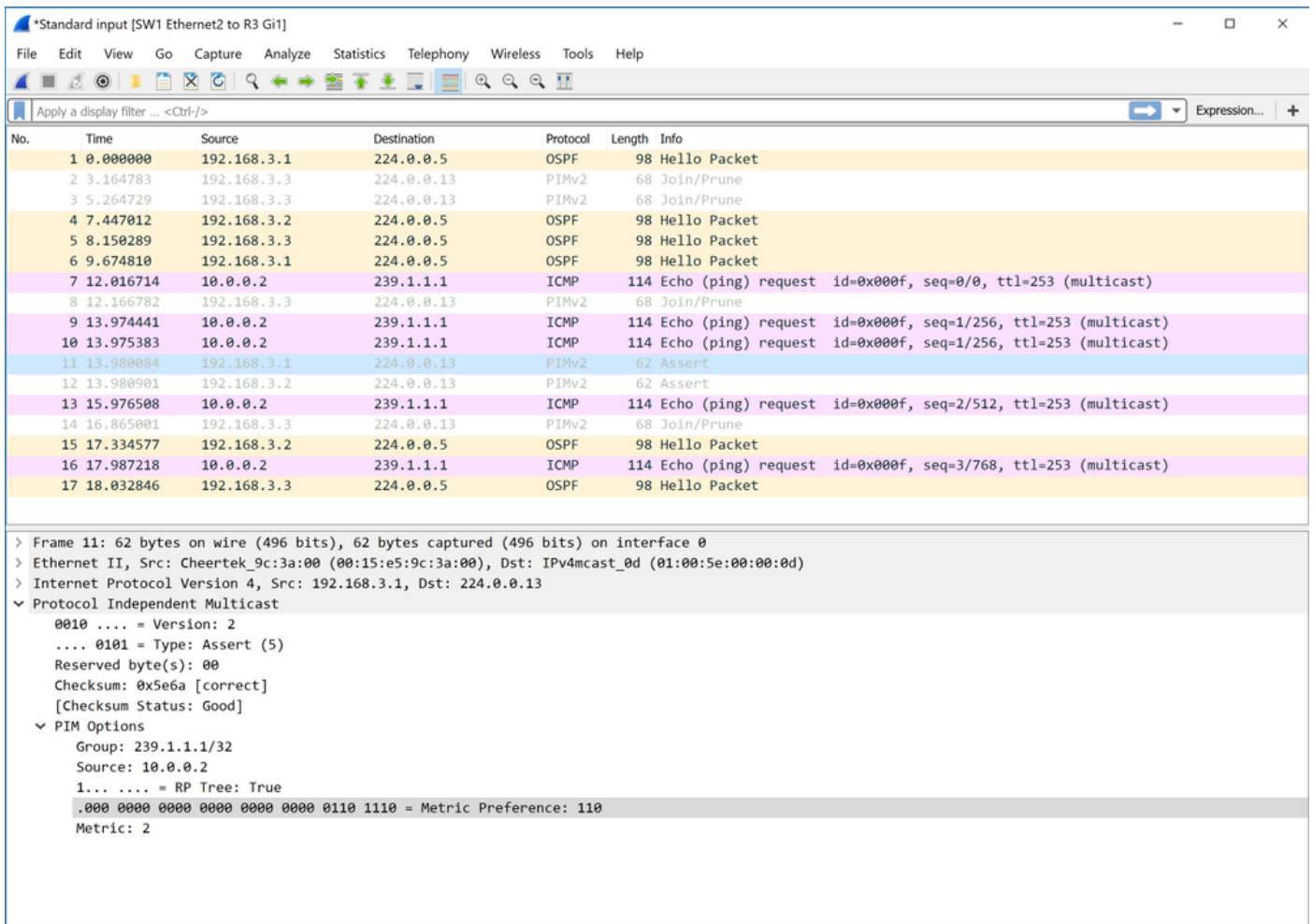
在啟用源之前，讓我們看一下R3上的mroute表，因為您可以看到組239.1.1.1已經存在(*,G)。這是
因為連線到LHR的接收方已經請求了指定組。

```
R3#show ip mroute
IP Multicast Routing Table
Outgoing interface flags: H - Hardware switched, A - Assert winner, p - PIM Join
 Timers: Uptime/Expires
 Interface state: Interface, Next-Hop or VCD, State/Mode

(*, 239.1.1.1), 00:00:57/00:02:32, RP 192.168.0.100, flags: S
  Incoming interface: GigabitEthernet1, RPF nbr 192.168.3.1
  Outgoing interface list:
    GigabitEthernet2, Forward/Sparse, 00:00:57/00:02:32

(*, 224.0.1.40), 00:11:24/00:02:41, RP 192.168.0.100, flags: SJCL
  Incoming interface: GigabitEthernet1, RPF nbr 192.168.3.1
  Outgoing interface list:
    GigabitEthernet2, Forward/Sparse, 00:02:02/00:02:41
```

現在，在R3介面Gi1上啟用源並捕獲資料包。

正如您在此資料包捕獲中看到的，PIM斷言資料包已經存在。

第11幀：

```
> Frame 11: 62 bytes on wire (496 bits), 62 bytes captured (496 bits) on interface 0
> Ethernet II, Src: Cheertek_9c:3a:00 (00:15:e5:9c:3a:00), Dst: IPv4mcast_0d (01:00:5e:00:00:0d)
> Internet Protocol Version 4, Src: 192.168.3.1, Dst: 224.0.0.13
∨ Protocol Independent Multicast
      0010 .... = Version: 2
      .... 0101 = Type: Assert (5)
      Reserved byte(s): 00
      Checksum: 0x5e6a [correct]
      [Checksum Status: Good]
    ∨ PIM Options
         Group: 239.1.1.1/32
         Source: 10.0.0.2
         1... .... = RP Tree: True
         .000 0000 0000 0000 0000 0000 0110 1110 = Metric Preference: 110
         Metric: 2
```

第12幀：

```
> Frame 12: 62 bytes on wire (496 bits), 62 bytes captured (496 bits) on interface 0
> Ethernet II, Src: Cheertek_8b:3e:00 (00:15:e5:8b:3e:00), Dst: IPv4mcast_0d (01:00:5e:00:00:0d)
> Internet Protocol Version 4, Src: 192.168.3.2, Dst: 224.0.0.13
∨ Protocol Independent Multicast
      0010 .... = Version: 2
      .... 0101 = Type: Assert (5)
      Reserved byte(s): 00
      Checksum: 0xde6a [correct]
      [Checksum Status: Good]
    ∨ PIM Options
         Group: 239.1.1.1/32
         Source: 10.0.0.2
         0... .... = RP Tree: False
         .000 0000 0000 0000 0000 0000 0110 1110 = Metric Preference: 110
         Metric: 2
```

當您檢視這些資料包時，您應該能夠確定誰是斷言獲勝者。現在讓我們看一下PIM斷言轉發器選擇。

度量首選項為管理距離(AD)。 這表示路由協定在路由表中安裝路由的管理距離，該距離用於查詢源IP地址，度量是路由開銷。

還有其它屬性用於確定誰是斷言獲勝者。您可以在RFC 7761中看到這些詳細資訊。

摘自RFC 7761第4.6.3節。

```
4.6.3.  Assert Metrics

   Assert metrics are defined as:

     struct assert_metric {
       rpt_bit_flag;
       metric_preference;
       route_metric;
       ip_address;
     };
   When comparing assert_metrics, the rpt_bit_flag, metric_preference,
   and route_metric fields are compared in order, where the first lower
```

```
      value wins.  If all fields are equal, the primary IP address of the
      router that sourced the Assert message is used as a tie-breaker, with
      the highest IP address winning.
```

使用這些已定義的欄位和路徑選擇可以確定在此場景中斷言獲勝者是誰。如果再次檢視斷言資料包，您會發現沒有比較度量首選項，因為決策是根據第一個選擇標準rpt_bit_flag做出的。

在此場景中，比較R1和R2。兩個路由器都傳送之前看到的斷言消息，一旦兩個裝置看到彼此的斷言消息，它們就可以比較彼此之間的度量，以確定贏家。

因為R2會傳送帶有RP樹的斷言消息：False（其值為0），實際上比R1與RP樹一起傳送的值低：值為1的True。RP樹位設定為0或1。

設定為1時RP樹位表示您當前位於共用樹上；清除的RPT位表示斷言的傳送方在介面上處於(S，G)轉發狀態。

正如(S，G)斷言優先於(*,G)斷言，R2應成為斷言獲勝者。過渡到「I am Assert Winner」狀態。如前面的RFC 7761語句中所述，更首選較低值。

讓我們看一下R1和R2，看看誰是公認的贏家。

```
R2#show ip mroute
IP Multicast Routing Table
Outgoing interface flags: H - Hardware switched, A - Assert winner, p - PIM Join
 Timers: Uptime/Expires
 Interface state: Interface, Next-Hop or VCD, State/Mode

(*, 239.1.1.1), 00:42:52/stopped, RP 192.168.0.100, flags: SP
  Incoming interface: GigabitEthernet2, RPF nbr 192.168.4.1
  Outgoing interface list: Null

(10.0.0.2, 239.1.1.1), 00:42:52/00:01:40, flags: T
  Incoming interface: GigabitEthernet2, RPF nbr 192.168.4.1
  Outgoing interface list:
    GigabitEthernet1, Forward/Sparse, 00:42:52/00:03:07, A

(*, 224.0.1.40), 00:43:23/00:02:25, RP 192.168.0.100, flags: SJPL
  Incoming interface: GigabitEthernet2, RPF nbr 192.168.4.1
  Outgoing interface list: Null
```

在此輸出中，您可以看到R2上的(S，G)在OIL上設定了A標誌，表示它是斷言獲勝者。在R1上，您在(S，G)上沒有OIL，並且設定了P標誌，這表示在本例中已將特定(S，G)剪除：它不是絕對的贏家。

> **附註**：當斷言存在於共用網段上時，下游鄰居會定期向適當的RPF鄰居（即斷言進程修改的RPF鄰居）傳送Join(*,G)和Join(S，G)消息。它們並不總是按照MRIB指示傳送到RPF鄰居。

```
R1#show ip mroute
IP Multicast Routing Table
Outgoing interface flags: H - Hardware switched, A - Assert winner, p - PIM Join
 Timers: Uptime/Expires
 Interface state: Interface, Next-Hop or VCD, State/Mode

(*, 239.1.1.1), 00:44:32/00:03:09, RP 192.168.0.100, flags: S
  Incoming interface: GigabitEthernet2, RPF nbr 192.168.5.2
```

```
   Outgoing interface list:
      GigabitEthernet1, Forward/Sparse, 00:44:32/00:03:09, A

(10.0.0.2, 239.1.1.1), 00:44:19/00:03:09, flags: PR
   Incoming interface: GigabitEthernet2, RPF nbr 192.168.5.2
   Outgoing interface list: Null

(*, 224.0.1.40), 00:44:50/00:02:53, RP 192.168.0.100, flags: SJCL
   Incoming interface: GigabitEthernet2, RPF nbr 192.168.5.2
   Outgoing interface list:
      GigabitEthernet1, Forward/Sparse, 00:43:56/00:02:53
```

如果R1和R2的RP樹位都設定為1。則可以考慮具有最低AD的路由器;如果相等,則檢視度量。如果兩台路由器上的RP樹位均為true,則將度量與RP IP地址進行比較。如果RP樹位為0,則將該度量與組播流的源進行比較。

如果所有這些值都相同,則最高IP地址來源斷言消息為獲勝者。

摘要

在案例1中,您並未觀察到斷言封包,但根據RFC,它們本應已觸發。如前所述,這是因為R3在構建(S,G)的控制平面之前正在修剪(*,G)。

在案例2中,您可以看到斷言封包。當LHR收到第一個資料包時,它會向R3傳送(S,G)加入/修剪以拉取源/組。然後R3將向R2傳送相同源/組的加入/修整資料包。這將導致R1和R2填充有效的OIL。現在,當T標誌填充到R3(S,G)狀態時,R3隻修剪(S,G)設定RP位。為此,您需要從共用網段接收另一個資料平面資料包。由於已經為(S,G)建立了控制平面,因此這會導致共用段上的重複觸發斷言消息。