



Multicast Case Sharing

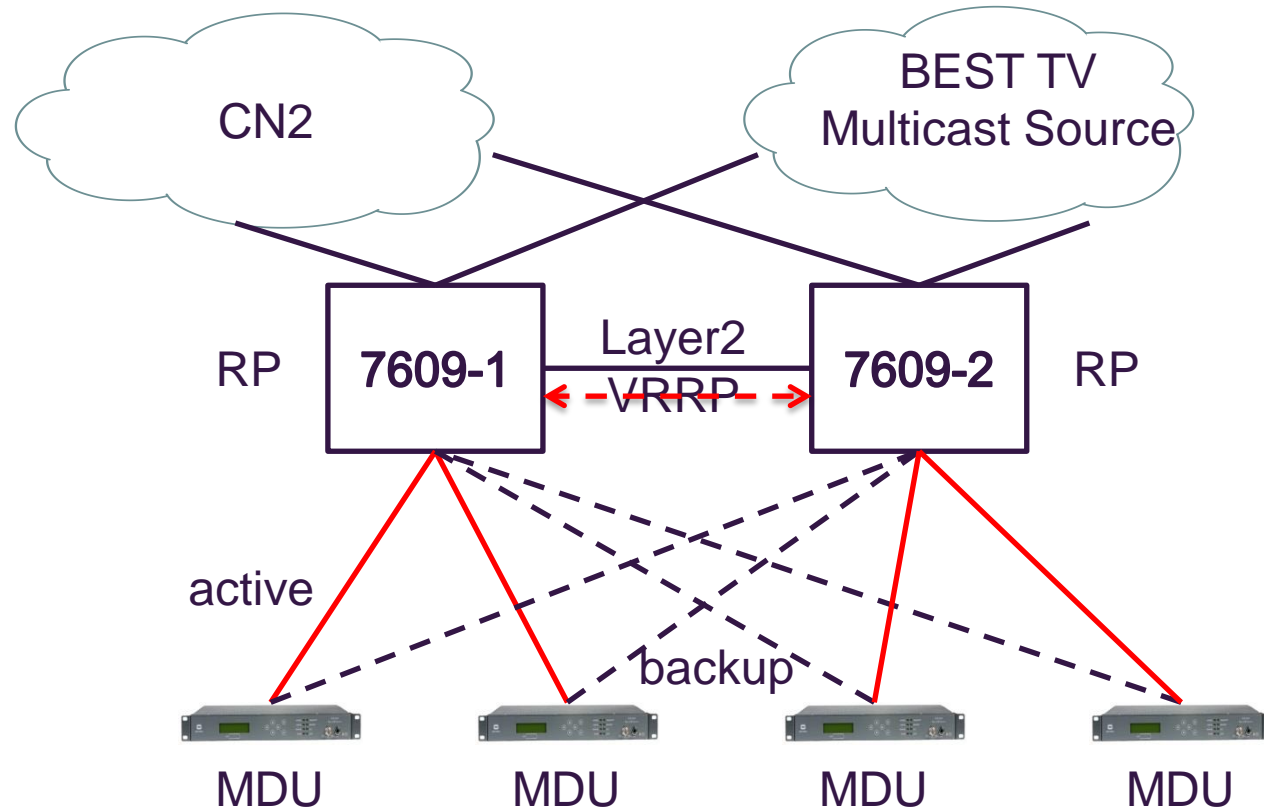
Chen Fei

2014/7

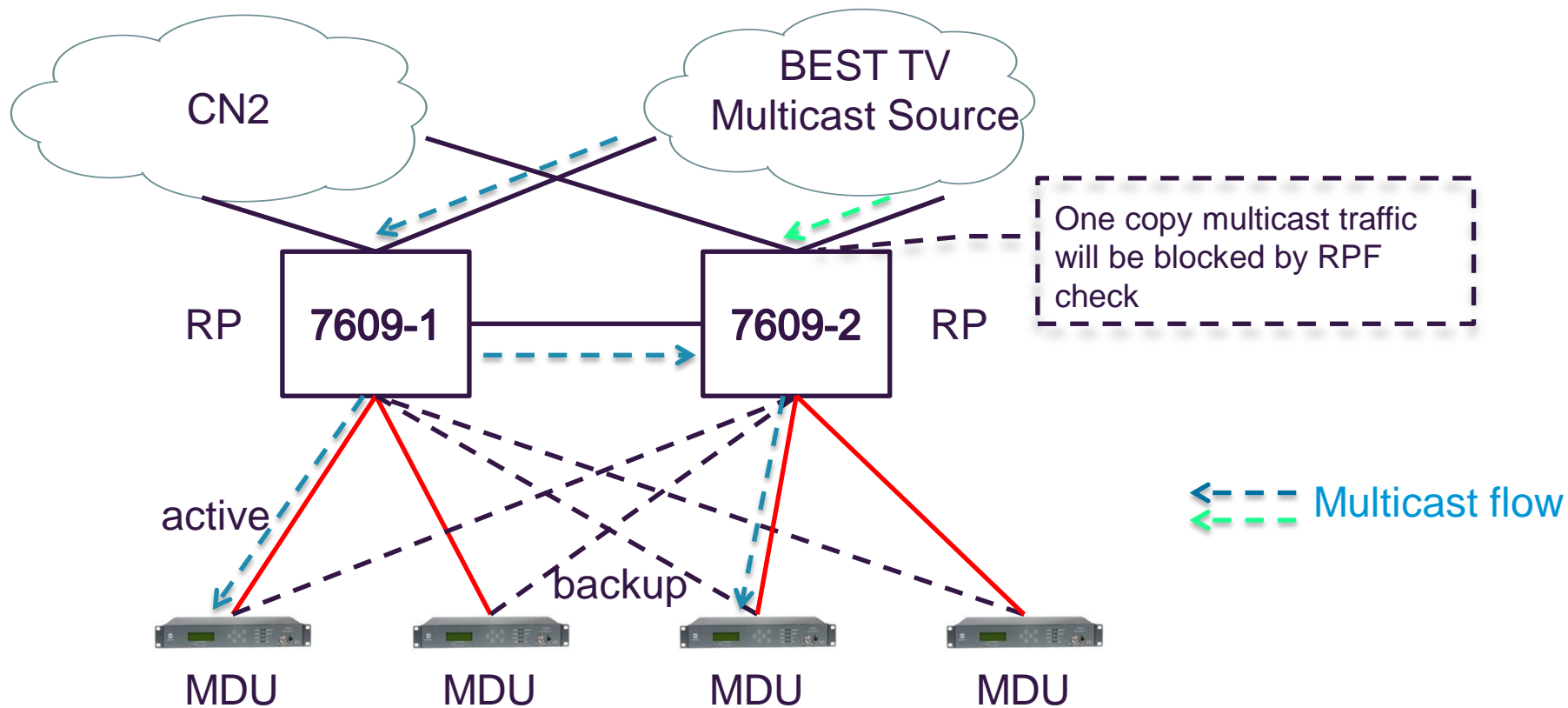
Case 1: Traffic Problem



Customer Topology



Traffic flow



Problem description

- MDU report that some multicast channel did not retrieve at MDU.
- Check the multicast entry on 7609-1, (*,G) entry exist, the interface VLAN 12 which MDU connected are listed on outgoing interface list. But (S,G) entry did not create.

```
7609-1#sh ip mroute 233.22.1.12
```

```
(*, 233.22.1.12), 00:45:58/00:02:52, RP 221.231.144.92, flags: SJC
```

```
Incoming interface: TenGigabitEthernet9/2, RPF nbr 58.223.143.249, Partial-SC
```

```
Outgoing interface list:
```

```
Vlan12, Forward/Sparse, 00:45:58/00:02:52, H
```

Analysis

- (*,G) entry is created by particular IGMP join request (IGMP report packet from client or static configure)
- From the (*,G) entry we can know 7609 can normally receive the IGMP report from MDU.
- If the multicast traffic arrived at 7609's CPU, based on STP switchover rule, the (S,G) entry will be created immediately.
- What we suspected that the device CPU did not receive 233.22.1.12 multicast.
suspected reason
 - a. Traffic did not receive on uplink interface
 - b. RPF check failure
 - c. Traffic reached at 7609 but did not punt to CPU

Analysis

- a. Traffic did not receive on uplink interface

Configured ACL on uplink interface

```
ip access-list 10 per 233.22.1.12 255.255.255.255
```

```
ip access-list 10 per any
```

can observed counter increasing on 233.22.1.12 ACE.

so it proves that the traffic is coming in uplink interface.

Analysis

- b. RPF check failure

Check the RPF interface on this multicast entry

Incoming interface is :

TenGigabitEthernet9/2, RPF nbr 58.223.143.249

Compared with unicast routing table , the RPF interface is correct.

Analysis

- c. Traffic did not punt to CPU

Check the multicast hardware entry on SP, If the MET programmed error

```
show mls cef ip multicast group 233.22.1.12 detail
```

```
(*, 233.22.1.12)
```

```
IOSVPN:0 (1) PI:1 (1) CR:0 (1) Recirc:0 (1)
```

```
Vlan:1023 AdjPtr:360481 FibRpfNf:1 FibRpfDf:1 FibAddr:0x3FB60
```

```
rwlans:1023 rwindex:0x7FFA adjmac:0000.0000.0000 rdt:1 E:0 CAP1:0
```

```
fmt:Mcast l3rwwld:1 DM:0 mtu:1518 rwtype:L2 met2:0x1306 met3:0x0
```

```
packets:00000000000000 bytes:000000000000000000
```

```
Starting Offset: 0x1306
```

```
V C:1016 I:0x07F08 ←--- egress vlan
P->0x01308
V E L0 C:1017 I:0x02063 ←---partial VLAN
```

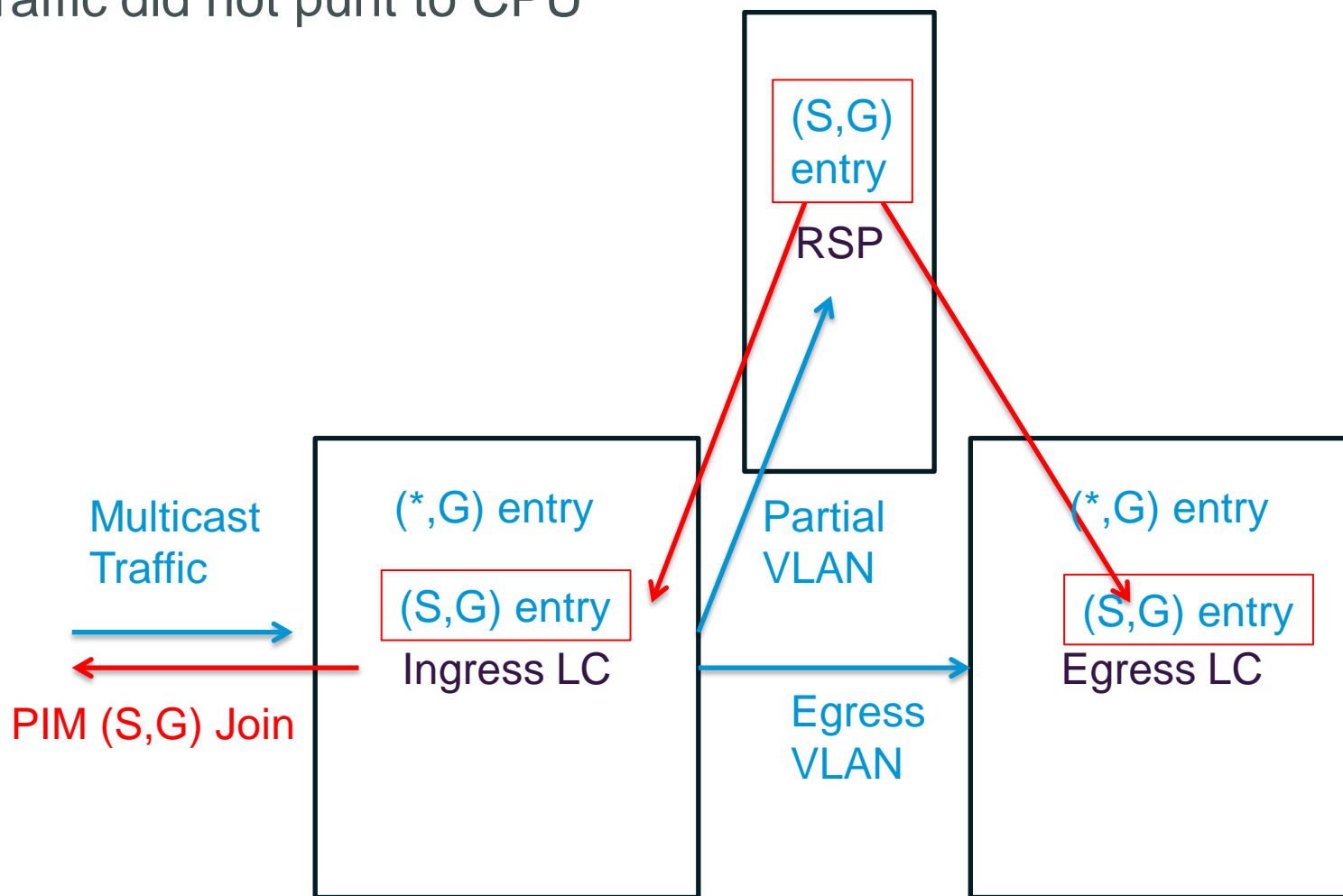
Content of MET2 index

Including: Egress VLAN and Partial VLAN

Egress VLAN - In Egress Replication mode, the replication takes place on the egress line-card .
And so the MET2 entry serves as a pointer to direct the packet to the egress line-card.
Partial VLAN – Partial VLAN is a special internal VLAN used to replicate packets to the RP in addition to local OIFs.

Analysis

- c. Traffic did not punt to CPU



Analysis

- c. Traffic did not punt to CPU

Check VLAN correction

Show vlan internal usage

1016 IPv4 VPN 0 Egress multicast ←----- used on egress replication mode

1017 IP Multicast Partial SC vpn(0)

Analysis

- c. Traffic did not punt to CPU

MET2 entry is correct, next action plan

Capture packet on RP via Netdr

interface Te9/2, routine process_rx_packet_inline, timestamp 04:20:23.516 dbus info:

src_vlan 0x3FF(1023), src_indx 0x201(513), len 0x552(1362) bpdv 0, index_dir 0, flood 1,

dont_lrn 0, dest_indx 0x43FF(17407)

A8020400 03FF0400 02010005 52000000 00110560 09000040 00000008 43FF0000 destmac

01.00.5E.16.01.0C, srcmac 00.21.A0.0B.06.00, protocol 0800 protocol ip: version 0x04, hlen

0x05, tos 0x00, totlen 1344, identifier 65361 df 1, mf 0, fo 0, ttl 1, src 172.27.111.207,

dst 233.22.1.12 udp src 56070, dst 5140 len1324 checksum 0x8A25

ttl=1, so multicast packets will drop on this 7609

Case 2: IGMP Attack



Fake IGMP Packets

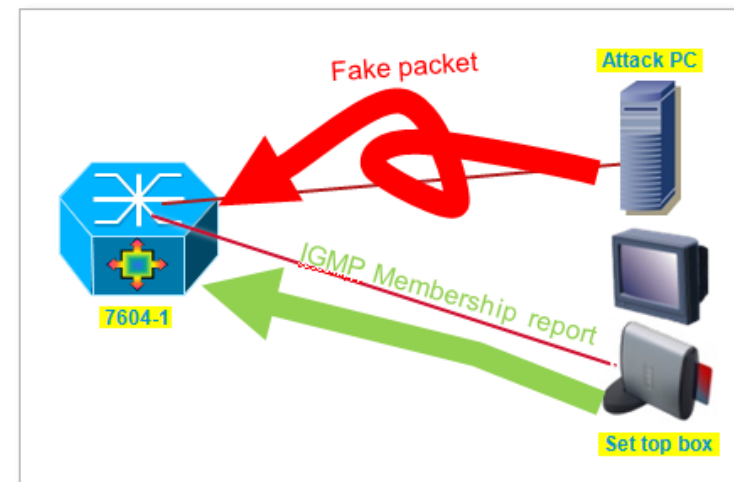
```
###[ IP ]###  
version= 4  
ttl= 1  
proto= igmp  
src= 192.168.213.2  
dst= 239.1.2.21  
###[ IGMP ]###  
type= Version 2 - Membership Report  
gaddr= 239.1.2.22
```

These two fields should be the same!!!

```
interface GigabitEthernet3/42  
description VHE-stream3 video port  
...  
ip igmp access-group allowed-group
```

```
7604-1#sh access-lists allowed-group  
Extended IP access list allowed-group  
10 permit ip any host 239.1.2.21  
20 deny ip any any
```

Now let's see which field is used for the IGMP ACLs



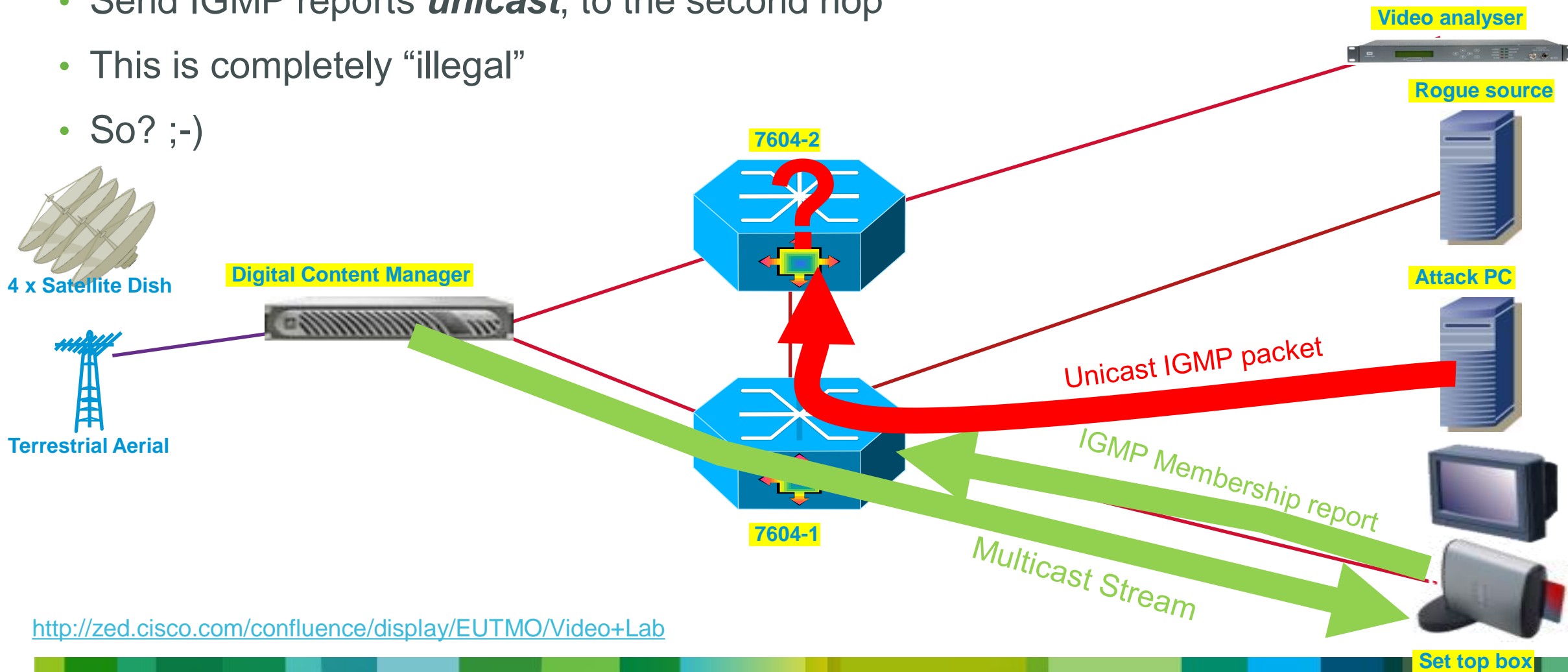
Fake IGMP Packets

IP header – destination	IGMP payload - group	Result on router	Comment
239.1.2.21	239.1.2. 21	state for .21	expected, permitted by ACL
239.1.2.22	239.1.2.22	no new state	expected, blocked by ACL
239.1.2.21	239.1.2.22	no new state	this means that the ACL looks at the group address in the IGMP payload, not the header
239.1.2.22	239.1.2. 21	state for .21	as above
192.168.213.1 (router's i/f)	239.1.2. 21	state for .21	IGMP packets can be sent to a unicast address

- IGMP Process looks at the group in the IGMP header.

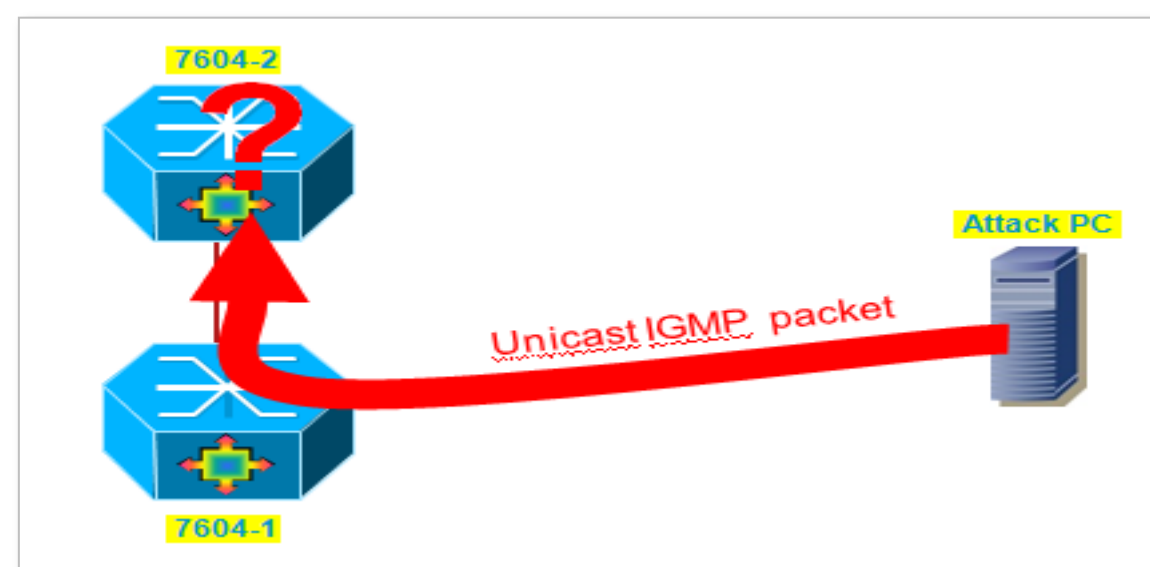
IGMP Unicast Attacks

- Send IGMP reports *unicast*, to the second hop
- This is completely “illegal”
- So? ;-)



<http://zed.cisco.com/confluence/display/EUTMO/Video+Lab>

IGMP Unicast Attacks



```
###[ IP ]###  
version= 4  
ttl= 12  
proto= igmp  
src= 192.168.213.2  
dst= 192.168.109.250  
options= "  
###[ IGMP ]###  
type= Version 2 - Membership Report  
gaddr= 239.1.2.21
```

Need higher TTL than 1 here

Illegal!! IGMP is *never* sent to unicast addresses!

Better not set the router alert option – otherwise every router on the path will inspect the packet

Here is a group address

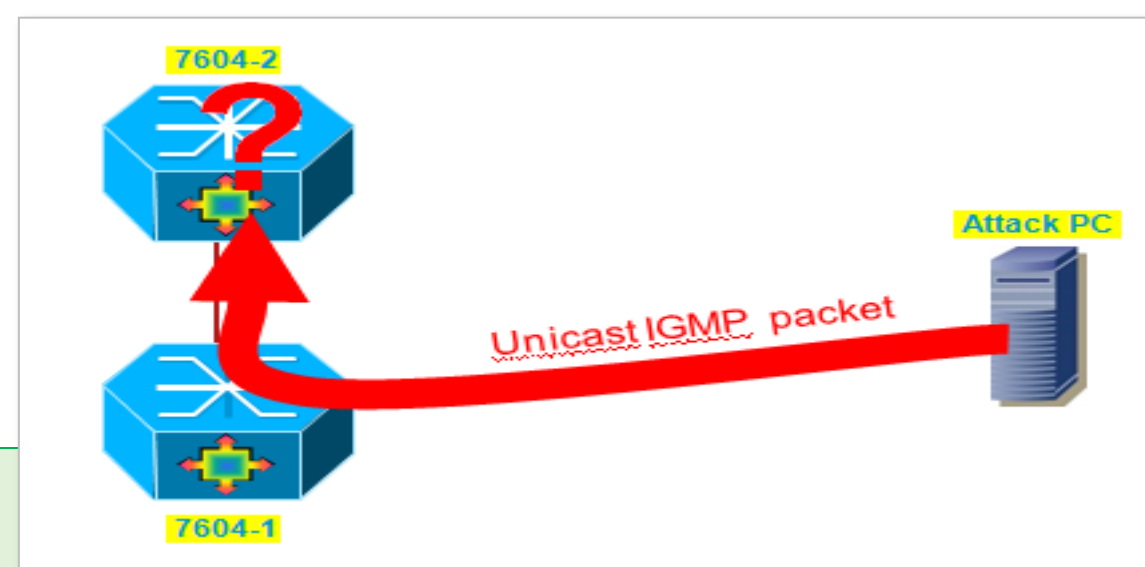
IGMP Unicast Attacks

This is the group we wanted to join (G) on 7604-2!!

```
7604-2>sh ip mroute 239.1.2.21
IP Multicast Routing Table
[...]
(*, 239.1.2.21), 5w0d/00:02:41, RP 192.168.255.2, flags: SJC
  Incoming interface: Null, RPF nbr 0.0.0.0
  Outgoing interface list:
    GigabitEthernet3/34, Forward/Sparse, 00:00:04/00:02:55
    GigabitEthernet2/21, Forward/Sparse, 5w0d/00:02:41

(192.168.183.2, 239.1.2.21), 2d18h/00:01:38, flags: PTX
  Incoming interface: GigabitEthernet3/34, RPF nbr 192.168.109.249
  Outgoing interface list: Null
```

- Can create IGMP state on “remote” routers!!!
- Need to protect with iACL (blocking all traffic to “internal” routers); and/or mls rate limiters;



We created outgoing state on the interface the IGMP was received on!!!

NOT EXPECTED

Case 3: Multicast Forwarder selection on LAN

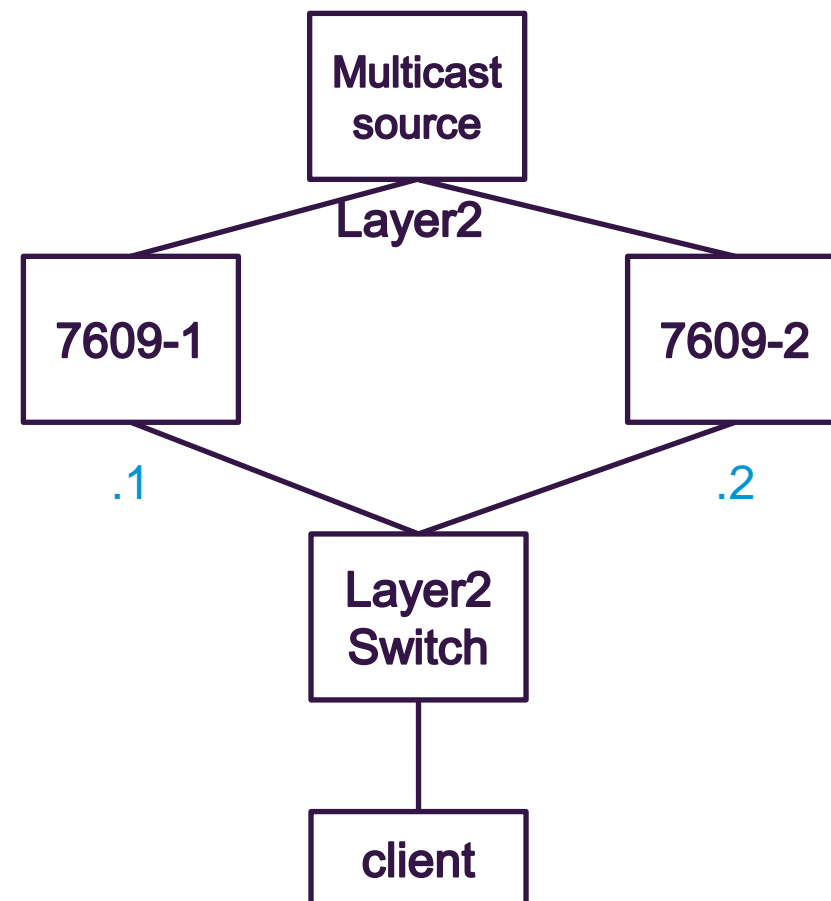


Testing Environment

1. All the device emulated by dynamips
2. Generated multicast traffic by ping one Multicast destination.
3. Individually Configure IGMP static join on each 7609 downstream interface which connected on downstream layer2 switch.
4. Between 7609-1 and 7609-2 running PIM protocol, PIM DR priority on 7609-1 is 100, 7609-2 is 50.

Question:

1. Who will be selected as multicast forwarder on LAN network.
2. If the layer2 switch reload, DR selection and Assert mechanism, which one will take effect



Testing Environment

1. When Lab setup

7609-1 downstream interface with high DR priority, will be selected with DR, set as forwarder interface on this LAN

2. Show ip mroute 224.1.1.1 on 7609-1

(*, 224.1.1.1), 00:05:06/stopped, RP 100.1.1.1, flags: SJC

Incoming interface: Null, RPF nbr 0.0.0.0

Outgoing interface list:

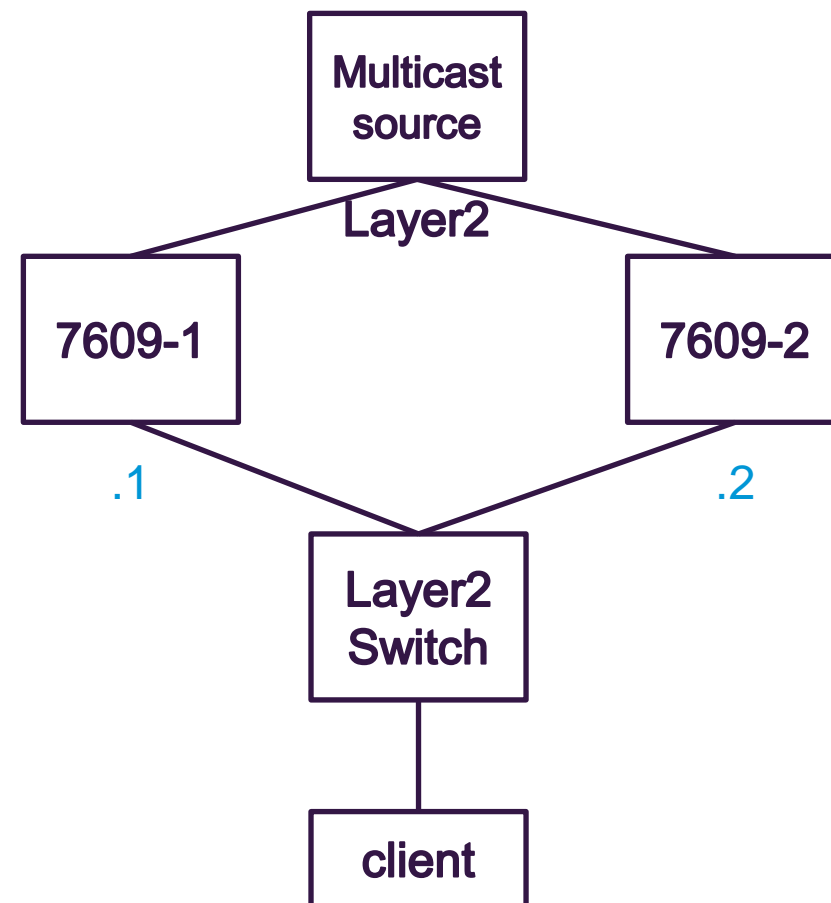
FastEthernet0/1, Forward/Sparse, 00:05:06/00:02:26

(10.1.1.2, 224.1.1.1), 00:05:05/00:02:01, flags: JT

Incoming interface: FastEthernet0/0, RPF nbr 0.0.0.0

Outgoing interface list:

FastEthernet0/1, Forward/Sparse, 00:05:05/00:02:26



Testing Environment

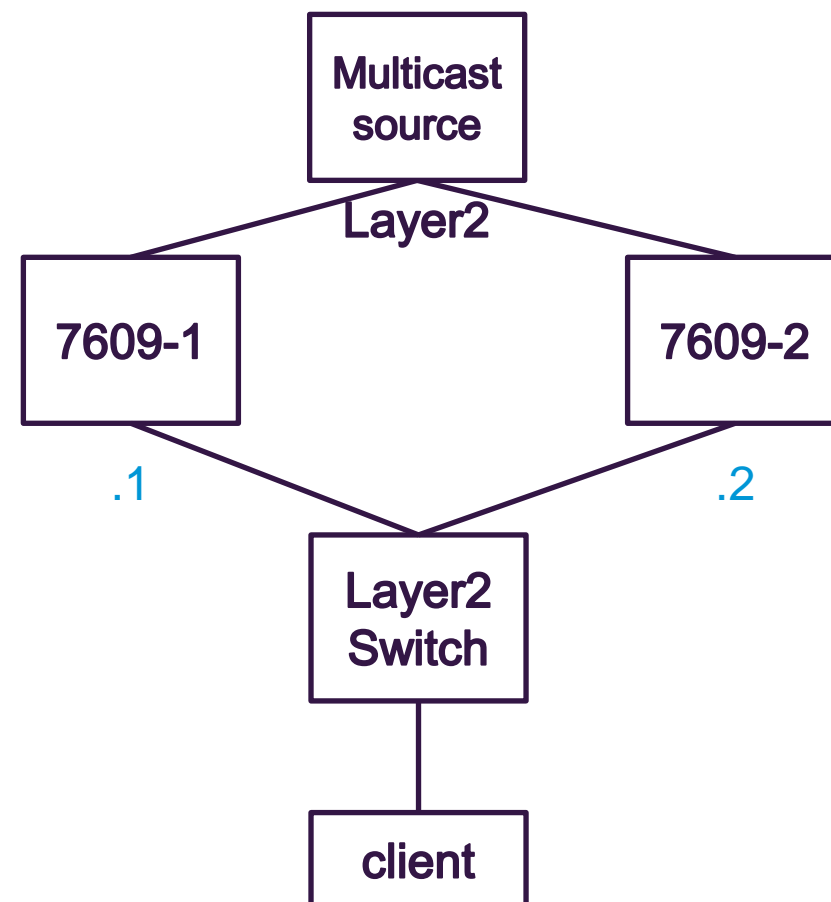
3. Show ip mroute 224.1.1.1 on 7609-2

(*, 224.1.1.1), 00:05:55/stopped, RP 100.1.1.1, flags: SP
Incoming interface: Null, RPF nbr 0.0.0.0
Outgoing interface list: Null

(10.1.1.2, 224.1.1.1), 00:05:55/00:01:00, flags: PT
Incoming interface: FastEthernet0/0, RPF nbr 0.0.0.0
Outgoing interface list: Null

Due to downstream interface is not DR interface, the IGMP info can not be used by this interface. So downstream interface is ineligible to add into Outgoing interface list.

Comply with our network design.



Testing Environment

4. After Layer2 switch reload.

Assumption: The downstream interface on 7609-1/7609-2 still UP.

During Layer 2 switch bootup, 7609-1 / 7609-2 both think itself is DR.
Due to layer 2 link broken, PIM hello packet can not exchange

So the traffic will be send out on both downstream interface.

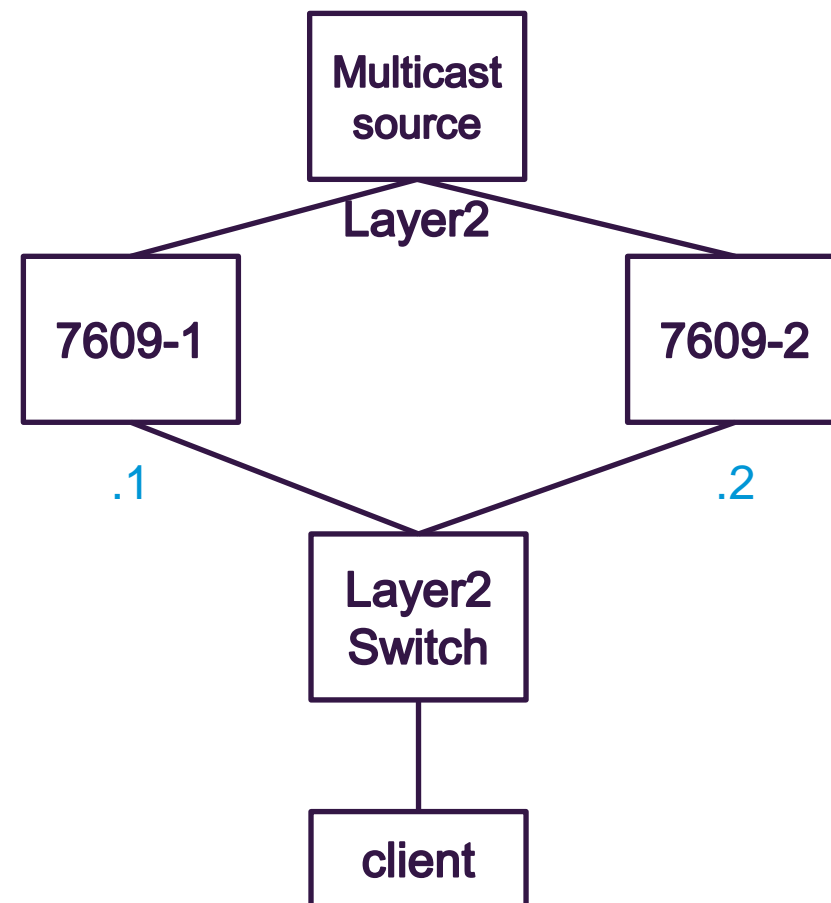
Syslog on 7609-2

*May 30 00:17:44.451: PIM(0): Neighbor 172.16.1.1 (FastEthernet0/1) timed out

*May 30 00:17:44.451: %PIM-5-NBRCHG: neighbor 172.16.1.1 DOWN on interface FastEthernet0/1 DR

*May 30 00:17:44.455: PIM(0): Changing DR for FastEthernet0/1, from 172.16.1.1 to 172.16.1.2 (this system)

*May 30 00:17:44.455: %PIM-5-DRCHG: DR change from neighbor 172.16.1.1 to 172.16.1.2 on interface FastEthernet0/1



Testing Environment

4. After Layer2 switch reload.

Check the mroute table on 7609-2

(*, 224.1.1.1), 00:15:42/stopped, RP 100.1.1.1, flags: SJC

Incoming interface: Null, RPF nbr 0.0.0.0

Outgoing interface list:

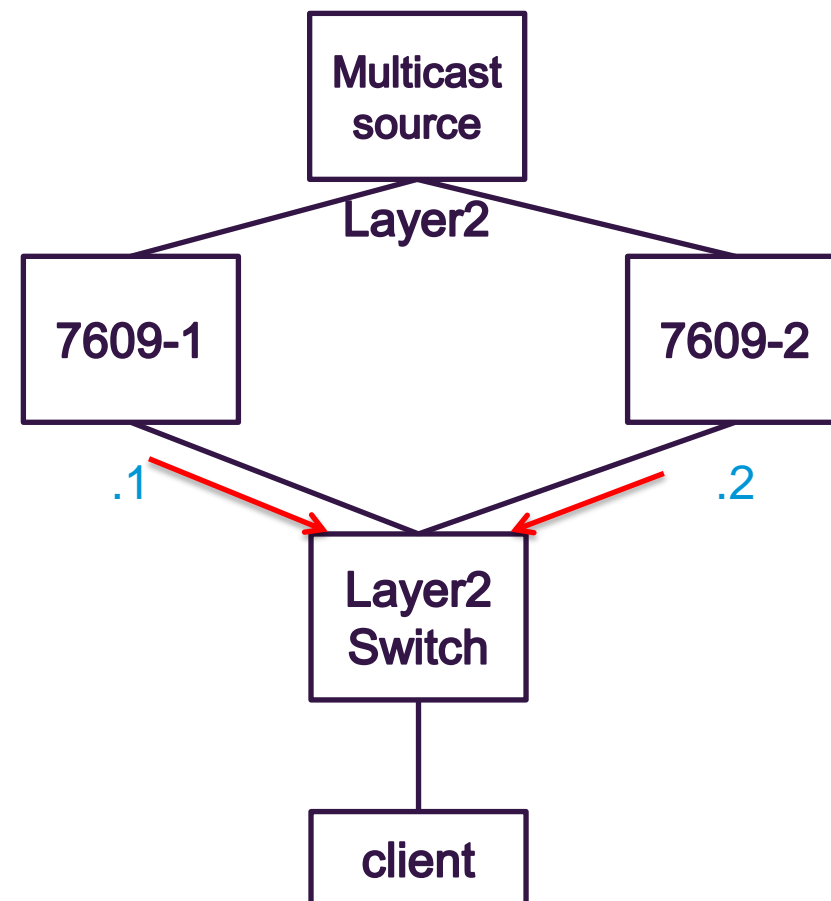
FastEthernet0/1, Forward/Sparse, 00:00:10/00:02:49

(10.1.1.2, 224.1.1.1), 00:15:41/00:01:14, flags: T

Incoming interface: FastEthernet0/0, RPF nbr 0.0.0.0

Outgoing interface list:

FastEthernet0/1, Forward/Sparse, 00:00:10/00:02:49



Testing Environment

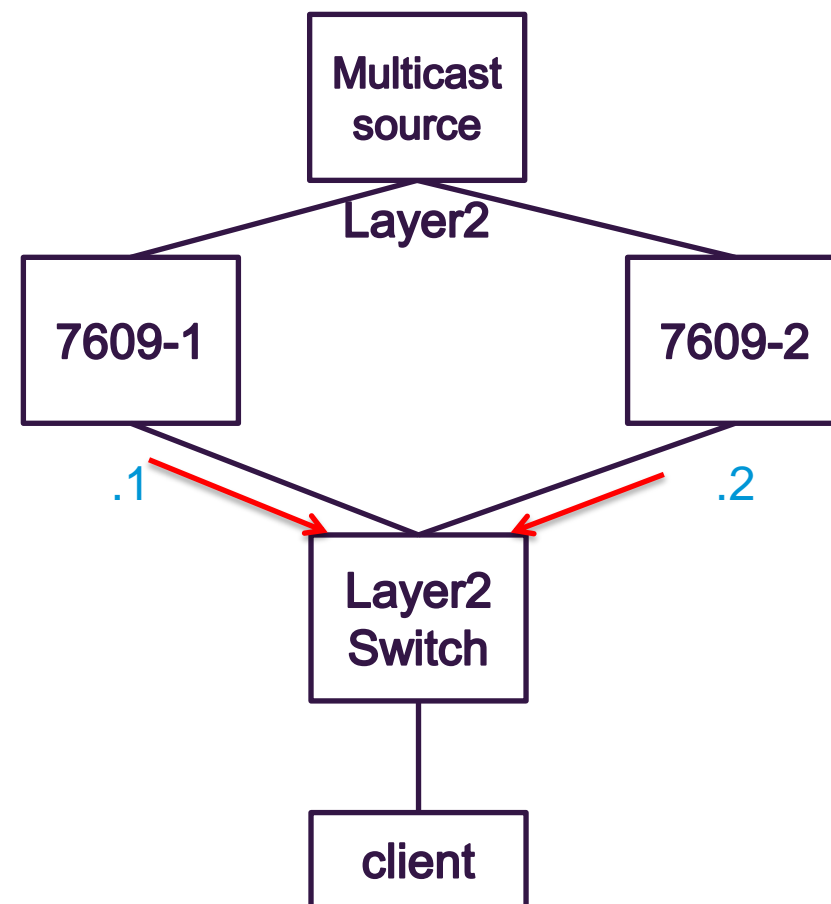
4. When the layer2 switch finished booting

There are two situation.

1) Opposite 7609 PIM hello packets before Multicast traffic be received by CPU

it will make DR selection process working. 7609-2 will remove the NON-DR interface from multicast forwarding entry.

Still 7609-1 as DR, and only forwarder on the LAN



Testing Environment

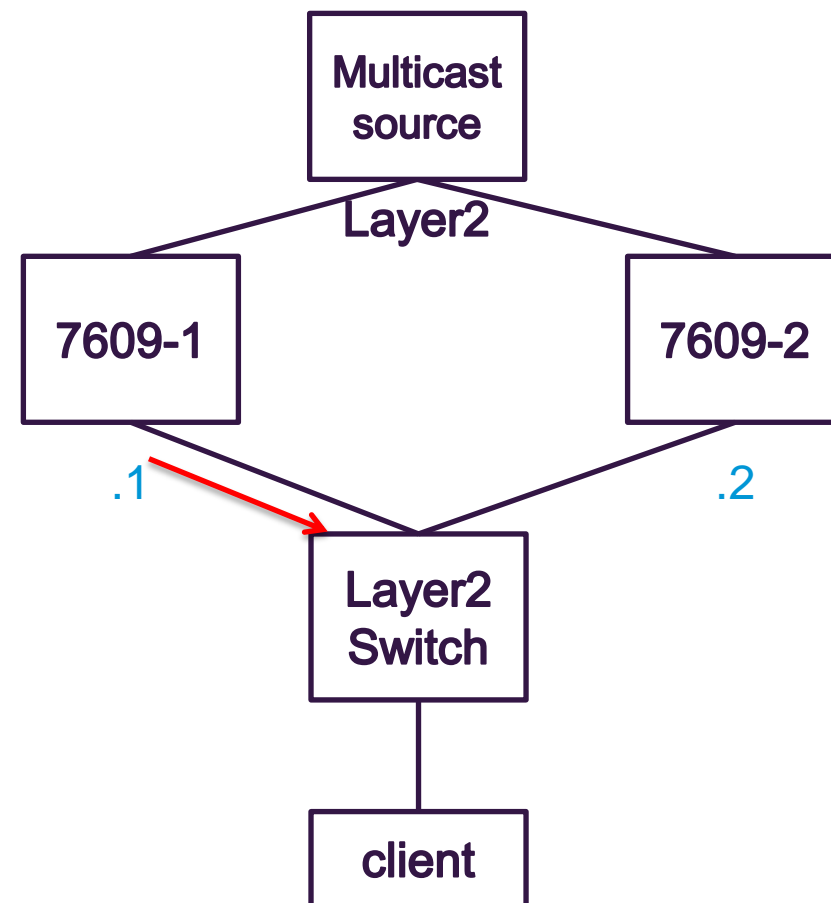
4. When the layer2 switch finished booting

There are two situation.

2) Multicast traffic before Opposite 7609 PIM hello packets be received by CPU

It will cause ASSERT mechanism working, the receive multicast traffic 7609 will send ASSERT packet out the downstream interface.

The ASSERT mechanism is totally difference to DR selection.



Testing Environment

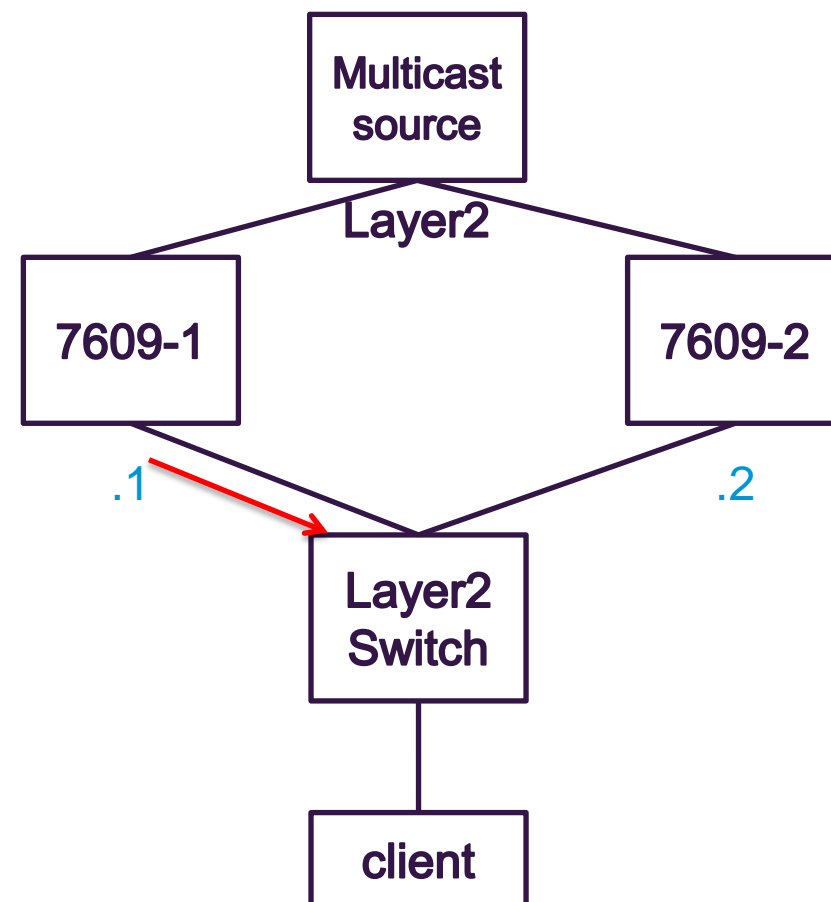
The ASSERT mechanism has three rules

The router generating an Assert with the **lowest Administrative distance** is elected the forwarder.

The best unicast routing metric is used to break a tie if the Administrative Distances are the same. The combination of AD and the **unicast routing metric** is referred to as a “tuple”. If metrics are the same then we move on to step 3.

The device with the **highest IP Address** will be elected as the PIM Forwarder.

The ASSERT result override DR selection



Testing Environment

On our testing environment

From both 7609 to source is layer2 link

So the AD and metric are both 0. directly connection link, so the interface with HIGH IP address, will be the forwarder.

It makes our network design ruin.

Check the syslog on 7609-1

*May 29 23:57:19.491: PIM(0): Received v2 Assert on FastEthernet0/1 from 172.16.1.2

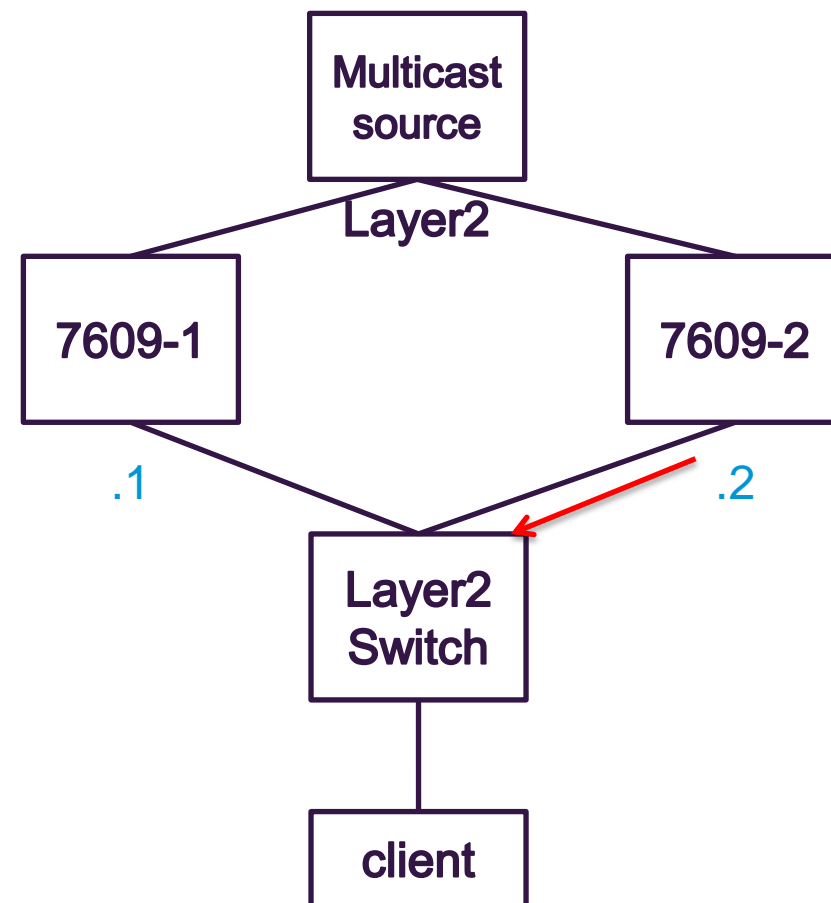
*May 29 23:57:19.491: PIM(0): Assert metric to source 10.1.1.2 is [0/0]

*May 29 23:57:19.495: PIM(0): **We lose, our metric [0/0]**

Check the syslog on 7609-2

*May 29 23:57:19.463: PIM(0): Send v2 Assert on FastEthernet0/1 for 224.1.1.1, source 10.1.1.2, metric [0/0]

*May 29 23:57:19.467: PIM(0): Assert metric to source 10.1.1.2 is [0/0]



Testing Environment

Check the show mroute on 7609-2

(*, 224.1.1.1), 00:20:44/stopped, RP 100.1.1.1, flags: SJC

Incoming interface: Null, RPF nbr 0.0.0.0

Outgoing interface list:

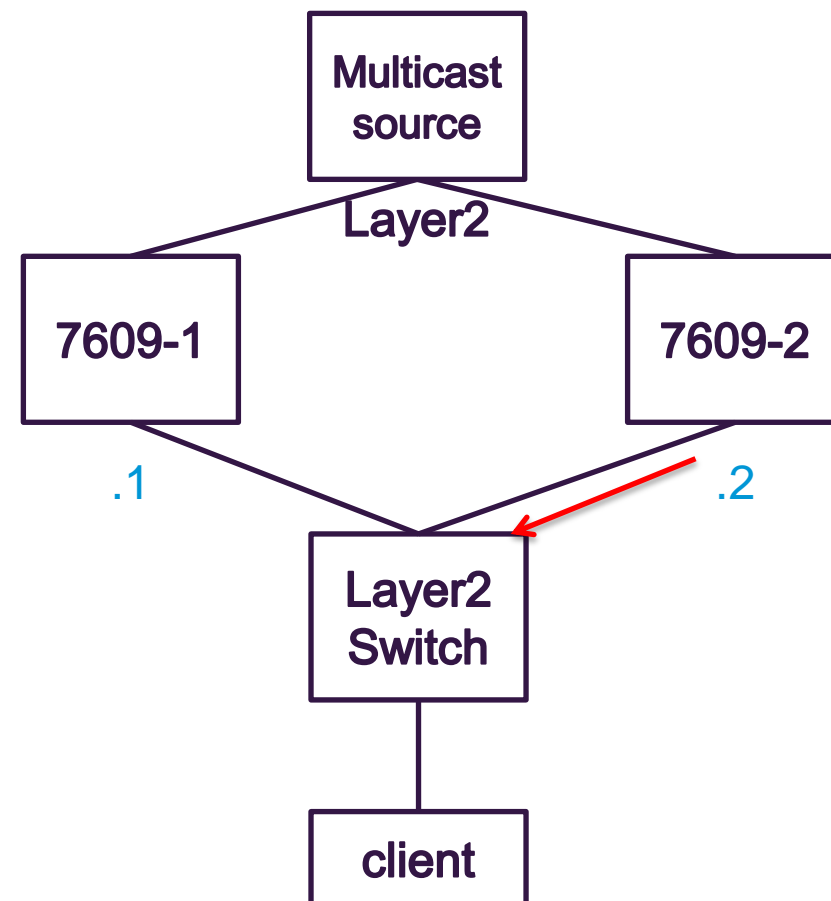
FastEthernet0/1, Forward/Sparse, 00:02:11/00:00:48

(10.1.1.2, 224.1.1.1), 00:15:13/00:02:24, flags: T

Incoming interface: FastEthernet0/0, RPF nbr 0.0.0.0

Outgoing interface list:

FastEthernet0/1, Forward/Sparse, 00:02:11/00:00:48, **A ←-- assert winner**



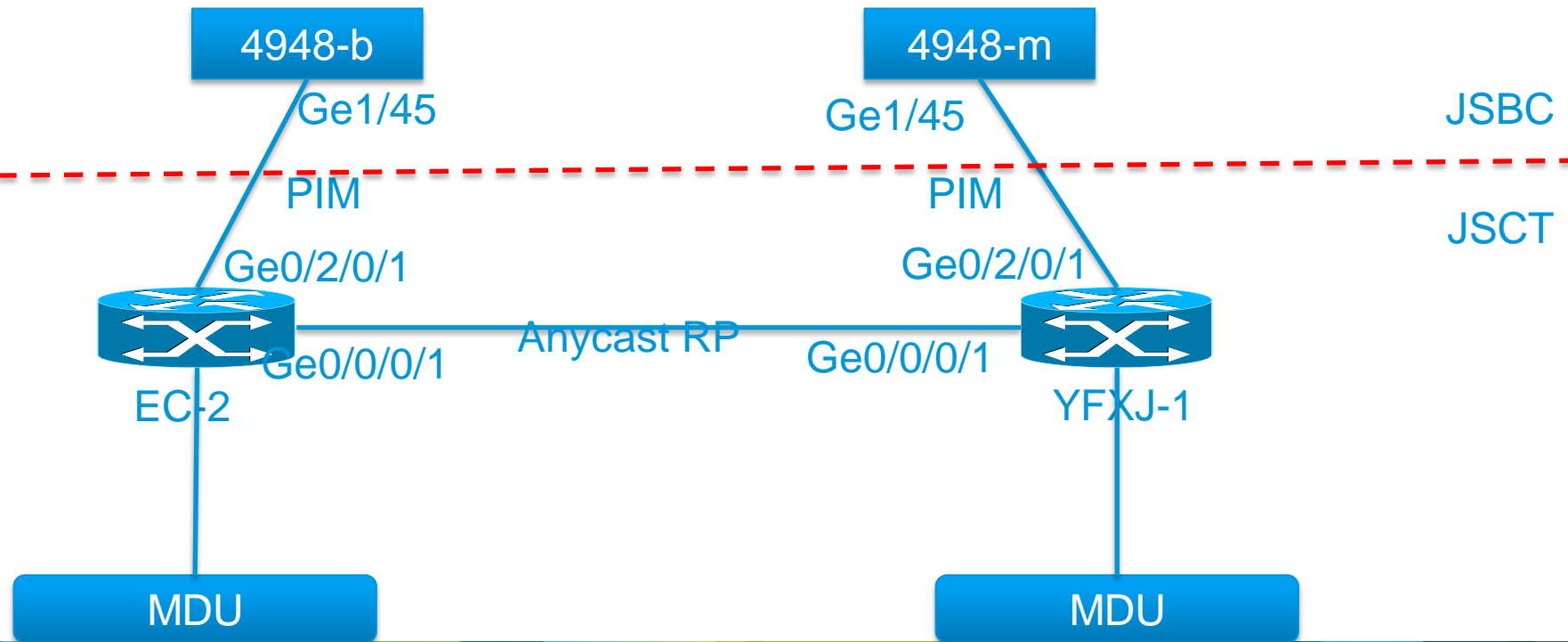
Problem Description

- ASR9K installed reload SMU on MW.
- After the device reloaded, the multicast service down in whole province.
- Based on the network HA design, the service can be recovered through the backup router. But it did not work.

Network Topology

--got from customer side

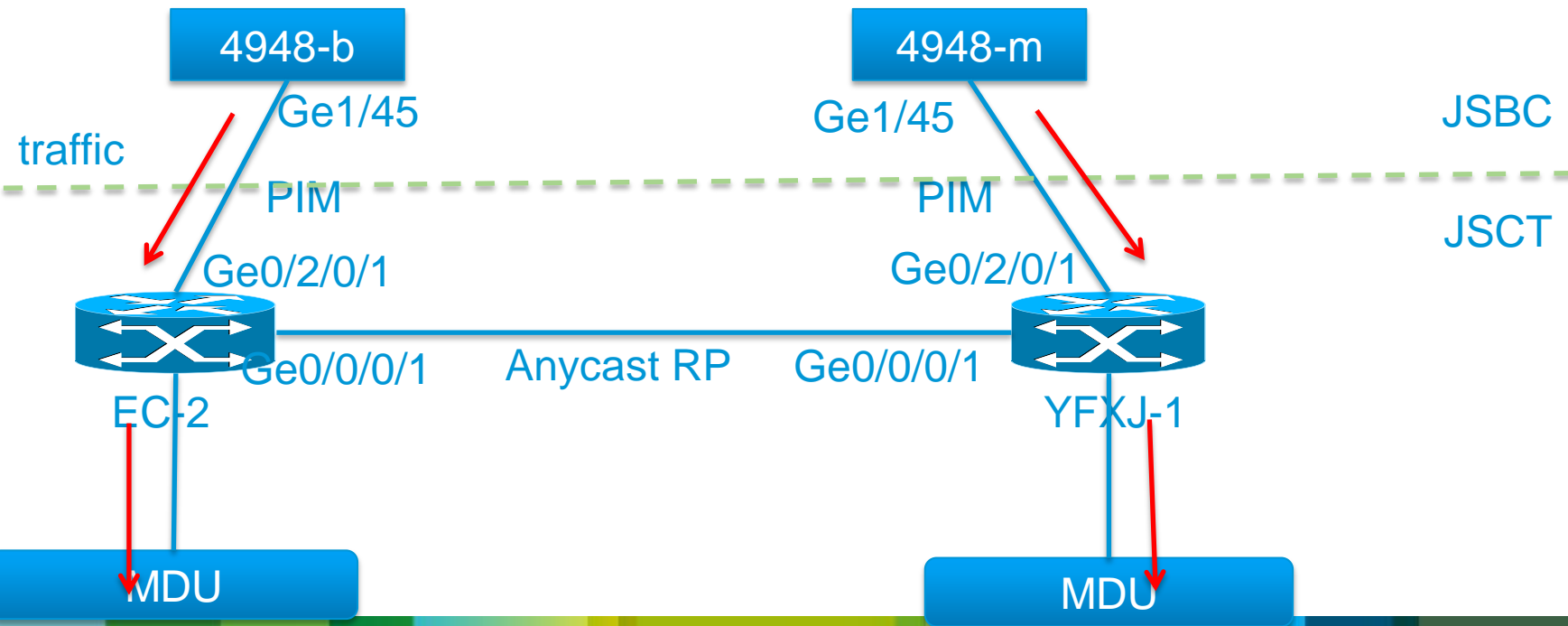
On my customer network two asr9k device is running Anycast RP to provide the protection for RP. IGP protocol is OSPF, running between them. Each device connected individually to one 4948 switch which belonged to content provider jiangsu broadcast. Between asr9k and 4948 is running PIM protocol.



Service traffic description

Normally each device received the traffic from directly connected switch, and forwarding to downstream MDU server. Each asr9k is manually configured static routes which pointing to multicast source ip address.

MDU is the server used to record and transform the multicast traffic to unicast. But only the MDU under EC-2 is working. Another site MDU is cold standby.



Status of PIM entry

PIM entry on EC-2 asr9k

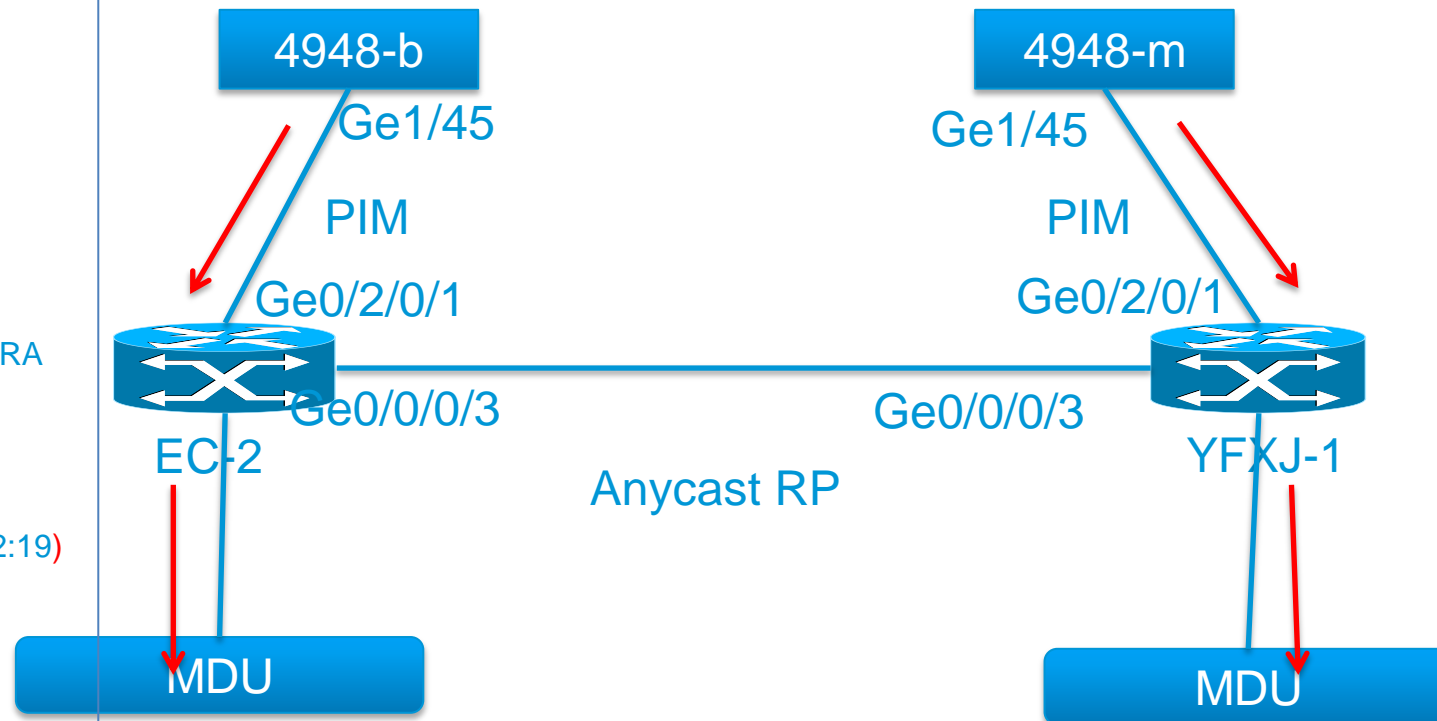
```
(* ,233.19.204.119) SM Up: 2w1d RP: 221.231.144.92*
JP: Join(never) RPF: Decapstunnel0,221.231.144.92 Flags: LH
BVI130          1w1d   fwd LI LH
GigabitEthernet0/2/0/1  1d05h  fwd Join(00:03:11)
GigabitEthernet0/4/0/19  04:19:45 fwd LI LH

(172.27.111.153,233.19.204.119)RPT SM Up: 1d05h RP: 221.231.144.92*
JP: Prune(never) RPF: Decapstunnel0,221.231.144.92 Flags: KAT(00:02:19) RA
RR (00:03:24)
GigabitEthernet0/2/0/1  1d05h  off Prune(00:03:11)

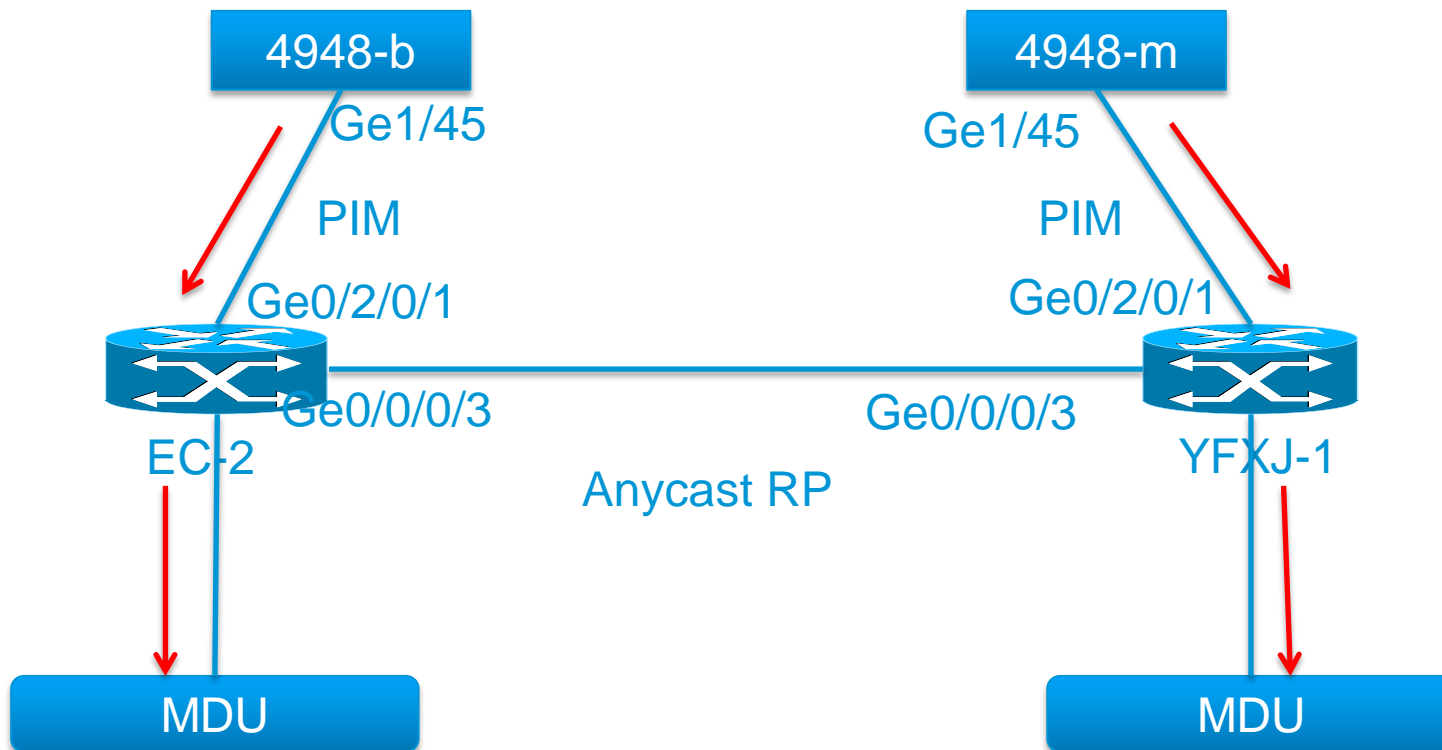
(172.27.111.153,233.19.204.119)SPT SM Up: 1d05h
JP: Join(00:00:10) RPF: GigabitEthernet0/2/0/1,10.99.99.14 Flags: KAT(00:02:19)
RA RR (00:03:24)
TenGigE0/0/0/1          1d05h  fwd Join(00:03:07)
```

From the pim entry we can know

1. EC-2 received the register packets from First hop router. The Flag entry has RR.
2. The (S,G) entry is created by local



Status of PIM entry



PIM entry on YFXJ-1

```
(* ,233.19.204.119) SM Up: 1d06h RP: 221.231.144.92*
JP: Join(never) RPF: Decapstunnel0,221.231.144.92 Flags:
  BVI32          1d06h  off LI
  GigabitEthernet0/2/0/1  1d06h  fwd Join(00:02:43)

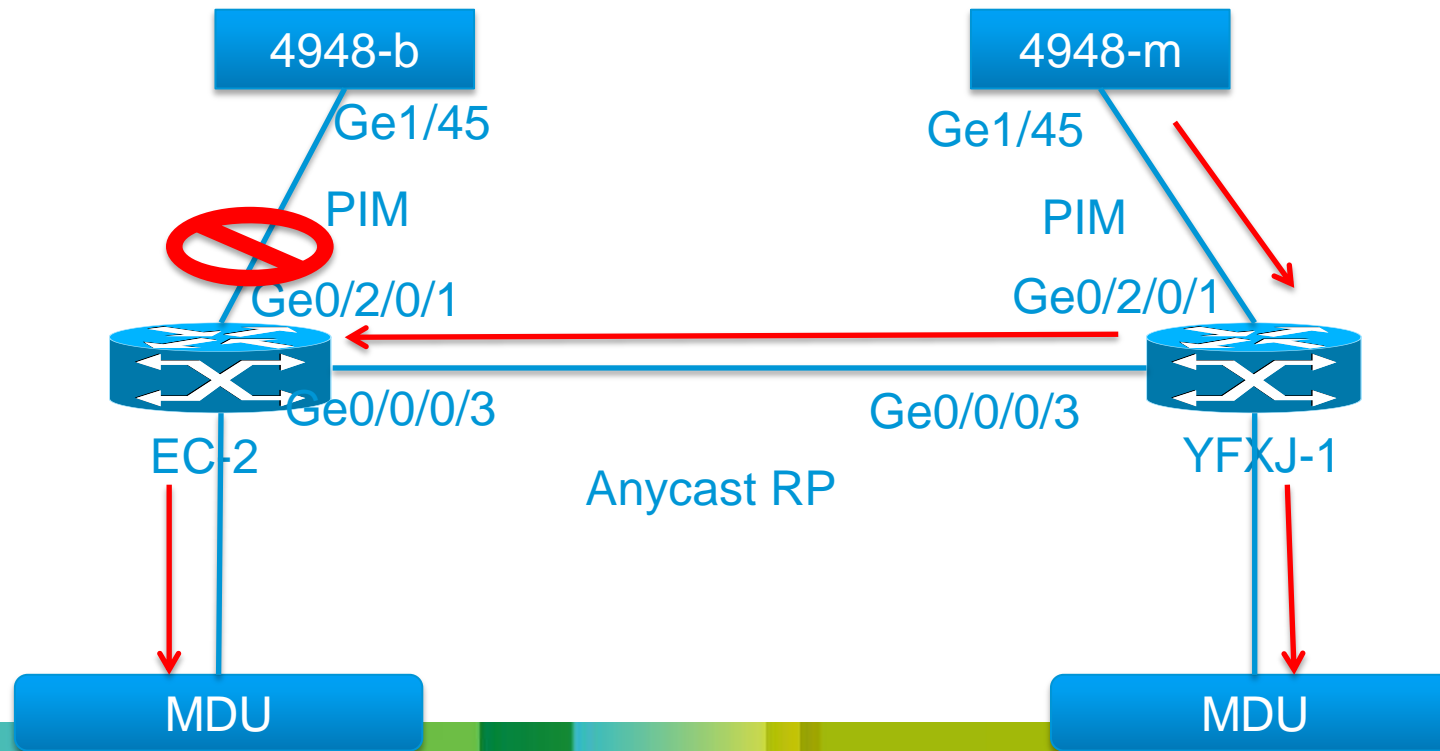
(172.27.111.153,233.19.204.119)RPT SM Up: 1d06h RP:
221.231.144.92*
JP: Prune(never) RPF: Decapstunnel0,221.231.144.92 Flags:
KAT(00:03:12) RA
  GigabitEthernet0/2/0/1  1d06h  off Prune(00:02:43)

(172.27.111.153,233.19.204.119)SPT SM Up: 1d06h
JP: Join(00:00:07) RPF: GigabitEthernet0/2/0/1,10.99.99.10 Flags:
KAT(00:03:12) E RA
  TenGigE0/0/0/3          10:27:40  fwd Join(00:02:41)
  Bundle-Ether1.200      10:27:40  fwd Join(00:02:51)
```

1. The (S,G) entry is created by remote MDSP info
2. Did not receive the register packets

HA Design

Based on the design, in case the link between EC-2 and 4948-b down, EC-2 should receive the MSDP sa-cache from YFXJ-1 and recreate the (S,G) entry then grasp the multicast traffic from YFXJ-1. The multicast source ip address is learned by OSPF routes from YFXJ-1. But actually it is unexpected.

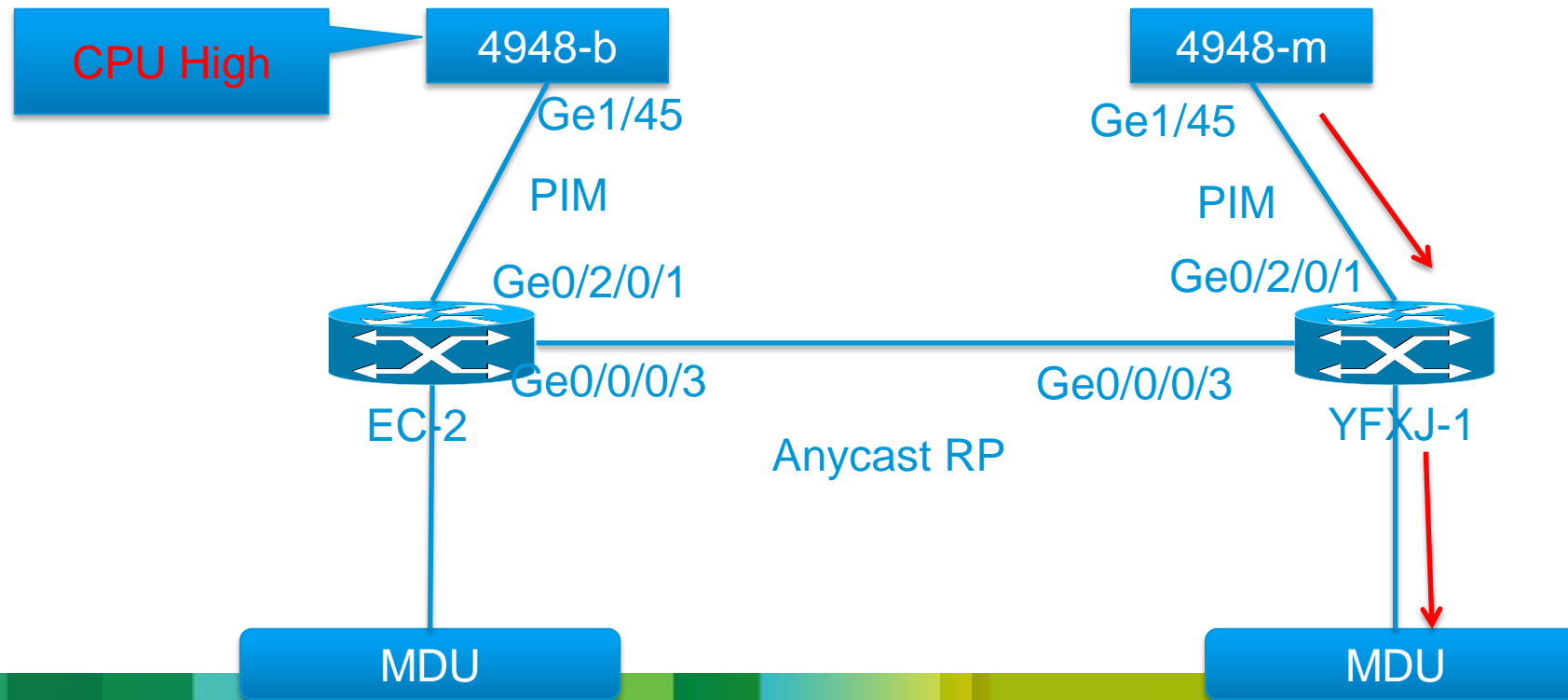


Issue occure

After EC-2 reload then boot up, check the pim entry on EC-2, did not observe any (S,G) entry. Check the interface connected to 0/2/0/1, do not receive any multicast traffic from 4948-b.

Ask JSBC to check the status of 4948-b, the device is hang due to high cpu utilization.

Then we shutdown the interface G0/2/0/1 to change the RPF interface and expect to recover the service.



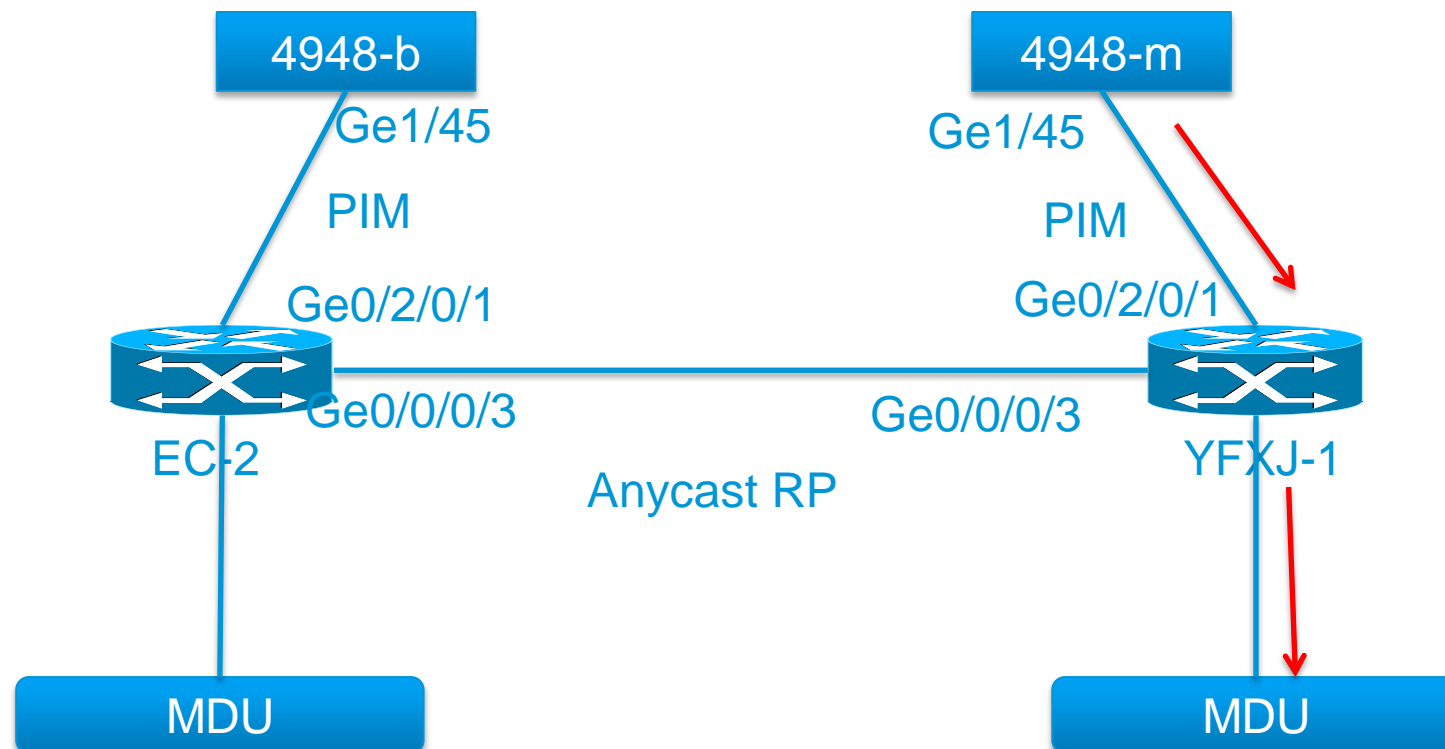
Issue occure

But the issue still exist

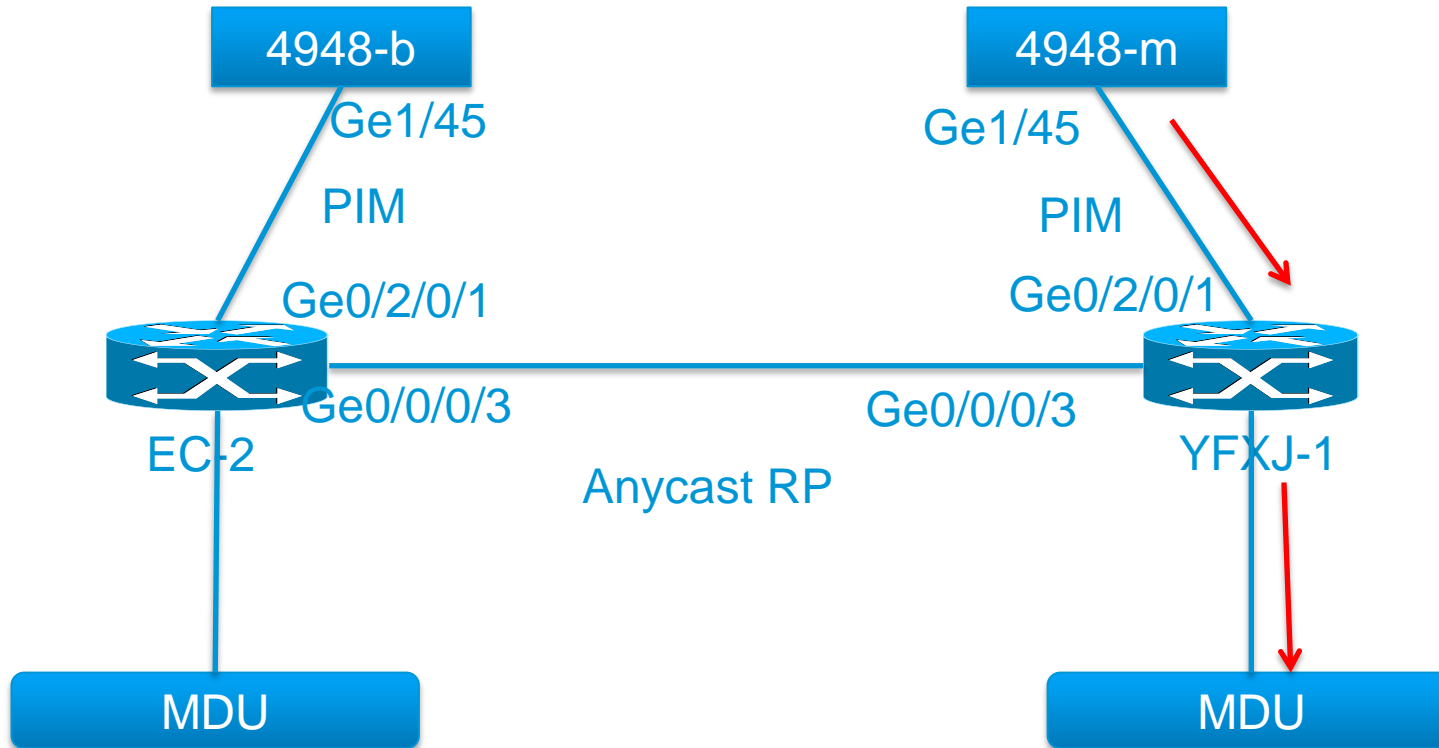
Firstly check MSDP info on EC-2, it did not receive any sa cache message from YFXJ-1.

Then we check the MSDP info on YFXJ-1, it did no generate any MSDP sa cache.

One strange clue we found on the YFXJ-1 device is all the (S,G) entry on YFXJ-1 is without any FLAG parameter.



Issue occure



PIM entry on YFXJ-1

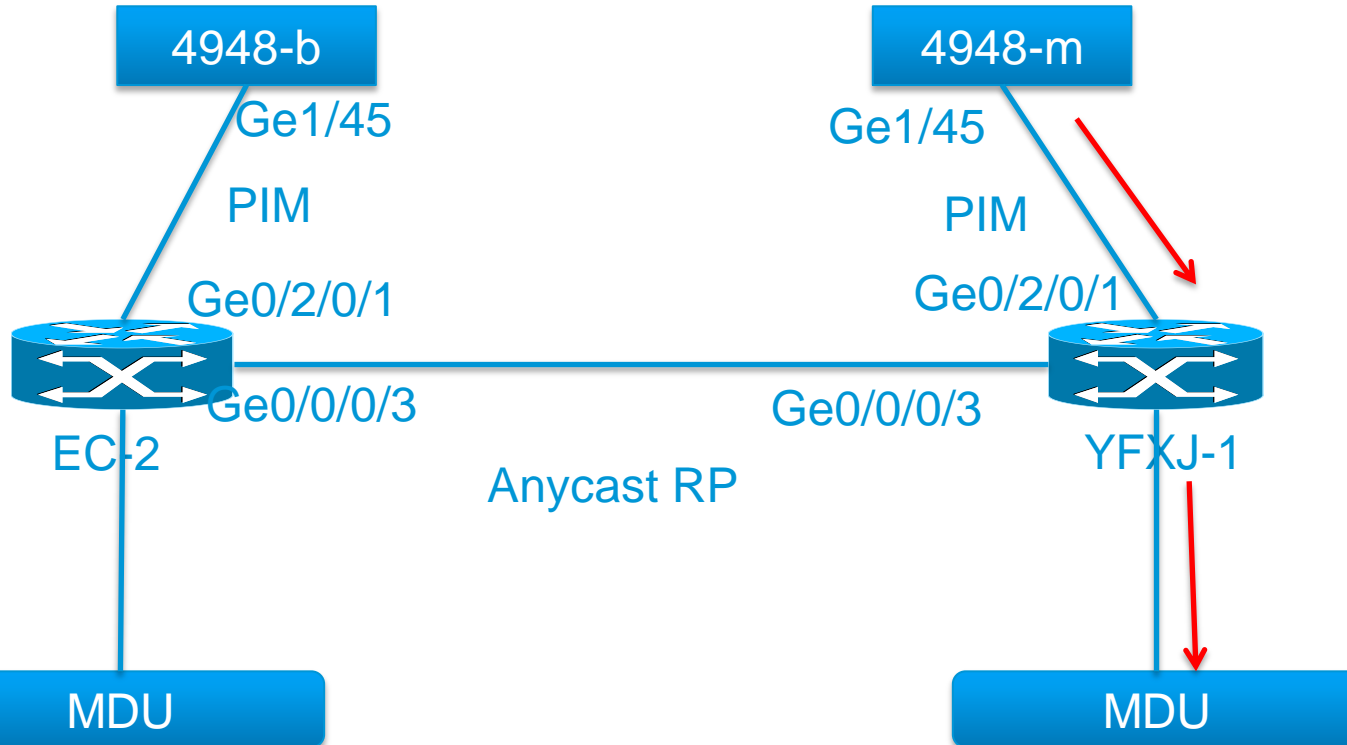
```
(*,233.19.204.119) SM Up: 00:24:32 RP: 221.231.144.92*
JP: Join(never) RPF: Decapstunnel0,221.231.144.92 Flags:
BVI32 00:24:32 off LI
GigabitEthernet0/2/0/1 00:24:05 fwd Join(00:03:05)
```

```
(172.27.111.153,233.19.204.119)RPT SM Up: 00:24:05 RP:
221.231.144.92*
JP: Prune(never) RPF: Decapstunnel0,221.231.144.92 Flags:
GigabitEthernet0/2/0/1 00:24:05 off Prune(00:03:05)
```

```
(172.27.111.153,233.19.204.119)SPT SM Up: 00:24:02
JP: Join(00:00:45) RPF: GigabitEthernet0/2/0/1,10.99.99.10 Flags:
TenGigE0/0/0/3 00:23:48 fwd Join(00:03:23)
Bundle-Ether1.200 00:24:02 fwd Join(00:03:28)
```

Issue occure

Due to missed the FLAG 'RR' On (S,G) entry, so it can no info device to generate the MSDP SA Cache.
'Clear pim topology' still can not recover.



PIM entry on YFXJ-1

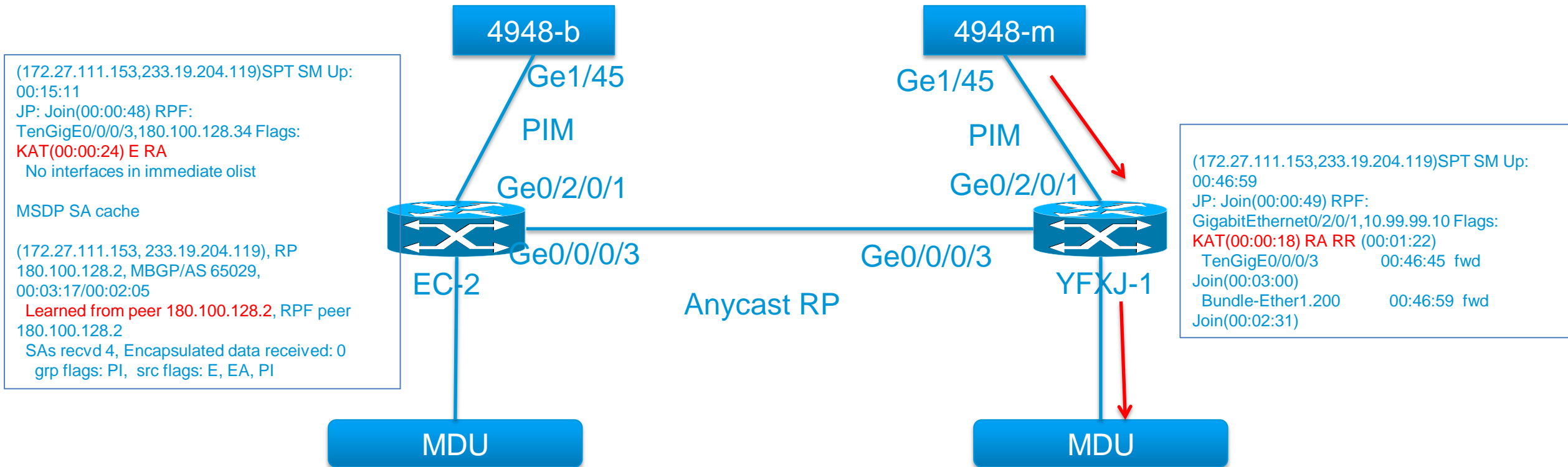
```
(* ,233.19.204.119) SM Up: 00:24:32 RP: 221.231.144.92*
JP: Join(never) RPF: Decapstunnel0,221.231.144.92 Flags:
  BVI32          00:24:32 off LI
  GigabitEthernet0/2/0/1 00:24:05 fwd Join(00:03:05)

(172.27.111.153,233.19.204.119)RPT SM Up: 00:24:05 RP:
221.231.144.92*
JP: Prune(never) RPF: Decapstunnel0,221.231.144.92 Flags:
  GigabitEthernet0/2/0/1 00:24:05 off Prune(00:03:05)

(172.27.111.153,233.19.204.119)SPT SM Up: 00:24:02
JP: Join(00:00:45) RPF: GigabitEthernet0/2/0/1,10.99.99.10 Flags:
  TenGigE0/0/0/3      00:23:48 fwd Join(00:03:23)
  Bundle-Ether1.200   00:24:02 fwd Join(00:03:28)
```

Issue Recovery

Only the recovery method is reload the CPU high device 4948-b.
 After reload the pim entry on YFXJ-1 will change to expectable status.
 Why reload the device 4948-b, will impact the status on YFXJ-1?



Meeting with customer

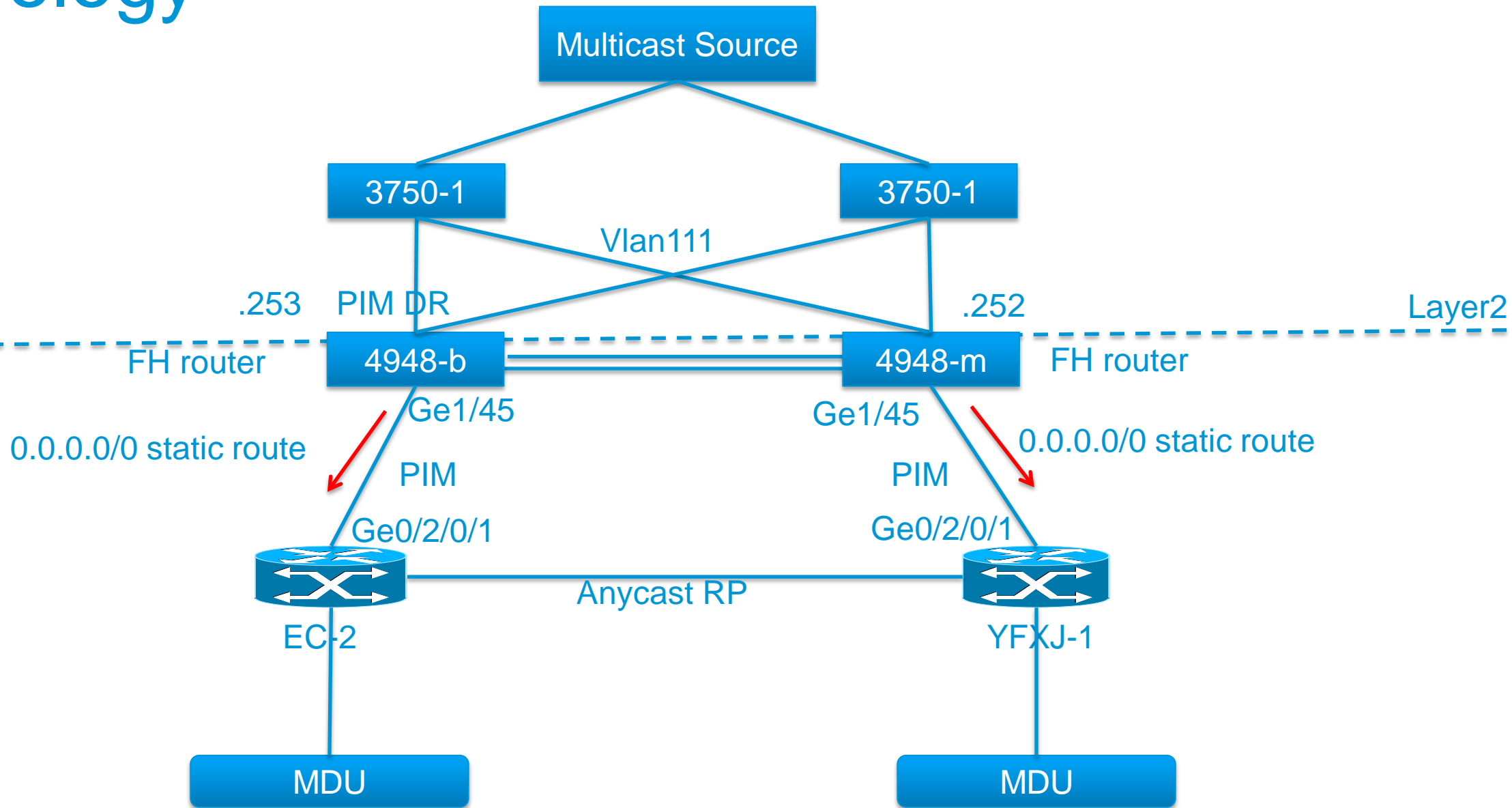
Meeting with customer to discuss the issue, expect to dig out the reason.

1. Firstly we ask customer to provide the detail network topology about multicast source.

2. Then we ask customer to capture the detail multicast related command on two 4948 devices.

after that , we figured out the new network topology.

Topology



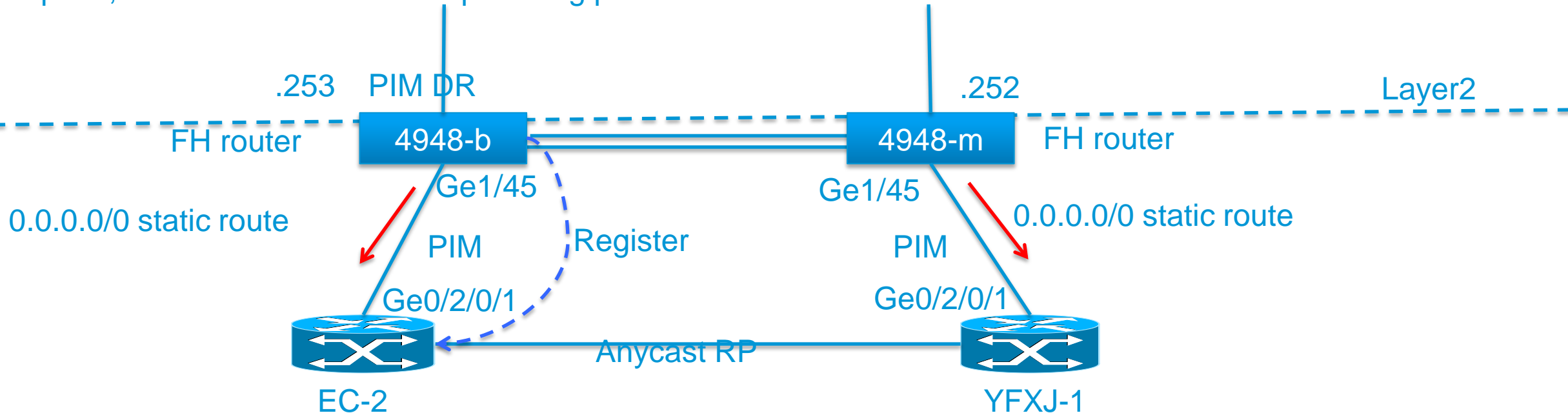
Topology

1. Two 4948 both connected to the same multicast source via VLAN 111. need selected DR interface, 4948-b's interface address is larger then 4948-m, it wins the DR selection.
2. 4948 device is the first hop router, directly connect to multicast source.
3. On 4948 device only configured one default static routes which next hop address is opposite asr9k interconnection interface address.
4. Not any backup routes point to RP's address.

Topology

The 4948-b send out register packets to nearest RP (EC-2) via the default static routes. When the SPT tree setup between RP and first hop router, RP will send out Register stop message to info first hop router stop send out register packets with multicast content.

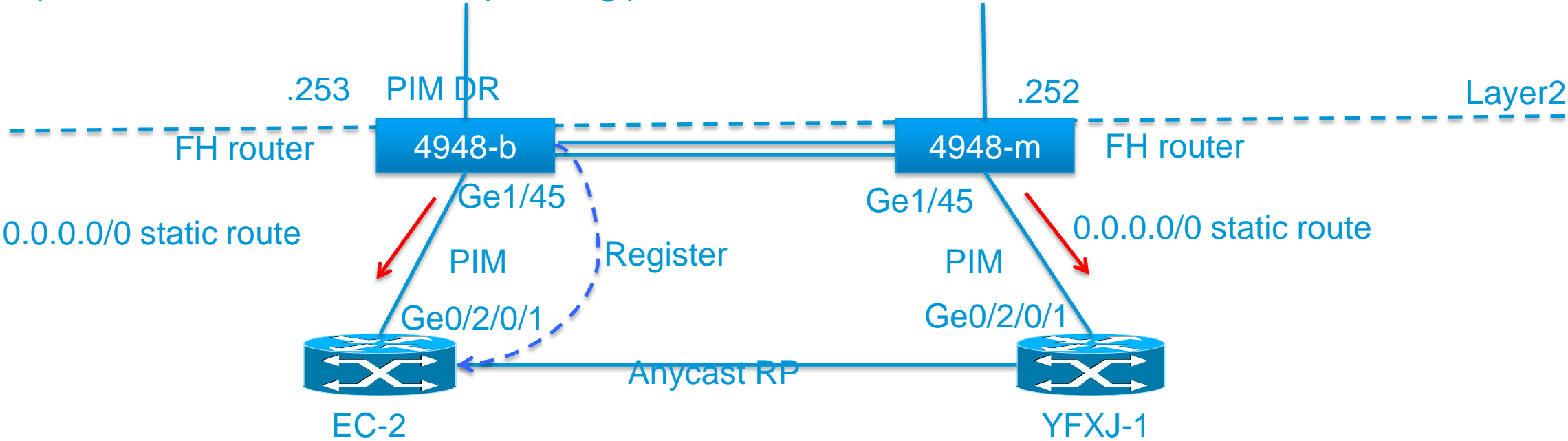
after received the stop message, First hop router will periodically(Register-Stop Timer) send Null-register packets to RP to refresh the the Register-Stop information at the DR. If the Register-Stop Timer actually expires, the DR will resume encapsulating packets from the source to the RP.



Root Cause

The 4948-b send out register packets to nearest RP (EC-2) via the default static routes. When the SPT tree setup between RP and first hop router, RP will send out Register stop message to info first hop router stop send out register packets with multicast content.

after received the stop message, First hop router will periodically(Register-Stop Timer) send Null-register packets to RP to refresh the the Register-Stop information at the DR. If the Register-Stop Timer actually expires, the DR will resume encapsulating packets from the source to the RP.



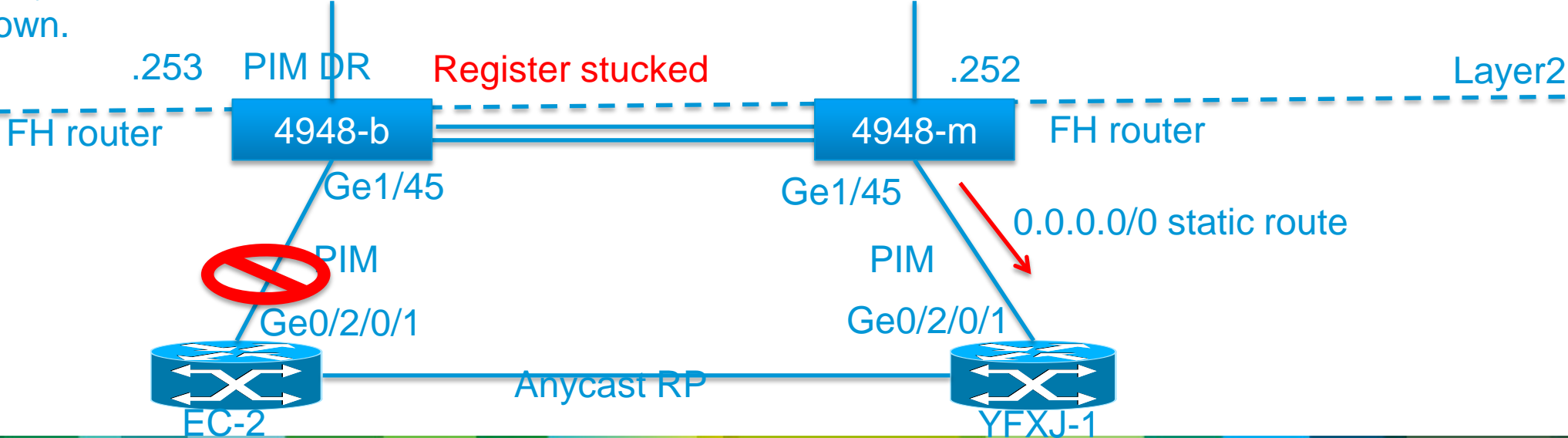
Root Cause

When the interface between 4948-b and EC-2 is down, 4948-b still is DR for VLAN 111, so it still need send out register packets to RP, but did not find any routes to RP. So the register packets stuck in CPU more and more. Leads to CPU high.

The same reason, due to YFXJ-1 RP did not receive any register packets from DR, the PIM entry can not set any flag.

In RP KeepaliveTimer(S,G) is restarted at the RP when packets arrive on the proper register tunnel interface. So the KAT and RR flag only set after receive the register packets from DR.

When the 4948-b restart the RP will be switched to 4948-m, so 4948-m sends out the register packets, and service recovery. But if the 4948-b boot up. The service still will be down, if the interface between EC-2 and 4948-b still down.



Conclusion

Propose customer to set up backup static routes pointed to RP.

Thank you.

