

ACI交換矩陣內轉發故障排除 — MultiPod轉發

目錄

[簡介](#)

[背景資訊](#)

[多Pod轉發概述](#)

[多Pod元件](#)

[多Pod的拓撲示例](#)

[對多Pod轉發進行故障排除的一般工作流程](#)

[多Pod單播故障排除工作流](#)

[1. 確認入口枝葉收到資料包。使用「工具」部分中顯示的ELAM CLI工具以及4.2中提供的ereport輸出。還使用ELAM Assistant應用。](#)

[2. 入口枝葉是否將目標作為入口VRF中的終結點學習？如果沒有，有路由嗎？](#)

[ELAM助理配置](#)

[驗證轉發決策](#)

[3. 確認主幹上的目標IP存在於COOP中，以便代理請求生效。](#)

[4. 多Pod主幹代理轉發決策](#)

[5. 檢驗主幹上的BGP EVPN](#)

[6. 驗證目標Pod中脊柱上的COOP。](#)

[7. 檢驗出口枝葉是否具有本地學習項。](#)

[使用fTriage驗證端到端流量](#)

[EP不在COOP中的代理請求](#)

[收集ARP驗證](#)

[多Pod故障排除場景#1 \(單播 \)](#)

[拓撲故障排除](#)

[原因：COOP中缺少終結點](#)

[其他可能的原因](#)

[Multi-Pod廣播、未知的單播和組播\(BUM\)轉發概述](#)

[GUI中的BD GIPo](#)

[IPN多點傳送控制平面](#)

[IPN多點傳送資料平面](#)

[虛擬RP配置](#)

[多Pod廣播、未知的單播和組播\(BUM\)故障排除工作流](#)

[1. 首先確認交換矩陣是否真正將流視為多目的地。](#)

[2. 確定BD GIPo。](#)

[3. 驗證該GIPo的IPN上的組播路由表。](#)

[多Pod故障排除場景#2 \(BUM流 \)](#)

[可能的原因1: 多台路由器擁有PIM RP地址](#)

[可能原因2: IPN路由器無法獲取RP地址的路由](#)

[可能原因3: IPN路由器未安裝GIPo路由或RPF指向ACI](#)

[其他參考資料](#)

簡介

本文檔介紹瞭解ACI多Pod轉發方案並對其進行故障排除的步驟。

背景資訊

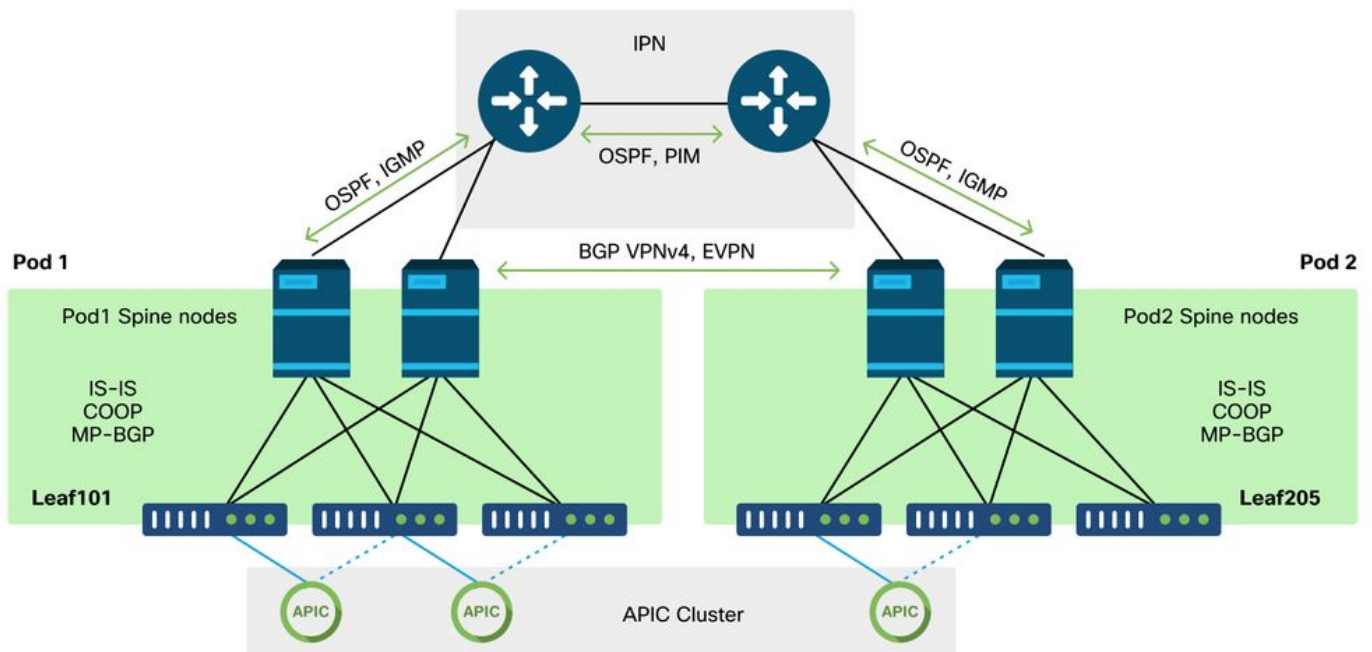
本文中的資料摘自 [思科以應用為中心的基礎設施第二版故障排除](#) 書，特別是 [交換矩陣內轉發 — 多Pod轉發](#) 章節。

多Pod轉發概述

本章將介紹如何對多Pod環境中各Pod之間的連線不正常的情況進行故障排除

在檢視具體的故障排除示例之前，花些時間瞭解高級別Multi-Pod元件非常重要。

多Pod元件



與傳統ACI交換矩陣類似，多Pod交換矩陣仍被視為單個ACI交換矩陣，並依賴單個APIC集群進行管理。

在每個單獨的Pod中，ACI利用重疊中的協定作為傳統交換矩陣。其中包括IS-IS，用於交換TEP資訊以及組播傳出介面(OIF)選擇、COOP用於全域性終端儲存庫，以及BGP VPNv4用於通過交換矩陣分發外部路由器。

多Pod基於這些元件構建，因為它必須將每個Pod連線在一起。

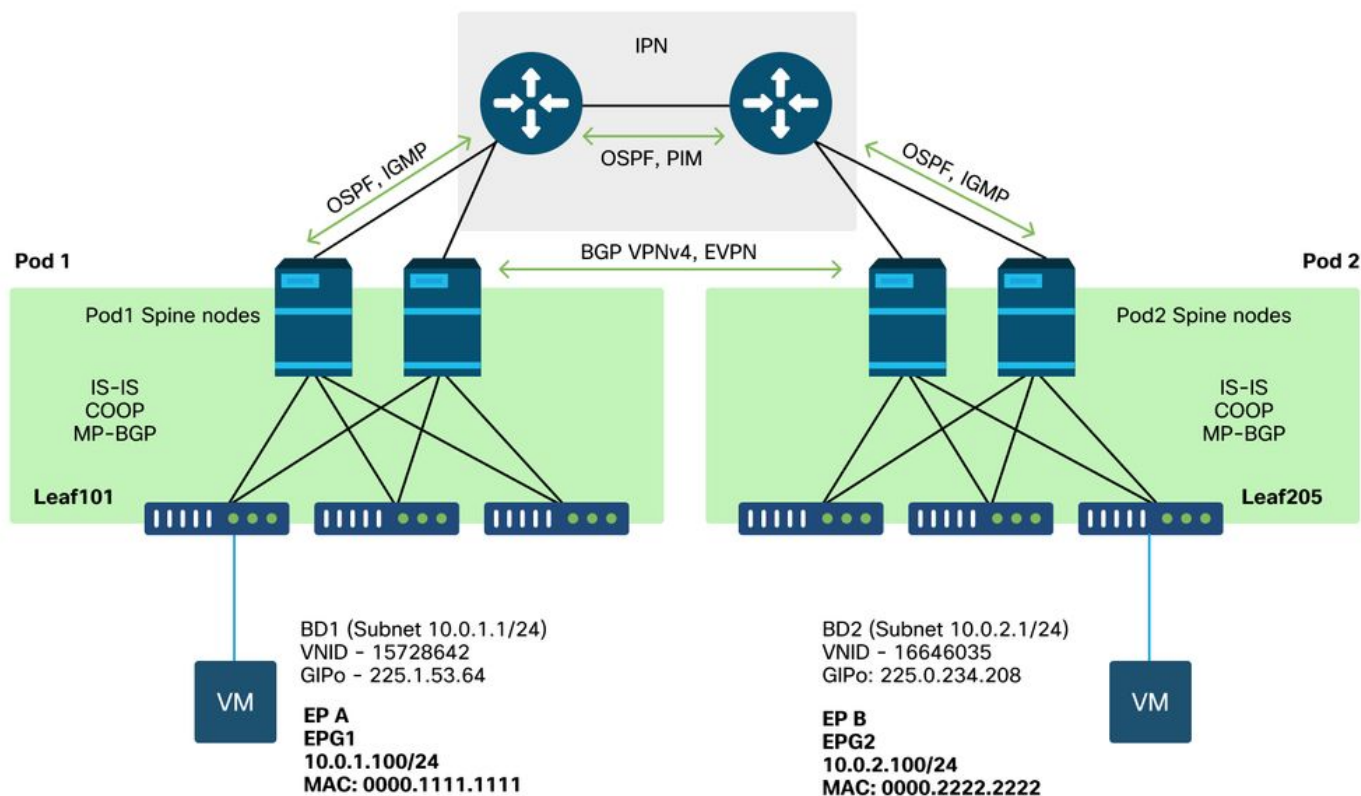
- 為了交換有關遠端Pod中TEP的路由資訊，OSPF用於通過IPN通告彙總TEP池。
- 為了交換從一個Pod獲知的外部路由到另一個Pod，BGP VPNv4地址系列在脊柱節點之間擴展。每個Pod都成為一個單獨的路由反射器集群。
- 為了跨Pod同步終端以及儲存在COOP中的其他資訊，BGP EVPN地址系列在脊柱節點之間擴展。
- 最後，為了處理跨Pod的廣播、未知單播和組播(BUM)流量的泛洪，每個Pod中的主幹節點充當

IGMP主機，IPN路由器通過雙向PIM交換組播路由資訊。

大部分多Pod故障排除場景和工作流程類似於單Pod ACI交換矩陣。此多Pod部分將重點介紹單Pod與多Pod轉發之間的區別。

多Pod的拓撲示例

與排除任何場景一樣，重要的是首先瞭解預期狀態。本章示例參考此拓撲。



對多Pod轉發進行故障排除的一般工作流程

在高級別，當調試多Pod轉發問題時，可以評估以下步驟：

1. 流是單播還是多目標？請記住，即使預期流在工作狀態下是單播，如果ARP未解析，則它是多目的地流。
2. 流是路由還是橋接？傳統上，從ACI的角度來看，路由流是指目的MAC地址是路由器MAC地址且由ACI上配置的網關擁有的任何流。此外，如果禁用ARP泛洪，則入口枝葉將基於目標IP地址進行路由。如果目的MAC地址不屬於ACI，則交換機將根據MAC地址轉發或遵循在網橋域上配置的「未知單播」行為。
3. 入口枝葉是否正在丟棄流？fTriage和ELAM是確認這一點的最佳工具。

如果流是第3層單播：

1. 入口枝葉是否具有與源EPG相同的VRF中的目標IP的終端學習？如果是，則此路由將始終優先於任何獲知的路由。枝葉將直接轉發到獲知端點的隧道地址或出口介面。
2. 如果沒有終端學習，輸入枝葉是否具有已設定「沉浸式」標誌的目標路由？這表示目的地子網已配置為橋接域子網，並且下一跳應是本地Pod中的主幹代理。

3. 如果沒有無處不在的路由，則最後的手段是通過L3Out獲知的任何路由。此部分與單Pod L3Out轉發相同。

如果流是第2層單播：

1. 入口枝葉是否具有與源EPG相同的網橋域中的目標MAC地址的終端學習？如果是，枝葉將轉發到遠端隧道IP或從其獲知端點的本地介面。
2. 如果在源網橋域中沒有獲知目的MAC地址，則枝葉將根據BD設定為「unknown-unicast」行為進行轉發。如果設定為「泛洪」，則枝葉將泛洪到分配給網橋域的GIPo組播組。本地和遠端Pod應獲得一個泛洪副本。如果設定為「Hardware Proxy」，則將幀傳送到主幹進行代理查詢，然後根據主幹的COOP條目進行轉發。

由於單播的故障排除輸出與BUM相比有很大不同，因此在進入BUM之前會考慮單播的工作輸出和場景。

多Pod單播故障排除 workflow

按照拓撲結構，遍歷從leaf205上的10.0.2.100到leaf101上的10.0.1.100的流。

請注意，繼續此處之前，請務必確認來源是否已為闡道（為路由流量）或目的地MAC位址（為橋接流量）解析ARP

1. 確認入口枝葉收到資料包。使用「工具」部分中顯示的ELAM CLI工具以及4.2中提供的ereport輸出。還使用ELAM Assistant應用。

```
module-1# debug platform internal tah elam asic 0
module-1(DBG-elam)# trigger reset
module-1(DBG-elam)# trigger init in-select 6 out-select 1
module-1(DBG-elam-insel6)# set outer ipv4 src_ip 10.0.2.100 dst_ip 10.0.1.100
module-1(DBG-elam-insel6)# start
module-1(DBG-elam-insel6)# status
ELAM STATUS
=====
Asic 0 Slice 0 Status Armed
Asic 0 Slice 1 Status Triggered
```

請注意，ELAM已觸發，可確認輸入交換器上已收到封包。現在看一下報告中的幾個欄位，因為輸出內容非常豐富。

```
=====
=====
Captured Packet
=====
=====
-----
Outer Packet Attributes
-----
-----
Outer Packet Attributes      : l2uc ipv4 ip ipuc ipv4uc
Opcode                       : OPCODE_UC
-----
```

```

-----
Outer L2 Header
-----
-----
Destination MAC          : 0022.BDF8.19FF
Source MAC               : 0000.2222.2222
802.1Q tag is valid     : yes( 0x1 )
CoS                      : 0( 0x0 )
Access Encap VLAN       : 1021( 0x3FD )
-----
-----
Outer L3 Header
-----
-----
L3 Type                  : IPv4
IP Version               : 4
DSCP                     : 0
IP Packet Length        : 84 ( = IP header(28 bytes) + IP payload )
Don't Fragment Bit      : not set
TTL                      : 255
IP Protocol Number      : ICMP
IP CheckSum              : 10988( 0x2AEC )
Destination IP          : 10.0.1.100
Source IP                : 10.0.2.100

```

報告中包含更多有關資料包去向的資訊，但ELAM助理應用程式目前對於解釋此資料更有用。此流程的ELAM Assistant輸出將在本章稍後部分顯示。

2.入口枝葉是否將目標作為入口VRF中的終結點學習？如果沒有，有路由嗎？

```

a-leaf205# show endpoint ip 10.0.1.100 detail
Legend:
s - arp                H - vtep                V - vpc-attached      p - peer-aged
R - peer-attached-rl  B - bounce                S - static            M - span
D - bounce-to-proxy   O - peer-attached        a - local-aged        m - svc-mgr
L - local              E - shared-service
-----+-----+-----+-----+-----+
VLAN/          Encap          MAC Address          MAC Info/
Interface      Endpoint Group      VLAN                IP Address           IP Info
Domain
Info
-----+-----+-----+-----+-----+

```

上述命令中沒有輸出表示未獲知目標IP。然後檢查路由表。

```

a-leaf205# show ip route 10.0.1.100 vrf Prod:Vrf1
IP Route Table for VRF "Prod:Vrf1"
'*' denotes best ucast next-hop
'***' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

10.0.1.0/24, ubest/mbest: 1/0, attached, direct, pervasive
  *via 10.0.120.34%overlay-1, [1/0], 01:55:37, static, tag 4294967294
    recursive next hop: 10.0.120.34/32%overlay-1

```

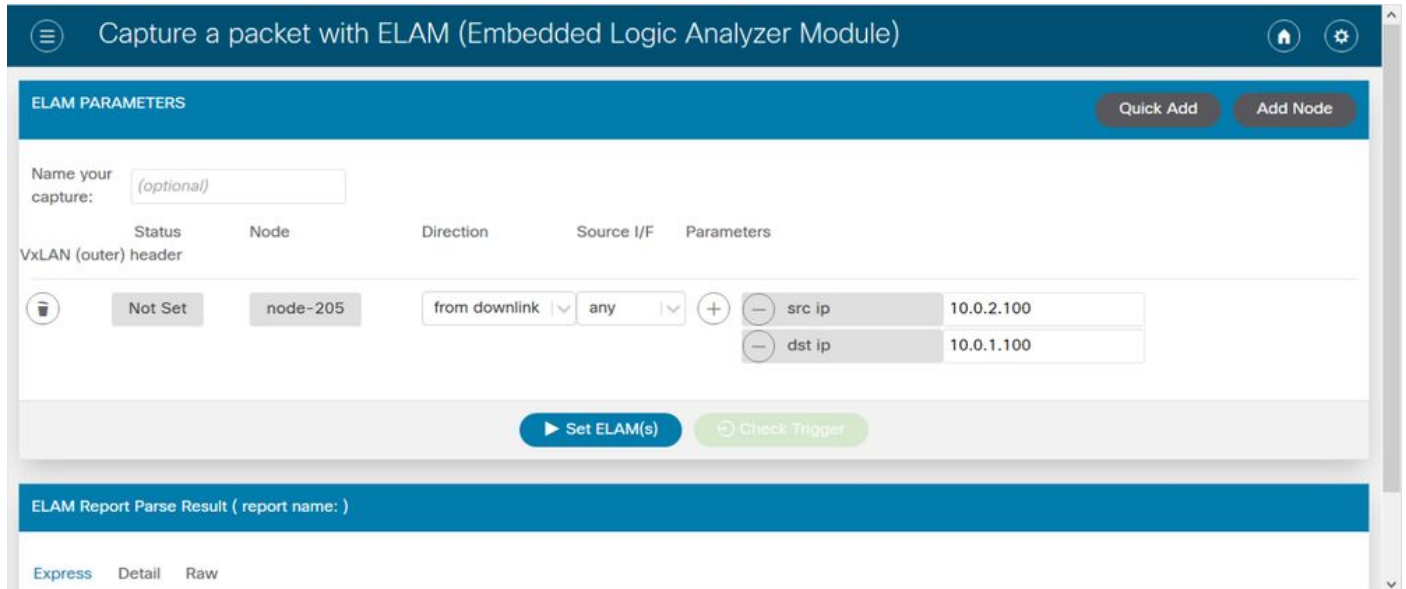
在以上輸出中，將會看到沈浸式標誌，表示這是網橋域子網路由。下一跳應是主幹上的任播代理地址。

```
a-leaf205# show isis dtep vrf overlay-1 | grep 10.0.120.34
10.0.120.34      SPINE    N/A      PHYSICAL, PROXY-ACAST-V4
```

請注意，如果在通道或實體介面上得知端點，則會優先使用，導致封包直接在那裡轉送。有關詳細資訊，請參閱本書的「外部轉發」一章。

使用ELAM助手確認上述輸出中顯示的轉發決策。

ELAM助理配置



驗證轉發決策

Packet Forwarding information	
Forward Result	
Destination Type	To another ACI node (LEAF, AVS/AVE etc.)
Destination TEP	10.0.120.34 (IPv4 Spine-Proxy)
Destination Physical Port	eth1/53
Contract	
Destination EPG pcTag (dclass)	0x1 / 1 (pcTag 1 is to ignore contract for special packets such as Spine-Proxy, ARP, Multicast etc..)
Source EPG pcTag (sclass)	0xC001 / 49153 (Prod.ap1.epg2)
Contract was applied	0 (Contract was not applied on this node)
Drop	
Drop Code	no drop

上面的輸出顯示，入口枝葉正在將資料包轉發到IPv4主幹代理地址。這是預期會發生的。

3. 確認主幹上的目標IP存在於COOP中，以便代理請求生效。

有多種方法可以獲取主幹上的COOP輸出，例如，使用「show coop internal info ip-db」命令檢視它：

```
a-spine4# show coop internal info ip-db | grep -B 2 -A 15 "10.0.1.100"
```

```
-----  
IP address : 10.0.1.100  
Vrf : 2392068 <-- This vnid should correspond to vrf where the IP is learned. Check operational  
tab of the tenant vrfs  
Flags : 0x2  
EP bd vnid : 15728642  
EP mac : 00:00:11:11:11:11  
Publisher Id : 192.168.1.254  
Record timestamp : 12 31 1969 19:00:00 0  
Publish timestamp : 12 31 1969 19:00:00 0  
Seq No: 0  
Remote publish timestamp: 09 30 2019 20:29:07 9900483  
URIB Tunnel Info  
Num tunnels : 1  
    Tunnel address : 10.0.0.34 <-- When learned from a remote pod this will be an External  
Proxy TEP. We'll cover this more  
    Tunnel ref count : 1  
-----
```

要在脊柱上運行的其他命令：

第2層條目的查詢COOP:

```
moquery -c coopEpRec -f 'coop.EpRec.mac=="00:00:11:11:22:22"
```

查詢I3條目的COOP並獲取父級I2條目：

```
moquery -c coopEpRec -x rsp-subtree=children 'rsp-subtree-  
filter=eq(coopIpv4Rec.addr,"192.168.1.1")' rsp-subtree-include=required
```

僅針對第3層條目的查詢COOP:

```
moquery -c coopIpv4Rec -f 'coop.Ipv4Rec.addr=="192.168.1.1"'
```

多資料查詢的有用之處在於，它們還可以直接在APIC上運行，並且使用者可以檢視在雞舍中擁有記錄的每個骨幹。

4.多Pod主幹代理轉發決策

如果脊柱的COOP條目指向本地Pod中的隧道，則轉發基於傳統的ACI行為。

請注意，可以通過從APIC運行來驗證交換矩陣中TEP的所有者：`moquery -c ipv4Addr -f 'ipv4.Addr.addr=="<tunnel address>"`

在代理方案中，隧道下一跳是10.0.0.34。此IP地址的所有者是誰？：

```
a-apic1# moquery -c ipv4Addr -f 'ipv4.Addr.addr=="10.0.0.34"' | grep dn  
dn : topology/pod-1/node-1002/sys/ipv4/inst/dom-overlay-1/if-[lo9]/addr-  
[10.0.0.34/32]  
dn : topology/pod-1/node-1001/sys/ipv4/inst/dom-overlay-1/if-[lo2]/addr-  
[10.0.0.34/32]
```

此IP由Pod 1中的兩個主幹節點擁有。這是一個稱為外部代理地址的特定IP。與ACI具有由Pod內的主幹節點擁有的代理地址一樣（請參見本節的步驟2），也有分配給Pod本身的代理地址。可通過運

行以下命令驗證此介面型別：

```
a-apic1# moquery -c ipv4If -x rsp-subtree=children 'rsp-subtree-
filter=eq(ipv4Addr.addr,"10.0.0.34")' rsp-subtree-include=required

...
# ipv4.If
mode          : anycast-v4,external

# ipv4.Addr
addr          : 10.0.0.34/32
dn           : topology/pod-1/node-1002/sys/ipv4/inst/dom-overlay-1/if-[lo9]/addr-
[10.0.0.34/32]
```

「external」標誌表示這是外部代理TEP。

5. 檢驗主幹上的BGP EVPN

應從脊柱上的BGP EVPN匯入雞舍端點記錄。以下命令可用於驗證它是否在EVPN中（如果它已經與遠端Pod外部代理TEP的下一跳在COOP中，則可以假定它來自EVPN）：

```
a-spine4# show bgp l2vpn evpn 10.0.1.100 vrf overlay-1
Route Distinguisher: 1:16777199
BGP routing table entry for [2]:[0]:[15728642]:[48]:[0000.1111.1111]:[32]:[10.0.1.100]/272,
version 689242 dest ptr 0xaf42a4ca
Paths: (2 available, best #2)
Flags: (0x000202 00000000) on xmit-list, is not in rib/evpn, is not in HW, is locked
Multipath: eBGP iBGP

  Path type: internal 0x40000018 0x2040 ref 0 adv path ref 0, path is valid, not best reason:
Router Id, remote nh not installed
AS-Path: NONE, path sourced internal to AS
  192.168.1.254 (metric 7) from 192.168.1.102 (192.168.1.102)
    Origin IGP, MED not set, localpref 100, weight 0
    Received label 15728642 2392068
    Received path-id 1
    Extcommunity:
      RT:5:16
      SOO:1:1
      ENCAP:8
      Router MAC:0200.0000.0000

      Advertised path-id 1
  Path type: internal 0x40000018 0x2040 ref 1 adv path ref 1, path is valid, is best path, remote
nh not installed
AS-Path: NONE, path sourced internal to AS
  192.168.1.254 (metric 7) from 192.168.1.101 (192.168.1.101)
    Origin IGP, MED not set, localpref 100, weight 0
    Received label 15728642 2392068
    Received path-id 1
    Extcommunity:
      RT:5:16
      SOO:1:1
      ENCAP:8
      Router MAC:0200.0000.0000

    Path-id 1 not advertised to any peer
```

請注意，上述命令也可針對MAC位址執行。

-192.168.1.254是在多埠設定期間配置的資料平面TEP。但是請注意，即使在BGP中通告它為NH，實際的下一躍點將是外部代理TEP。

-192.168.1.101和。102是通告此路徑的Pod 1主幹節點。

6.驗證目標Pod中脊柱上的COOP。

可以使用與前面相同的命令：

```
a-spine2# show coop internal info ip-db | grep -B 2 -A 15 "10.0.1.100"
```

```
-----  
IP address : 10.0.1.100  
Vrf : 2392068  
Flags : 0  
EP bd vnid : 15728642  
EP mac : 00:50:56:81:3E:E6  
Publisher Id : 10.0.72.67  
Record timestamp : 10 01 2019 15:46:24 502206158  
Publish timestamp : 10 01 2019 15:46:24 524378376  
Seq No: 0  
Remote publish timestamp: 12 31 1969 19:00:00 0  
URIB Tunnel Info  
Num tunnels : 1  
    Tunnel address : 10.0.72.67  
    Tunnel ref count : 1  
-----
```

通過在APIC上運行以下命令來驗證隧道地址的所有者：

```
a-apic1# moquery -c ipv4Addr -f 'ipv4.Addr.addr=="10.0.72.67"'  
Total Objects shown: 1  
  
# ipv4.Addr  
addr : 10.0.72.67/32  
childAction :  
ctrl :  
dn : topology/pod-1/node-101/sys/ipv4/inst/dom-overlay-1/if-[lo0]/addr-[10.0.72.67/32]  
ipv4CfgFailedBmp :  
ipv4CfgFailedTs : 00:00:00:00.000  
ipv4CfgState : 0  
lcOwn : local  
modTs : 2019-09-30T18:42:43.262-04:00  
monPolDn : uni/fabric/monfab-default  
operSt : up  
operStQual : up  
pref : 0  
rn : addr-[10.0.72.67/32]  
status :  
tag : 0  
type : primary  
vpcPeer : 0.0.0.0
```

上面的命令顯示從COOP到枝葉101的隧道。這意味著枝葉101應該具有目標端點的本地學習。

7.檢驗出口枝葉是否具有本地學習項。

這可通過「show endpoint」命令完成：

```
a-leaf101# show endpoint ip 10.0.1.100 detail
```

Legend:

```
s - arp          H - vtep          V - vpc-attached    p - peer-aged
R - peer-attached-rl B - bounce        S - static          M - span
D - bounce-to-proxy O - peer-attached  a - local-aged     m - svc-mgr
L - local        E - shared-service
```

```
+-----+-----+-----+-----+-----+
-----+-----+-----+-----+-----+
      VLAN/                               Encap           MAC Address       MAC Info/
Interface   Endpoint Group                VLAN             IP Address        IP
Domain                               Info
Info
+-----+-----+-----+-----+-----+
-----+-----+-----+-----+-----+
341                               vlan-1075        0000.1111.1111 LV
po5                Prod:ap1:epg1
Prod:Vrf1                               vlan-1075        10.0.1.100 LV
po5
```

請注意，終端已獲取。應基於已設定VLAN標籤1075的port-channel 5轉發資料包。

使用fTriage驗證端到端流量

如本章「工具」一節所述，fTriage可用於對映現有端到端流量，並瞭解路徑中的每台交換機對資料包的操作。這在大型和更複雜的部署（如多面板）中尤其有用。

請注意，fTriage需要一些時間才能完全運行（可能為15分鐘）。

對示例流運行fTriage時：

```
a-apic1# ftrriage route -ii LEAF:205 -dip 10.0.1.100 -sip 10.0.2.100
```

```
fTriage Status: {"dbgFtrriage": {"attributes": {"operState": "InProgress", "pid": "7297",
"apicId": "1", "id": "0"}}}
```

Starting ftrriage

Log file name for the current run is: ftlog_2019-10-01-16-04-15-438.txt

```
2019-10-01 16:04:15,442 INFO /controller/bin/ftrriage route -ii LEAF:205 -dip 10.0.1.100 -sip
10.0.2.100
```

```
2019-10-01 16:04:38,883 INFO ftrriage: main:1165 Invoking ftrriage with default password
and default username: apic#fallback\admin
```

```
2019-10-01 16:04:54,678 INFO ftrriage: main:839 L3 packet Seen on a-leaf205 Ingress:
Eth1/31 Egress: Eth1/53 Vnid: 2392068
```

```
2019-10-01 16:04:54,896 INFO ftrriage: main:242 ingress encap string vlan-1021
```

```
2019-10-01 16:04:54,899 INFO ftrriage: main:271 Building ingress BD(s), Ctx
```

```
2019-10-01 16:04:56,778 INFO ftrriage: main:294 Ingress BD(s) Prod:Bd2
```

```
2019-10-01 16:04:56,778 INFO ftrriage: main:301 Ingress Ctx: Prod:Vrf1
```

```
2019-10-01 16:04:56,887 INFO ftrriage: pktrec:490 a-leaf205: Collecting transient losses
snapshot for LC module: 1
```

```
2019-10-01 16:05:22,458 INFO ftrriage: main:933 SIP 10.0.2.100 DIP 10.0.1.100
```

```
2019-10-01 16:05:22,459 INFO ftrriage: unicast:973 a-leaf205: <- is ingress node
```

```
2019-10-01 16:05:25,206 INFO ftrriage: unicast:1215 a-leaf205: Dst EP is remote
```

```
2019-10-01 16:05:26,758 INFO ftrriage: misc:657 a-leaf205: DMAC(00:22:BD:F8:19:FF) same
as RMAC(00:22:BD:F8:19:FF)
```

```
2019-10-01 16:05:26,758 INFO ftrriage: misc:659 a-leaf205: L3 packet getting
routed/bounced in SUG
```

```
2019-10-01 16:05:27,030 INFO ftrriage: misc:657 a-leaf205: Dst IP is present in SUG L3
tbl
```

```
2019-10-01 16:05:27,473 INFO ftrriage: misc:657 a-leaf205: RwdMAC DIPo(10.0.72.67) is
```

one of dst TEPs ['10.0.72.67']
2019-10-01 16:06:25,200 INFO ftriage: main:622 Found peer-node a-spine3 and IF: Eth1/31
in candidate list
2019-10-01 16:06:30,802 INFO ftriage: node:643 a-spine3: Extracted Internal-port GPD
Info for lc: 1
2019-10-01 16:06:30,803 INFO ftriage: fcls:4414 a-spine3: LC trigger ELAM with IFS:
Eth1/31 Asic :3 Slice: 1 Srcid: 24
2019-10-01 16:07:05,717 INFO ftriage: main:839 L3 packet Seen on a-spine3 Ingress:
Eth1/31 Egress: LC-1/3 FC-24/0 Port-1 Vnid: 2392068
2019-10-01 16:07:05,718 INFO ftriage: pktrec:490 a-spine3: Collecting transient losses
snapshot for LC module: 1
2019-10-01 16:07:28,043 INFO ftriage: fib:332 a-spine3: Transit in spine
2019-10-01 16:07:35,902 INFO ftriage: unicast:1252 a-spine3: Enter dbg_sub_nextthop with
Transit inst: ig infra: False glbs.dipo: 10.0.72.67
2019-10-01 16:07:36,018 INFO ftriage: unicast:1417 a-spine3: EP is known in COOP (DIPo =
10.0.72.67)
2019-10-01 16:07:40,422 INFO ftriage: unicast:1458 a-spine3: Infra route 10.0.72.67 present
in RIB
2019-10-01 16:07:40,423 INFO ftriage: node:1331 a-spine3: Mapped LC interface: LC-1/3
FC-24/0 Port-1 to FC interface: FC-24/0 LC-1/3 Port-1
2019-10-01 16:07:46,059 INFO ftriage: node:460 a-spine3: Extracted GPD Info for fc: 24
2019-10-01 16:07:46,060 INFO ftriage: fcls:5748 a-spine3: FC trigger ELAM with IFS: FC-
24/0 LC-1/3 Port-1 Asic :0 Slice: 1 Srcid: 40
2019-10-01 16:08:06,735 INFO ftriage: unicast:1774 L3 packet Seen on FC of node: a-spine3
with Ingress: FC-24/0 LC-1/3 Port-1 Egress: FC-24/0 LC-1/3 Port-1 Vnid: 2392068
2019-10-01 16:08:06,735 INFO ftriage: pktrec:487 a-spine3: Collecting transient losses
snapshot for FC module: 24
2019-10-01 16:08:09,123 INFO ftriage: node:1339 a-spine3: Mapped FC interface: FC-24/0
LC-1/3 Port-1 to LC interface: LC-1/3 FC-24/0 Port-1
2019-10-01 16:08:09,124 INFO ftriage: unicast:1474 a-spine3: Capturing Spine Transit pkt-
type L3 packet on egress LC on Node: a-spine3 IFS: LC-1/3 FC-24/0 Port-1
2019-10-01 16:08:09,594 INFO ftriage: fcls:4414 a-spine3: LC trigger ELAM with IFS: LC-
1/3 FC-24/0 Port-1 Asic :3 Slice: 1 Srcid: 48
2019-10-01 16:08:44,447 INFO ftriage: unicast:1510 a-spine3: L3 packet Spine egress
Transit pkt Seen on a-spine3 Ingress: LC-1/3 FC-24/0 Port-1 Egress: Eth1/29 Vnid: 2392068
2019-10-01 16:08:44,448 INFO ftriage: pktrec:490 a-spine3: Collecting transient losses
snapshot for LC module: 1
2019-10-01 16:08:46,691 INFO ftriage: unicast:1681 a-spine3: Packet is exiting the fabric
through {a-spine3: ['Eth1/29']} Dipo 10.0.72.67 and filter SIP 10.0.2.100 DIP 10.0.1.100
2019-10-01 16:10:19,947 INFO ftriage: main:716 Capturing L3 packet Fex: False on node:
a-spine1 IF: Eth2/25
2019-10-01 16:10:25,752 INFO ftriage: node:643 a-spine1: Extracted Internal-port GPD
Info for lc: 2
2019-10-01 16:10:25,754 INFO ftriage: fcls:4414 a-spine1: LC trigger ELAM with IFS:
Eth2/25 Asic :3 Slice: 0 Srcid: 24
2019-10-01 16:10:51,164 INFO ftriage: main:716 Capturing L3 packet Fex: False on node:
a-spine2 IF: Eth1/31
2019-10-01 16:11:09,690 INFO ftriage: main:839 L3 packet Seen on a-spine2 Ingress:
Eth1/31 Egress: Eth1/25 Vnid: 2392068
2019-10-01 16:11:09,690 INFO ftriage: pktrec:490 a-spine2: Collecting transient losses
snapshot for LC module: 1
2019-10-01 16:11:24,882 INFO ftriage: fib:332 a-spine2: Transit in spine
2019-10-01 16:11:32,598 INFO ftriage: unicast:1252 a-spine2: Enter dbg_sub_nextthop with
Transit inst: ig infra: False glbs.dipo: 10.0.72.67
2019-10-01 16:11:32,714 INFO ftriage: unicast:1417 a-spine2: EP is known in COOP (DIPo =
10.0.72.67)
2019-10-01 16:11:36,901 INFO ftriage: unicast:1458 a-spine2: Infra route 10.0.72.67 present
in RIB
2019-10-01 16:11:47,106 INFO ftriage: main:622 Found peer-node a-leaf101 and IF:
Eth1/54 in candidate list
2019-10-01 16:12:09,836 INFO ftriage: main:839 L3 packet Seen on a-leaf101 Ingress:
Eth1/54 Egress: Eth1/30 (Po5) Vnid: 11470
2019-10-01 16:12:09,952 INFO ftriage: pktrec:490 a-leaf101: Collecting transient losses
snapshot for LC module: 1

```

2019-10-01 16:12:30,991 INFO      ftriage:      nxos:1404 a-leaf101: nxos matching rule id:4659
scope:84 filter:65534
2019-10-01 16:12:32,327 INFO      ftriage:      main:522 Computed egress encap string vlan-1075
2019-10-01 16:12:32,333 INFO      ftriage:      main:313 Building egress BD(s), Ctx
2019-10-01 16:12:34,559 INFO      ftriage:      main:331 Egress Ctx Prod:Vrfl
2019-10-01 16:12:34,560 INFO      ftriage:      main:332 Egress BD(s): Prod:Bd1
2019-10-01 16:12:37,704 INFO      ftriage:      unicast:1252 a-leaf101: Enter dbg_sub_nextthop with
Local inst: eg infra: False glbs.dipo: 10.0.72.67
2019-10-01 16:12:37,705 INFO      ftriage:      unicast:1257 a-leaf101: dbg_sub_nextthop invokes
dbg_sub_eg for ptep
2019-10-01 16:12:37,705 INFO      ftriage:      unicast:1784 a-leaf101: <- is egress node
2019-10-01 16:12:37,911 INFO      ftriage:      unicast:1833 a-leaf101: Dst EP is local
2019-10-01 16:12:37,912 INFO      ftriage:      misc:657 a-leaf101: EP if(Po5) same as egr
if(Po5)
2019-10-01 16:12:38,172 INFO      ftriage:      misc:657 a-leaf101: Dst IP is present in SUG L3
tbl
2019-10-01 16:12:38,564 INFO      ftriage:      misc:657 a-leaf101: RW seg_id:11470 in SUG same
as EP segid:11470
fTriage Status: {"dbgFtriage": {"attributes": {"operState": "Idle", "pid": "0", "apicId": "0",
"id": "0"}}}
fTriage Status: {"dbgFtriage": {"attributes": {"operState": "Idle", "pid": "0", "apicId": "0",
"id": "0"}}}

```

fTriage中有大量資料。其中重點介紹一些最重要的領域。請注意，資料包的路徑為「leaf205(Pod 2)> spine3(Pod 2)> spine2(Pod 1)> leaf101(Pod 1)」。

沿途作出的所有轉發決策和合約查詢也都顯示。

請注意，如果這是第2層流，則需要將fTriage的語法設定為：

```
ftriage bridge -ii LEAF:205 -dmac 00:00:11:11:22:22
```

EP不在COOP中的代理請求

在考慮特定故障場景之前，還需要討論一個與通過多Pod的單播轉發相關的內容。如果目標端點未知、請求被代理且端點不在COOP中，會發生什麼情況？

在此案例中，封包/訊框會傳送到主幹，並產生收集請求。

當主幹產生收集請求時，原始資料包仍保留在請求中，但資料包會收到ethertype 0xffff2，這是保留給gleans的自定義Ethertype。因此，在Wireshark等資料包捕獲工具中解釋這些消息並不容易。

第3層外部目的地也設定為239.255.255.240，這是專門用於收集消息的保留組播組。這些流量應在交換矩陣中泛洪，任何已部署收集請求的目標子網的出口枝葉交換機都將生成ARP請求以解析目標。這些ARP從配置的BD子網IP地址傳送（因此，如果在橋接域上禁用了單播路由，則代理請求無法解析無提示/未知端點的位置）。

可以通過以下命令驗證在出口枝葉上接收聚合消息以及隨後生成的ARP和接收到的ARP響應：

收集ARP驗證

```

a-leaf205# show ip arp internal event-history event | grep -F -B 1 192.168.21.11
...
73) Event:E_DEBUG_DSF, length:127, at 316928 usecs after Wed May 1 08:31:53 2019
Updating epm ifidx: 1a01e000 vlan: 105 ip: 192.168.21.11, ifMode: 128 mac: 8c60.4f02.88fc <<<
Endpoint is learned
75) Event:E_DEBUG_DSF, length:152, at 316420 usecs after Wed May 1 08:31:53 2019
log_collect_arp_pkt; sip = 192.168.21.11; dip = 192.168.21.254; interface = Vlan104;info = Garp

```

```

Check adj:(nil) <<< Response received
77) Event:E_DEBUG_DSF, length:142, at 131918 usecs after Wed May 1 08:28:36 2019
log_collect_arp_pkt; dip = 192.168.21.11; interface = Vlan104;iod = 138; Info = Internal Request
Done <<< ARP request is generated by leaf
78) Event:E_DEBUG_DSF, length:136, at 131757 usecs after Wed May 1 08:28:36 2019 <<< Glean
received, Dst IP is in BD subnet
log_collect_arp_glean;dip = 192.168.21.11;interface = Vlan104;info = Received pkt Fabric-Glean:
1
79) Event:E_DEBUG_DSF, length:174, at 131748 usecs after Wed May 1 08:28:36 2019
log_collect_arp_glean; dip = 192.168.21.11; interface = Vlan104; vrf = CiscoLive2019:vrf1; info
= Address in PSVI subnet or special VIP <<< Glean Received, Dst IP is in BD subnet

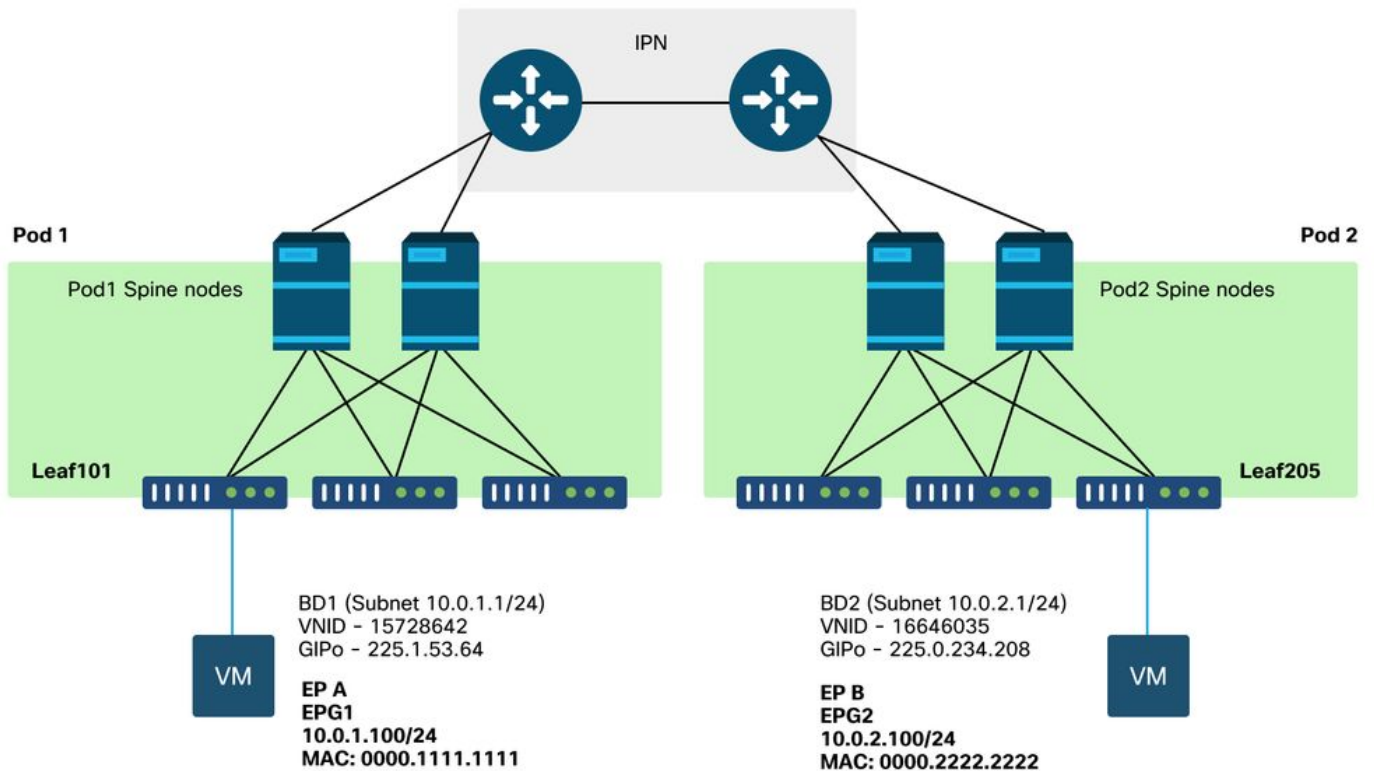
```

作為參考，傳送到239.255.255.240的彙總消息是需要在IPN上的雙向PIM組範圍內包含此組的原因。

多Pod故障排除場景#1 (單播)

在以下拓撲中，EP B無法與EP A通訊。

拓撲故障排除



請注意，在多Pod轉發中出現的許多問題與在單Pod中發現的問題相同。因此，需要重點解決多Pod的特定問題。

執行前面介紹的單播故障排除工作流程時，請注意，請求是代理的，但Pod 2中的主幹節點在COOP中沒有目標IP。

原因：COOP中缺少終結點

如前所述，系統會根據BGP EVPN資訊填充遠端Pod終端的COOP條目。因此，必須確定：

a.) 源Pod(Pod 2)主幹是否在EVPN中包含？

```
a-spine4# show bgp l2vpn evpn 10.0.1.100 vrf overlay-1
<no output>
```

b.) 遠端Pod(Pod 1)主幹是否在EVPN中包含？

```
a-spine1# show bgp l2vpn evpn 10.0.1.100 vrf overlay-1
Route Distinguisher: 1:16777199 (L2VNI 1)
BGP routing table entry for [2]:[0]:[15728642]:[48]:[0050.5681.3ee6]:[32]:[10.0.1.100]/272,
version 11751 dest ptr 0xafbf8192
Paths: (1 available, best #1)
Flags: (0x00010a 00000000) on xmit-list, is not in rib/evpn
Multipath: eBGP iBGP
```

```
Advertised path-id 1
Path type: local 0x4000008c 0x0 ref 0 adv path ref 1, path is valid, is best path
AS-Path: NONE, path locally originated
0.0.0.0 (metric 0) from 0.0.0.0 (192.168.1.101)
Origin IGP, MED not set, localpref 100, weight 32768
Received label 15728642 2392068
Extcommunity:
RT:5:16
```

Path-id 1 advertised to peers:

Pod 1主幹已安裝，下一跳IP為0.0.0.0;這意味著它從COOP本地匯出。但是請注意，「通告給對等體」部分不包括Pod 2脊柱節點。

c.) Pod之間的BGP EVPN是否啟動？

```
a-spine4# show bgp l2vpn evpn summ vrf overlay-1
```

Neighbor	V	AS	MsgRcvd	MsgSent	TblVer	InQ	OutQ	Up/Down	State/PfxRcd
192.168.1.101	4	65000	57380	66362		0	0	0 00:00:21	Active
192.168.1.102	4	65000	57568	66357		0	0	0 00:00:22	Active

請注意，在上面的輸出中，Pod之間的BGP EVPN對等已關閉。State/PfxRcd列中除數值以外的任何內容均表示鄰接關係未啟動。Pod 1的EP不是通過EVPN學習的，也不是匯入到COOP中。

如果出現此問題，請驗證以下內容：

1. 脊柱節點和連線的IPN之間是否啟用OSPF？
2. 脊柱節點是否通過OSPF獲知遠端脊柱IP的路由？
3. 通過IPN的完整路徑是否支援巨量MTU？
4. 所有協定鄰接關係是否穩定？

其他可能的原因

如果終端不在任何Pod的COOP資料庫中，並且目標裝置是靜默主機（在交換矩陣中的任何枝葉交換機上未獲知），請驗證交換矩陣收集過程是否正常工作。要使此項工作：

- 必須在BD上啟用單播路由。
- 目標必須位於BD子網中。

- IPN必須為239.255.255.240組提供組播路由服務。

組播部分將在下一節中詳細介紹。

Multi-Pod廣播、未知的單播和組播(BUM)轉發概述

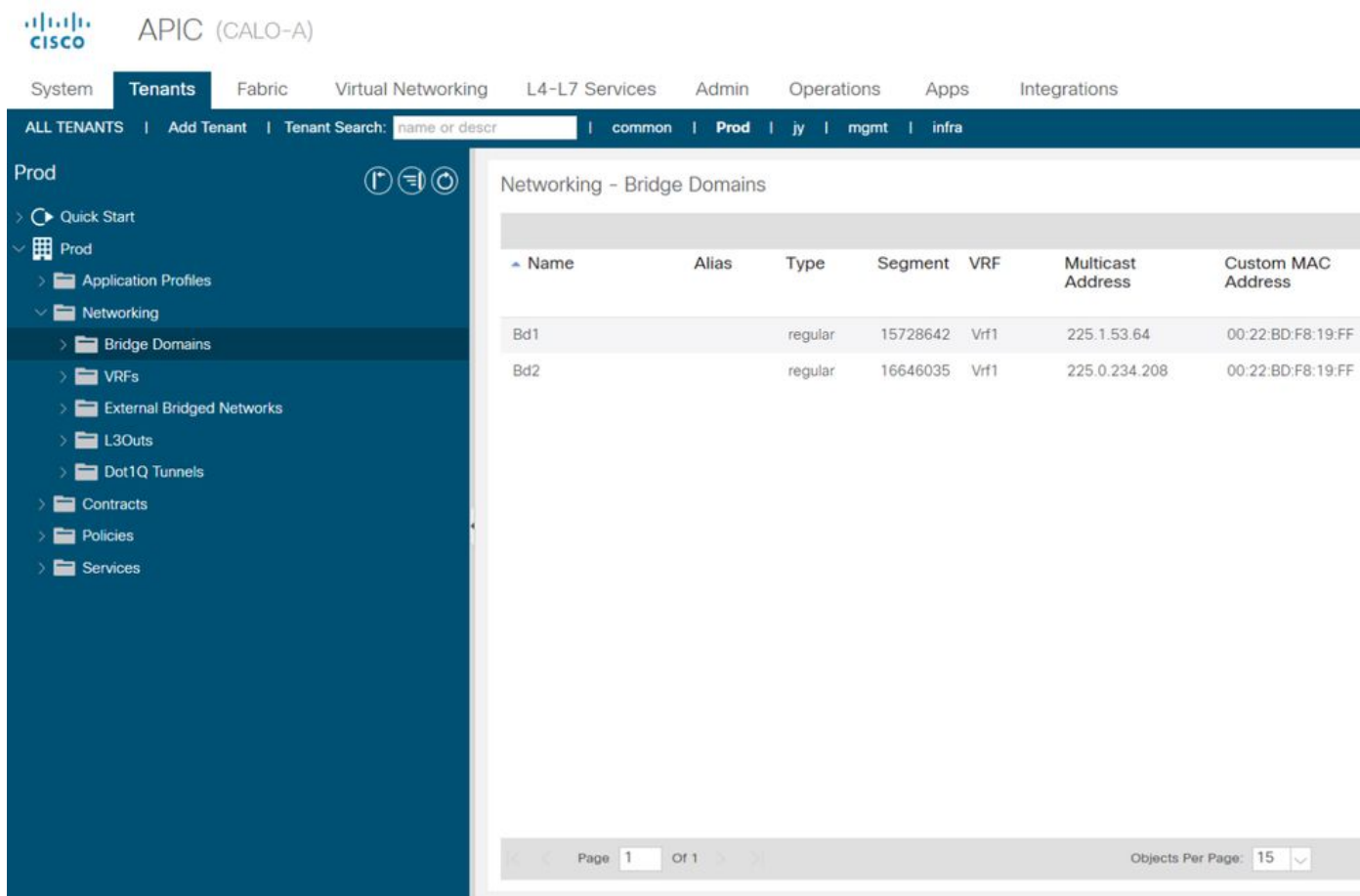
在ACI中，流量在許多不同的情況下通過重疊組播組泛洪。例如，發生泛濫：

- 組播和廣播流量。
- 必須泛洪的未知單播。
- 交換矩陣ARP收集消息。
- EP通告消息。

許多特性和功能依賴於BUM轉發。

在ACI中，所有網橋域都分配一個組播地址，稱為組IP外部（或GIPo）地址。網橋域內必須泛洪的所有流量都泛洪到此GIPo上。

GUI中的BD GIPo



The screenshot shows the Cisco APIC (CALO-A) GUI. The left sidebar is expanded to 'Prod' > 'Networking' > 'Bridge Domains'. The main content area displays a table titled 'Networking - Bridge Domains' with the following data:

Name	Alias	Type	Segment	VRF	Multicast Address	Custom MAC Address
Bd1		regular	15728642	Vrf1	225.1.53.64	00:22:BD:F8:19:FF
Bd2		regular	16646035	Vrf1	225.0.234.208	00:22:BD:F8:19:FF

At the bottom of the table, it shows 'Page 1 Of 1' and 'Objects Per Page: 15'.

可以直接在一個APIC上查詢該對象。

Moquery中的BD GIPo

```
a-apid1# moquery -c fvBD -f 'fv.BD.name=="Bd1"'  
Total Objects shown: 1
```

```

# fv.BD
name : Bd1
OptimizeWanBandwidth : no
annotation :
arpFlood : yes
bcastP : 225.1.53.64
childAction :
configIssues :
descr :
dn : uni/tn-Prod/BD-Bd1
epClear : no
epMoveDetectMode :
extMngdBy :
hostBasedRouting : no
intersiteBumTrafficAllow : no
intersiteL2Stretch : no
ipLearning : yes
ipv6McastAllow : no
lcOwn : local
limitIpLearnToSubnets : yes
llAddr : ::
mac : 00:22:BD:F8:19:FF
mcastAllow : no
modTs : 2019-09-30T20:12:01.339-04:00
monPolDn : uni/tn-common/monepg-default
mtu : inherit
multiDstPktAct : bd-flood
nameAlias :
ownerKey :
ownerTag :
pcTag : 16387
rn : BD-Bd1
scope : 2392068
seg : 15728642
status :
type : regular
uid : 16011
unicastRoute : yes
unkMacUcastAct : proxy
unkMcastAct : flood
v6unkMcastAct : flood
vmac : not-applicable

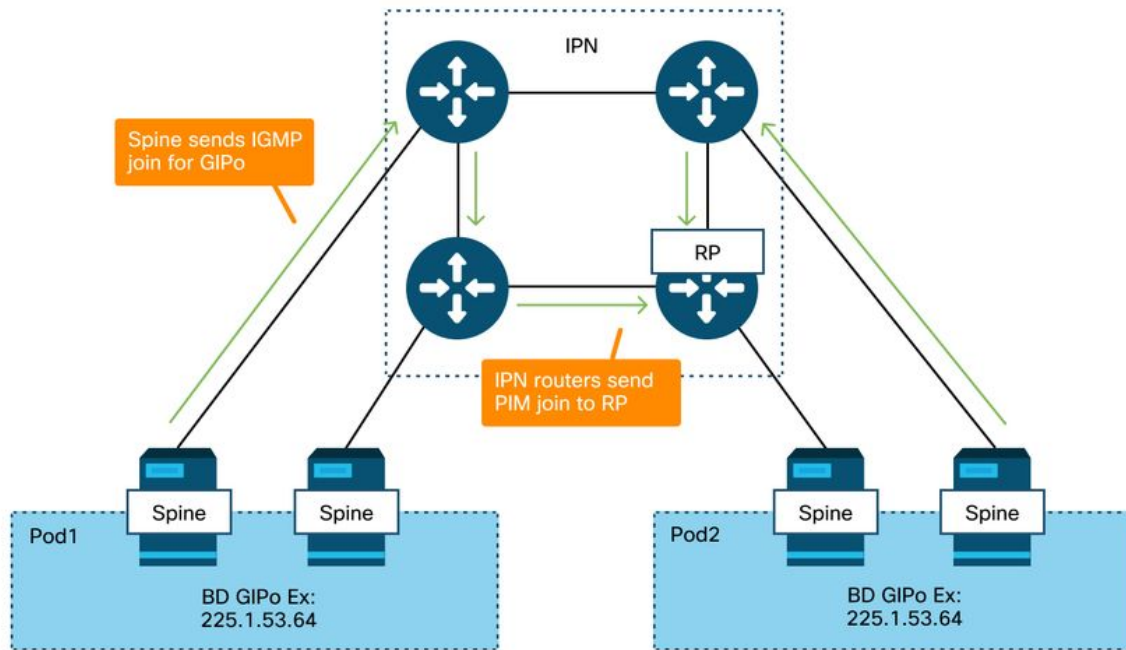
```

無論是否使用Multi-Pod，以上關於GIPo泛洪的資訊都是正確的。與多Pod相關的這一部分是IPN上的組播路由。

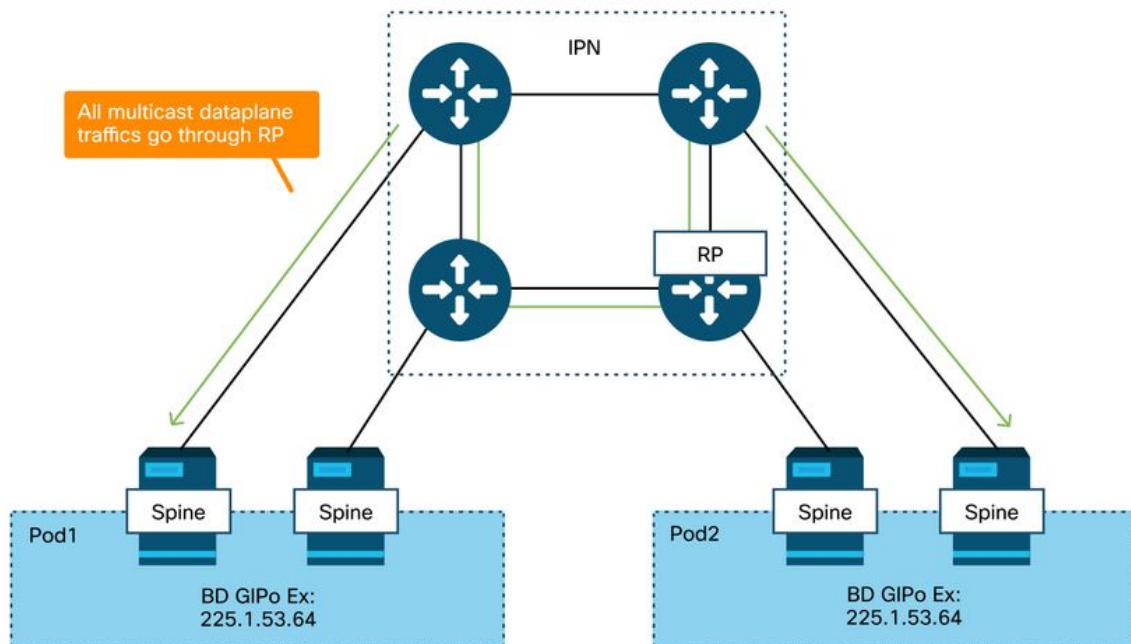
IPN多點傳送路由涉及以下內容：

- 主幹節點充當組播主機（僅限IGMP）。它們不運行PIM。
- 如果BD部署在Pod中，則該Pod的一個主幹將在其一個面向IPN的介面上傳送IGMP連線。此功能跨所有主幹節點和許多組上的面向IPN的介面進行條帶化。
- IPN接收這些連線，並向雙向PIM RP傳送PIM連線。
- 由於使用了PIM Bidir，因此沒有(S, G)樹。PIM Bidir中只使用(*,G)樹。
- 傳送到GIPo的所有資料平面流量都會通過RP。

IPN多點傳送控制平面



IPN多點傳送資料平面

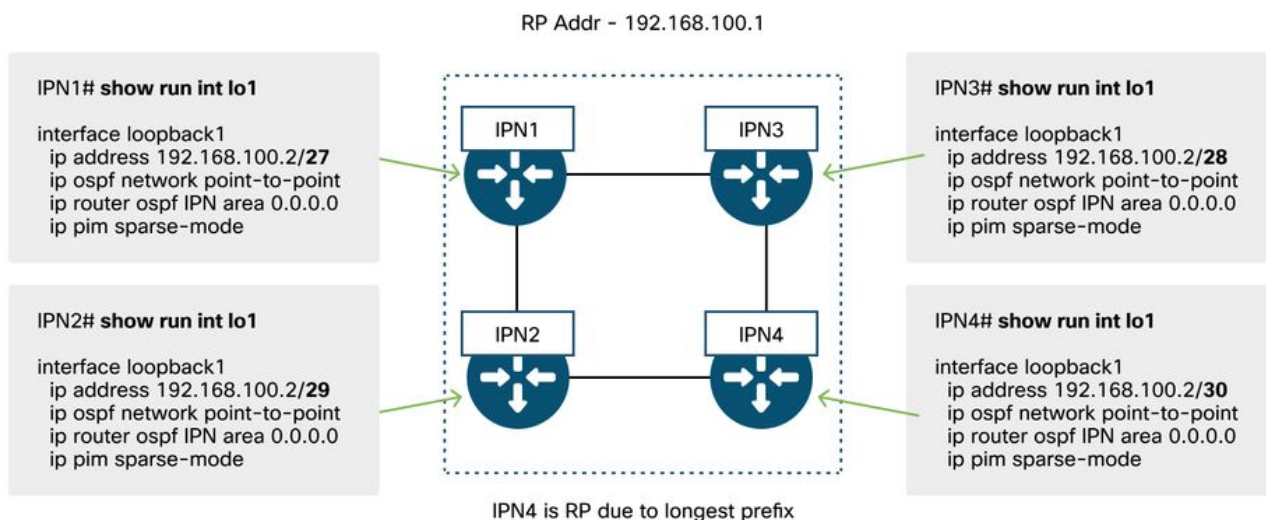


使用PIM Bidir的RP冗餘的唯一方法是使用Phantom。本書的「Multi-Pod Discovery (多容器發現)」部分對此進行了詳細介紹。快速總結一下，請注意，對於虛擬RP：

- 所有IPN必須配置相同的RP地址。

- 任何裝置上都不必須存在確切的RP地址。
- 多台裝置向包含虛擬RP IP地址的子網通告可達性。通告的子網長度應有所不同，以便所有路由器都同意誰將通告RP的最佳路徑。如果此路徑丟失，則收斂取決於IGP。

虛擬RP配置



多Pod廣播、未知的單播和組播(BUM)故障排除工作流

1. 首先確認交換矩陣是否真正將流視為多目的地。

在以下常見示例中，流將在BD中泛洪：

- 幀是ARP廣播，並且BD上啟用了ARP泛洪。
- 該幀將發往組播組。請注意，即使啟用了IGMP監聽，流量仍會總是湧入GIPo上的交換矩陣。
- 流量將發往ACI為其提供組播路由服務的組播組。
- 流是第2層（橋接流），且目標MAC地址未知，並且BD上的未知單播行為設定為「泛洪」。

確定做出轉發決策的最簡單方法是使用ELAM。

2. 確定BD GIPo。

請參閱本章前面討論此問題的部分。還可以通過ELAM助理應用運行骨幹ELAM，以驗證是否正在接收泛洪流量。

3. 驗證該GIPo的IPN上的組播路由表。

執行此操作的輸出將根據所使用的IPN平台而有所不同，但級別較高：

- 所有IPN路由器必須在RP上達成一致，此GIP的RPF必須指向此樹。
- 連線到每個Pod的一台IPN路由器應為該組獲得IGMP加入。

多Pod故障排除場景#2 (BUM流)

此案例包括涉及未在多個Pod或BUM案例（未知的單播等）中解析ARP的任何案例。

這裡有幾個可能的原因。

可能的原因1:多台路由器擁有PIM RP地址

在此情境中，輸入枝葉會泛洪流量（使用ELAM驗證），來源Pod會接收並泛洪流量，但遠端Pod無法取得。對於某些bd，泛洪有效，但對於另一些不是。

在IPN上，為GIPo運行「show ip mroute <GIPo address>」以檢視RPF樹指向多個不同的路由器。

如果是這種情況，請檢查以下內容：

- 驗證是否未在任何位置配置實際PIM RP地址。擁有該實際RP地址的任何裝置都會看到其本地/32路由。
- 驗證在幻影RP場景中，多個IPN路由器未通告RP的相同字首長度。

可能原因2:IPN路由器無法獲取RP地址的路由

與第一個可能的原因相同，這裡泛洪流量無法離開IPN。每個IPN路由器上的「show ip route <rp address>」輸出將僅顯示本地配置的字首長度，而不是顯示其他路由器正在通告的字首長度。

其結果是，即使未在任何位置配置實際RP IP地址，每台裝置仍認為自己是RP。

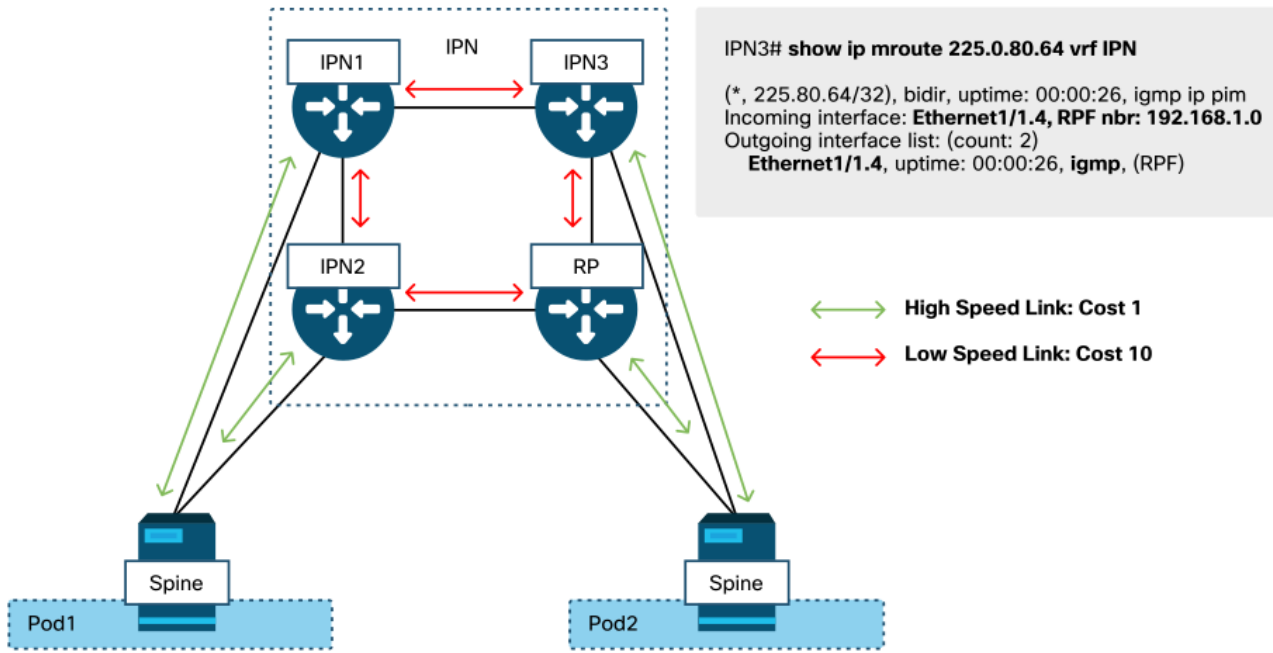
如果事實如此。檢查以下內容：

- 檢驗IPN路由器之間的路由鄰接關係是否為up。檢驗該路由是否位於實際的協定資料庫中（如OSPF資料庫）。
- 驗證所有被認為是候選RP的環回都配置為OSPF點對點網路型別。如果未配置此網路型別，則無論實際配置的內容如何，每台路由器都會始終通告/32字首長度。

可能原因3:IPN路由器未安裝GIPo路由或RPF指向ACI

如前所述，ACI不會在其面向IPN的鏈路上運行PIM。這意味著IPN指向RP的最佳路徑永遠不應指向ACI。如果多個IPN路由器連線到同一個主幹，則可能發生這種情況。與直接在IPN路由器之間連線相比，通過主幹可以發現更好的OSPF度量。

面向ACI的RPF介面



要解決此問題：

- 確保IPN路由器之間的路由協定鄰接關係已啟動。
- 將主幹節點上面向IPN的鏈路的OSPF開銷度量增加到某個值，使該度量不如IPN到IPN鏈路的優選。

其他參考資料

在ACI軟體4.0之前，外部裝置使用COS 6時遇到了一些挑戰。大多數這些問題都通過4.0增強功能得以解決，但是有關詳細資訊，請參閱CiscoLive會話「BRKACI-2934 — 對多裝置故障排除」和「服務品質」部分。

關於此翻譯

思科已使用電腦和人工技術翻譯本文件，讓全世界的使用者能夠以自己的語言理解支援內容。請注意，即使是最佳機器翻譯，也不如專業譯者翻譯的內容準確。Cisco Systems, Inc. 對這些翻譯的準確度概不負責，並建議一律查看原始英文文件（提供連結）。