ACIイントラファブリックフォワーディングのトラブルシューティング：MultiPod Forwarding

内容

概要

このドキュメントでは、ACIマルチポッド転送のシナリオを理解し、トラブルシューティングする手順について説明します。

背景説明

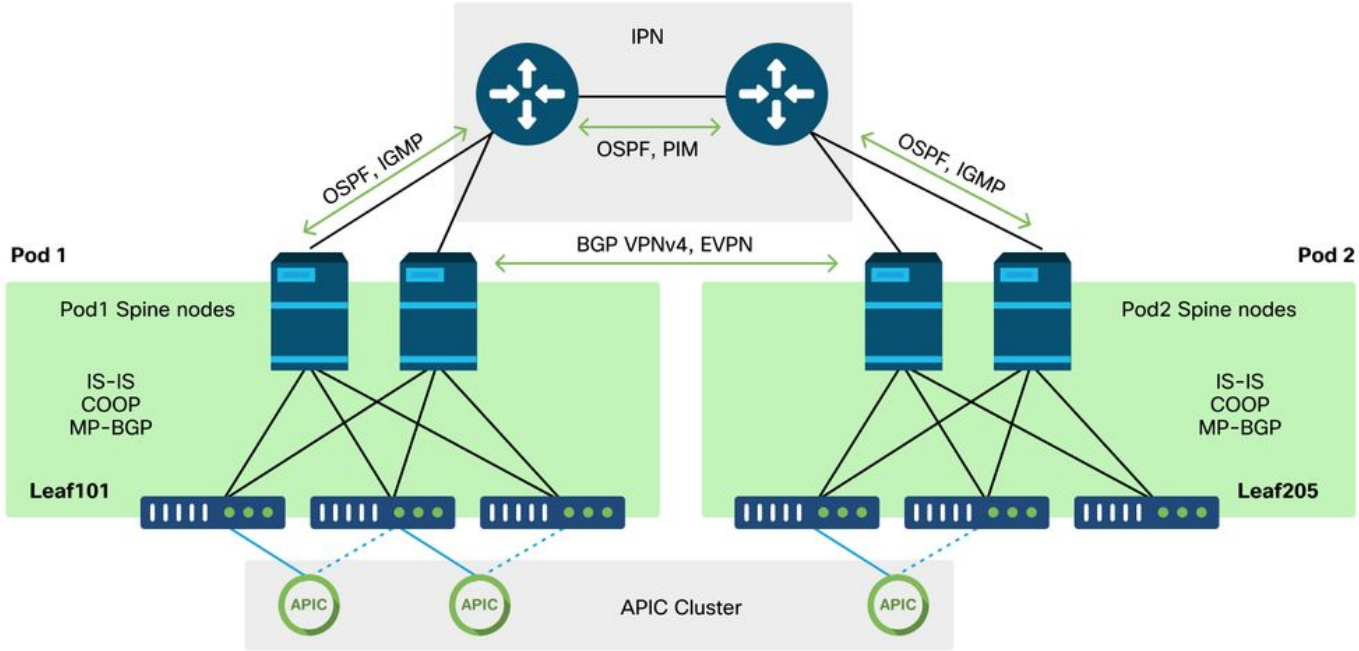
このドキュメントの内容は、 Troubleshooting Cisco Application Centric Infrastructure, Second Edition 特に **Intra-Fabric Forwarding：マルチポッド転送** 章

マルチポッドフォワーディングの概要

この章では、マルチポッド環境でポッド間の接続が正しく機能しないシナリオのトラブルシューティング方法について説明します

特定のトラブルシューティング例を見る前に、マルチポッドコンポーネントの概要を理解することが重要です。

マルチポッドコンポーネント



従来のACIファブリックと同様に、マルチポッドファブリックは引き続き単一のACIファブリックと見なされ、単一のAPICクラスタに管理を依存します。

各ポッド内で、ACIは従来のファブリックと同じプロトコルをオーバーレイで利用します。これには、TEP情報の交換、マルチキャスト発信インターフェイス(OIF)の選択、グローバルエンドポ

イントリポジトリのCOOP、ファブリックを介した外部ルータの配布のためのBGP VPNv4などのIS-ISが含まれます。

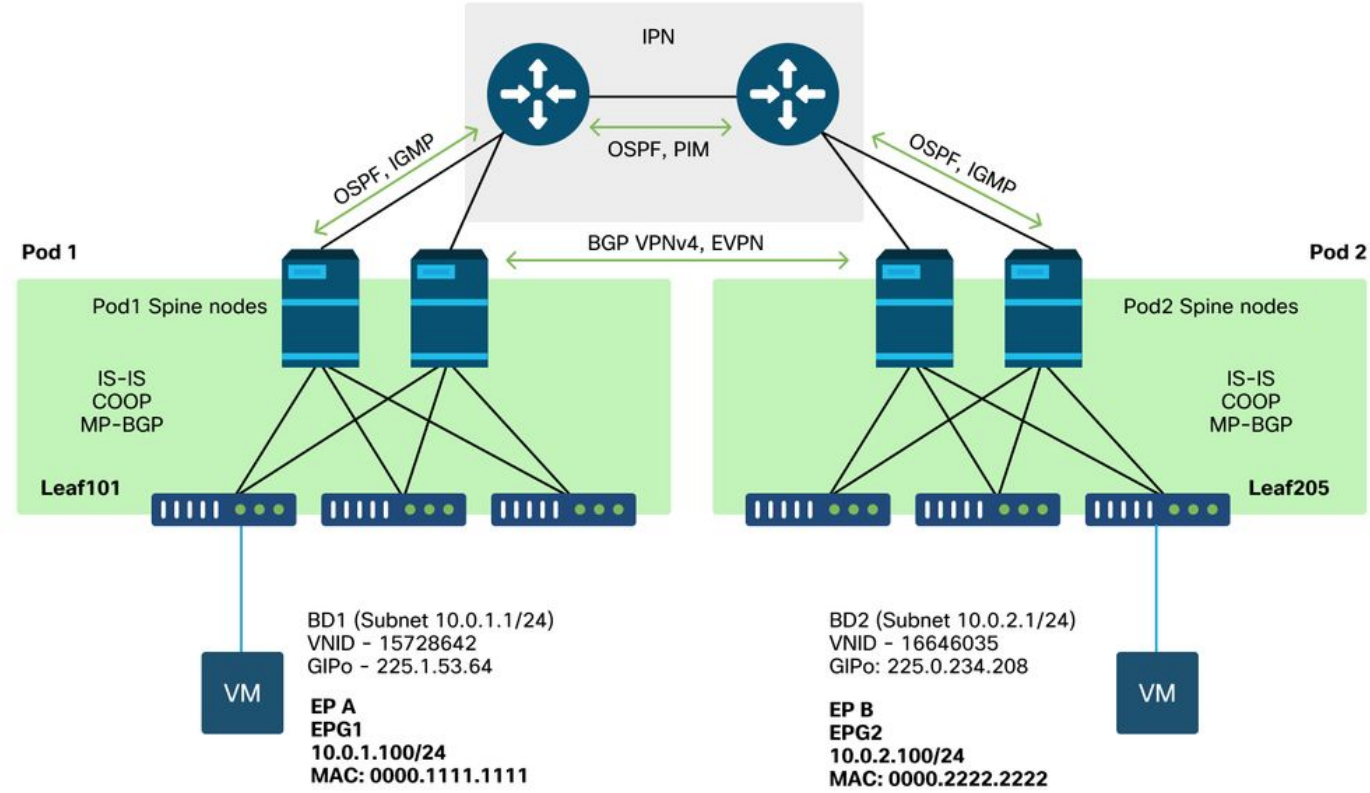
マルチポッドは、各ポッドを接続する必要があるため、これらのコンポーネント上に構築されます。

- リモートポッド内のTEPに関するルーティング情報を交換するために、OSPFはIPNを介してサマリーTEPプールをアドバタイズするために使用されます。
- あるポッドから別のポッドに学習した外部ルートを交換するために、BGP VPNv4アドレスファミリがスパインノード間で拡張されます。各ポッドは、個別のルートリフレクタクラスタになります。
- COOPに保存されているエンドポイントおよびその他の情報をポッド間で同期するために、BGP EVPNアドレスファミリがスパインノード間で拡張されます。
- 最後に、ポッド間でのブロードキャスト、不明なユニキャスト、およびマルチキャスト(BUM)トラフィックのフラッディングを処理するために、各ポッドのスパインノードはIGMPホストとして機能し、IPNルータは双方向PIMを介してマルチキャストルーティング情報を交換します。

マルチポッドのトラブルシューティングシナリオとワークフローの大部分は、シングルポッドACIファブリックに似ています。このマルチポッドのセクションでは、主にシングルポッドとマルチポッド転送の違いに焦点を当てます。

マルチポッドの例のトポロジ

すべてのシナリオのトラブルシューティングと同様に、予想される状態を理解することから始めることが重要です。この章の例では、このトポロジを参照してください。



マルチポッド転送のトラブルシューティングの一般的なワークフロー

マルチポッド転送の問題をデバッグする場合、大まかに見て、次の手順を評価できます。

1. フローはユニキャストですか、それとも複数宛先ですか。フローが動作状態のユニキャストであると想定される場合でも、ARPが解決されない場合は複数宛先フローであることを忘れないでください。
2. フローはルーティングされますか、それともブリッジされますか。従来、ACIの観点からのルーテッドフローは、宛先MACアドレスがACIで設定されたゲートウェイによって所有されるルータMACアドレスであるフローです。また、ARPフラッディングがディセーブルになっている場合、入力リーフはターゲットIPアドレスに基づいてルーティングします。宛先MACアドレスがACIによって所有されていない場合、スイッチはMACアドレスに基づいて転送するか、ブリッジドメインに設定された「不明なユニキャスト」動作に従います。
3. 入力リーフがフローをドロップしていますか。これを確認するには、TriageとELAMが最適なツールです。

フローがレイヤ3ユニキャストの場合：

1. 入力リーフには、送信元EPGと同じVRF内の宛先IPを学習するエンドポイントがありますか。その場合、学習されたルートよりも常に優先されます。リーフは、エンドポイントが学習されるトンネルアドレスまたは出力インターフェイスに直接転送します。
2. エンドポイント学習がない場合、入力リーフには「Pervasive」フラグが設定された宛先へのルートがありますか。これは、宛先サブネットがブリッジドメインサブネットとして設定されており、ネクストホップがローカルPodのスパインプロキシである必要があることを示しています。
3. Pervasiveルートがない場合、最後の手段はL3Outを通じて学習されたルートです。この部分は、シングルポッドL3Out転送と同じです。

フローがレイヤ2ユニキャストの場合：

1. 入力リーフには、送信元EPGと同じブリッジドメイン内の宛先MACアドレスを学習するエンドポイントがありますか。その場合、リーフはリモートトンネルIPに転送されるか、エンドポイントが学習されるローカルインターフェイスから転送されます。
2. 送信元ブリッジドメインに宛先MACアドレスの学習がない場合、リーフはBDが設定されている「不明なユニキャスト」動作に基づいて転送されます。[Flood]に設定すると、リーフはブリッジドメインに割り当てられたGIPoマルチキャストグループにフラッディングされます。ローカルおよびリモートのポッドには、フラッディングされたコピーが必要です。[ハードウェアプロキシ(Hardware Proxy)]に設定されている場合、フレームはプロキシルックアップのためにスパインに送信され、スパインのCOOPエントリに基づいて転送されます。

ユニキャストのトラブルシューティング出力はBUMとは大きく異なるため、ユニキャストの動作出力とシナリオはBUMの前に検討され、その後BUMに移行されます。

マルチポッドユニキャストのトラブルシューティングワークフロー

トポロジに従って、leaf205の10.0.2.100からleaf101の10.0.1.100までのフローを確認します。

ここで進む前に、送信元のARPがゲートウェイ（ルーテッドフローの場合）または宛先MACアドレス（ブリッジドフローの場合）で解決されているかどうかを確認することが重要です

1.入力リーフがパケットを受信していることを確認します。「ツール」セクションに示すELAM CLIツールと、4.2で使用可能なレポート出力を使用します。ELAM Assistantアプリケーションも使用します。

```
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
```

入力スイッチでパケットが受信されたことを確認するELAMがトリガーされていることに注意してください。次に、出力が広範囲に及ぶため、レポート内のいくつかのフィールドを確認します。

```
================================================================================================
===========
                                                        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
```

パケットの送信先に関するレポートにはさらに多くの情報がありますが、ELAMアシスタントアプリケーションは現在、このデータの解釈に役立ちます。このフローに対するELAM Assistantの出力は、この章の後半で示されます。

2.入力リーフは、入力VRFのエンドポイントとして宛先を学習していますか。そうでない場合、ルートはありますか。

```
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
    Domain                          VLAN            IP Address        IP Info
                  Info
+------------------------------------+---------------+-----------------+--------------+----------
---+-----------------------------+
```

上記のコマンドの出力は、宛先IPが学習されていないことを意味します。次に、ルーティングテーブルを確認します。

```
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
```

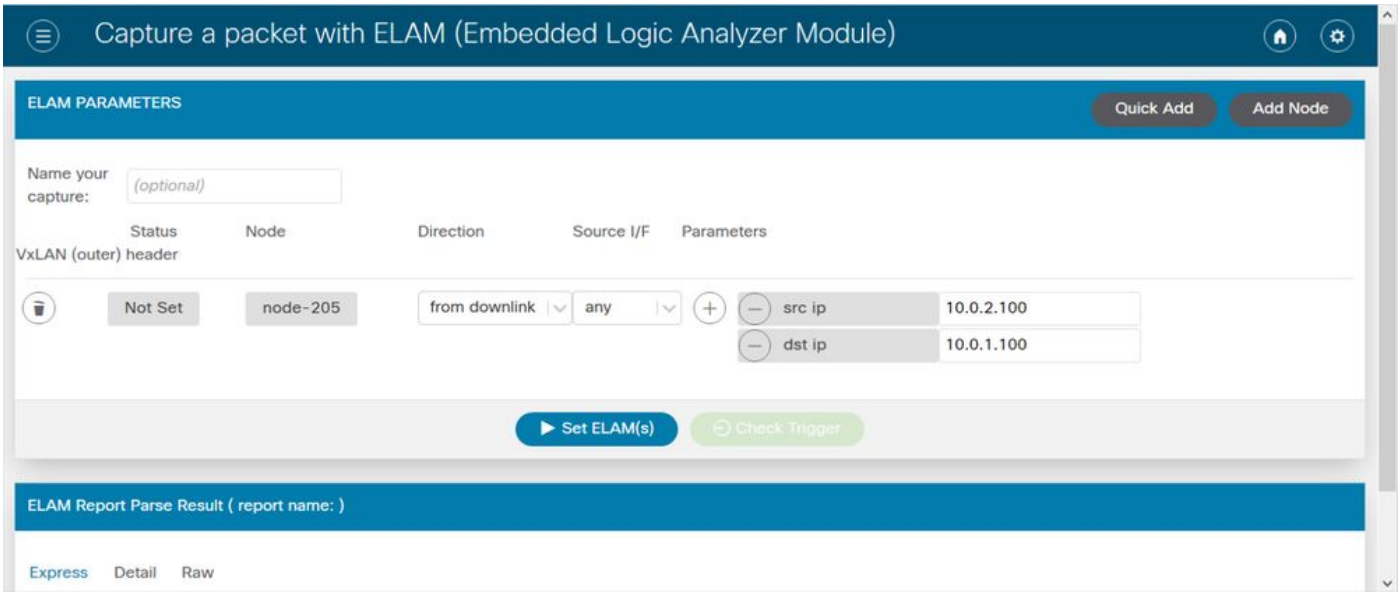
上記の出力では、これがブリッジドメインサブネットルートであることを示すPervasiveフラグが表示されています。ネクストホップは、スパイン上のエニーキャストプロキシアドレスである必要があります。

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

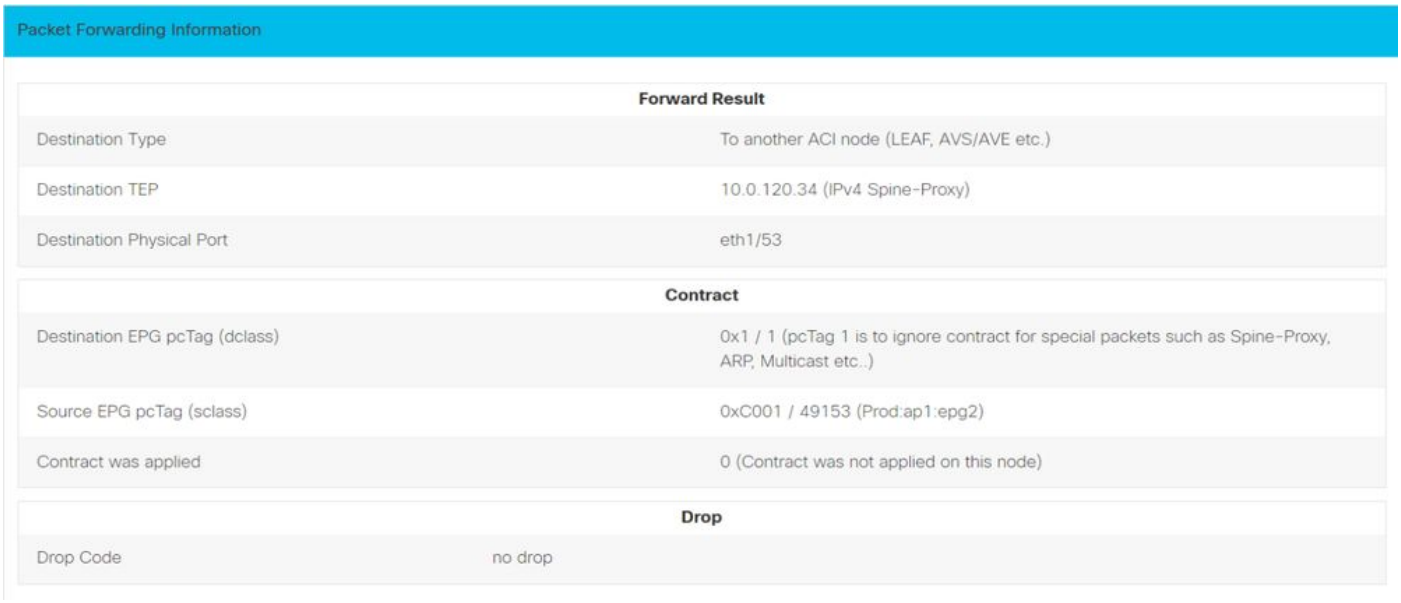
エンドポイントがトンネルまたは物理インターフェイスで学習された場合、これが優先され、パケットが直接そこで転送されます。詳細については、このマニュアルの「外部転送」の章を参照してください。

ELAM Assistantを使用して、上記の出力に示されている転送の決定を確認します。

ELAM Assistantの設定



転送決定の確認



上記の出力は、入力リーフがIPv4スパインプロキシアドレスにパケットを転送していることを示しています。これが起こると予想されていることです。

3.プロキシ要求が機能するように、スパインで宛先IPがCOOPに存在することを確認します。

スパインのCOOP出力を取得するには複数の方法があります。たとえば、「show coop internal info ip-db」コマンドを使用して表示します。

```
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

------------------------------
```

スパインで実行するその他のコマンド：

l2エントリのCOOPを照会します。

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

l3エントリのCOOPを照会し、親l2エントリを取得します。

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

l3エントリ専用のCOOPを照会します。

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

複数のmoqueryの便利な点は、APIC上で直接実行でき、coopにレコードを持つすべてのスパインを表示できることです。

4.マルチポッドスパインプロキシ転送の決定

スパインのCOOPエントリがローカルポッド内のトンネルを指している場合、転送は従来のACIの動作に基づきます。

TEPの所有者は、APICから次のコマンドを実行してファブリック内で確認できます。**moquery -c ipv4Addr -f 'ipv4.Addr.addr=="<tunnel address>"'**

プロキシシナリオでは、トンネルのネクストホップは10.0.0.34です。このIPアドレスの所有者は誰ですか。

```
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]
```

このIPは、ポッド1の両方のスパインノードによって所有されます。これは、外部プロキシアドレスと呼ばれる特定のIPです。ACIがポッド内のスパインノードによって所有されるプロキシアドレスを持っているのと同じように（このセクションのステップ2を参照）、ポッド自体に割り当てられたプロキシアドレスもあります。このインターフェイスタイプは、次のコマンドを実行して確認できます。

```
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」フラグは、これが外部プロキシTEPであることを示します。

5.スパインでのBGP EVPNの確認

coopエンドポイントレコードは、スパインのBGP EVPNからインポートする必要があります。次のコマンドを使用して、EVPN内にあることを確認できます (ただし、すでにリモートPod外部プロキシTEPのネクストホップとCOOP内にある場合は、それがEVPNから送信されたものと見なすことができます) 。

```
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
```

上記のコマンドは、MACアドレスに対しても実行できます。

-192.168.1.254は、マルチポッドセットアップ中に設定されたデータプレーンTEPです。ただし、BGPでNHとしてアドバタイズされていても、実際のネクストホップは外部プロキシTEPであることに注意してください。

-192.168.1.101および。102は、このパスをアドバタイズしているPod 1スパインノードです。

6.宛先ポッドのスパインでCOOPを確認します。

前述と同じコマンドを使用できます。

```
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

------------------------------
```

APICで次のコマンドを実行して、トンネルアドレスの所有者を確認します。

```
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
```

上記のコマンドは、COOPからのトンネルがleaf101を指していることを示しています。これは、leaf101に宛先エンドポイントのローカル学習が必要であることを意味します。

7.出力リーフにローカル学習があることを確認します。

これは、「show endpoint」コマンドで実行できます。

```
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
    Domain                          VLAN            IP Address        IP
Info                        Info
+------------------------------------+--------------+----------------+-------------+----------
---+----------------------------+
341                                     vlan-1075    0000.1111.1111 LV
po5                 Prod:ap1:epg1
Prod:Vrf1                               vlan-1075        10.0.1.100 LV
po5
```

エンドポイントが学習されることに注意してください。VLANタグ1075が設定されたポートチャネル5からパケットを転送する必要があります。

fTriageを使用したエンドツーエンドフローの確認

この章の「ツール」の項で説明したように、fTriageを使用して既存のフローをエンドツーエンドでマッピングし、パス内のすべてのスイッチがパケットに対して行っている処理を理解できます。これは、マルチポッドなど、大規模で複雑な導入で特に役立ちます。

fTriageが完全に実行されるまでに時間がかかることに注意してください（15分の可能性があります）。

サンプルフローでfTriageを実行する場合：

```
a-apic1# ftriage route -ii LEAF:205 -dip 10.0.1.100 -sip 10.0.2.100
fTriage Status: {"dbgFtriage": {"attributes": {"operState": "InProgress", "pid": "7297",
"apicId": "1", "id": "0"}}}
Starting ftriage
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/ftriage route -ii LEAF:205 -dip 10.0.1.100 -sip
10.0.2.100
2019-10-01 16:04:38,883 INFO     ftriage:     main:1165 Invoking ftriage with default password
and default username: apic#fallback\\admin
2019-10-01 16:04:54,678 INFO     ftriage:     main:839  L3 packet Seen on a-leaf205 Ingress:
Eth1/31  Egress: Eth1/53 Vnid: 2392068
2019-10-01 16:04:54,896 INFO     ftriage:     main:242  ingress encap string vlan-1021
2019-10-01 16:04:54,899 INFO     ftriage:     main:271  Building ingress BD(s), Ctx
2019-10-01 16:04:56,778 INFO     ftriage:     main:294  Ingress BD(s) Prod:Bd2
2019-10-01 16:04:56,778 INFO     ftriage:     main:301  Ingress Ctx: Prod:Vrf1
2019-10-01 16:04:56,887 INFO     ftriage:   pktrec:490  a-leaf205: Collecting transient losses
snapshot for LC module: 1
2019-10-01 16:05:22,458 INFO     ftriage:     main:933  SIP 10.0.2.100 DIP 10.0.1.100
2019-10-01 16:05:22,459 INFO     ftriage:  unicast:973  a-leaf205: <- is ingress node
2019-10-01 16:05:25,206 INFO     ftriage:  unicast:1215 a-leaf205: Dst EP is remote
2019-10-01 16:05:26,758 INFO     ftriage:     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     ftriage:     misc:659  a-leaf205: L3 packet getting
routed/bounced in SUG
2019-10-01 16:05:27,030 INFO     ftriage:     misc:657  a-leaf205: Dst IP is present in SUG L3
tbl
```

```
2019-10-01 16:05:27,473 INFO     ftriage:     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_nexthop 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_nexthop 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:Vrf1
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_nexthop 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_nexthop 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"}}}
```

Triageには大量のデータがあります。最も重要なフィールドのいくつかが強調表示されています。パケットのパスが「leaf205(Pod 2) > spine3(Pod 2) > spine2(Pod 1) > leaf101(Pod 1)」であることに注意してください。 転送に関する決定や、その途中で行われた契約検索もすべて表示されます。

これがレイヤ2フローである場合、fTriageの構文を次のように設定する必要があることに注意してください。

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

EPがCOOPでないプロキシされた要求

特定の障害シナリオを検討する前に、マルチポッド上のユニキャスト転送に関連する説明がもう1つあります。宛先エンドポイントが不明で、要求がプロキシされ、エンドポイントがCOOPでない場合はどうなりますか。

このシナリオでは、パケット/フレームがスパインに送信され、収集リクエストが生成されます。

スパインがグリーニング要求を生成すると、元のパケットは引き続き要求内に保持されますが、パケットはイーサタイプ0xfff2を受信します。これは、グリーニング用に予約されたカスタムEthertypeです。このため、Wiresharkなどのパケットキャプチャツールでこれらのメッセージを解釈することは容易ではありません。

外側のレイヤ3宛先も239.255.255.240に設定されます。これは、特に収集メッセージ用に予約されたマルチキャストグループです。これらはファブリック全体にフラッディングされる必要があり、収集リクエストの宛先サブネットが展開されている出力リーフスイッチは、宛先を解決するためのARPリクエストを生成します。これらのARPは、設定されているBDサブネットIPアドレスから送信されます（したがって、ブリッジドメインでユニキャストルーティングが無効になっている場合、プロキシ要求はサイレント/不明エンドポイントの場所を解決できません）。

出力リーフでの収集メッセージの受信と、その後に生成されたARPおよび受信されたARP応答は、次のコマンドで確認できます。

Glean ARP検証

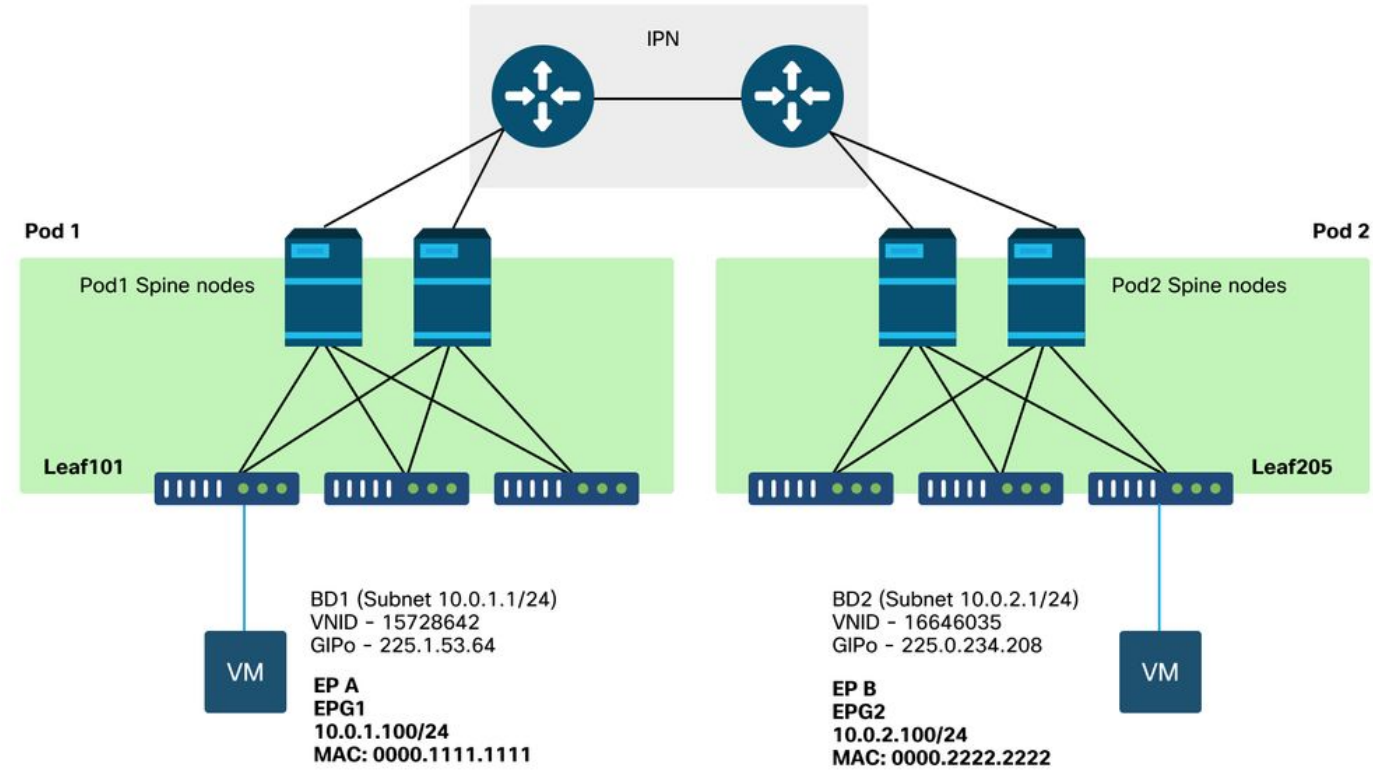
```
a-leaf205# show ip arp internal event-history event | grep -F -B 1 192.168.21.11
...
73) Event:E_DEBUG_DSF, length:127, at 316928 usecs after Wed May  1 08:31:53 2019
Updating epm ifidx: 1a01e000 vlan: 105 ip: 192.168.21.11, ifMode: 128 mac: 8c60.4f02.88fc   <<<
Endpoint is learned
75) Event:E_DEBUG_DSF,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

参考として、239.255.255.240に送信される収集メッセージは、このグループをIPNの双方向PIMグループ範囲に含める必要がある理由です。

マルチポッドトラブルシューティングシナリオ#1（ユニキャスト）

次のトポロジでは、EP BはEP Aと通信できません。

トポロジのトラブルシューティング



マルチポッド転送で見られる問題の多くは、単一ポッドで見られる問題と同じであることに注意してください。このため、マルチポッド固有の問題に焦点を当てています。

前述したユニキャストのトラブルシューティングワークフローに従う際には、要求はプロキシされますが、ポッド2のスパインノードにはCOOPの宛先IPがないことに注意してください。

原因：COOPにエンドポイントがない

前述したように、リモートポッドエンドポイントのCOOPエントリは、BGP EVPN情報から入力されます。その結果、次のことを判断することが重要です。

a.) ソースポッド (ポッド2) スパインはEVPNに含まれていますか。

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

b.) リモートポッド (ポッド1) スパインはEVPNに含まれていますか。

```
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:
```

ポッド1スパインにはそれが設定されており、ネクストホップIPは0.0.0.0です。これは、COOPからローカルにエクスポートされたことを意味します。ただし、「ピアへのアドバタイズ」セクションにはPod 2スパインノードは含まれていません。

c.) BGP EVPNはポッド間でアップになっていますか。

```
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
```

上記の出力で、BGP EVPNピアリングがポッド間でダウンしていることに注目してください。State/PfxRcdカラムの数値以外の値は、隣接関係がアップしていないことを示します。Pod 1 EPはEVPNを通じて学習されず、COOPにインポートされません。

この問題が発生する場合は、次の点を確認します。

1. スパインノードと接続されたIPNの間でOSPFはアップになっていますか。

2. スパインノードには、リモートスパインIPに対してOSPFを通じて学習されたルートがありますか。
3. IPN上のフルパスはジャンボMTUをサポートしますか。
4. すべてのプロトコル隣接関係は安定していますか。

考えられる他の原因

エンドポイントがどのポッドのCOOPデータベースにも存在せず、宛先デバイスがサイレントホスト（ファブリック内のどのリーフスイッチでも学習されていない）である場合は、ファブリック収集プロセスが正常に機能していることを確認します。これを実行するには、次のようにします。

- BDでユニキャストルーティングを有効にする必要があります。
- 宛先はBDサブネットにある必要があります。
- IPNは239.255.255.240グループに対してマルチキャストルーティングサービスを提供している必要があります。

マルチキャスト部分については、次のセクションで詳しく説明します。

マルチポッドブロードキャスト、不明なユニキャスト、およびマルチキャスト(BUM)転送の概要

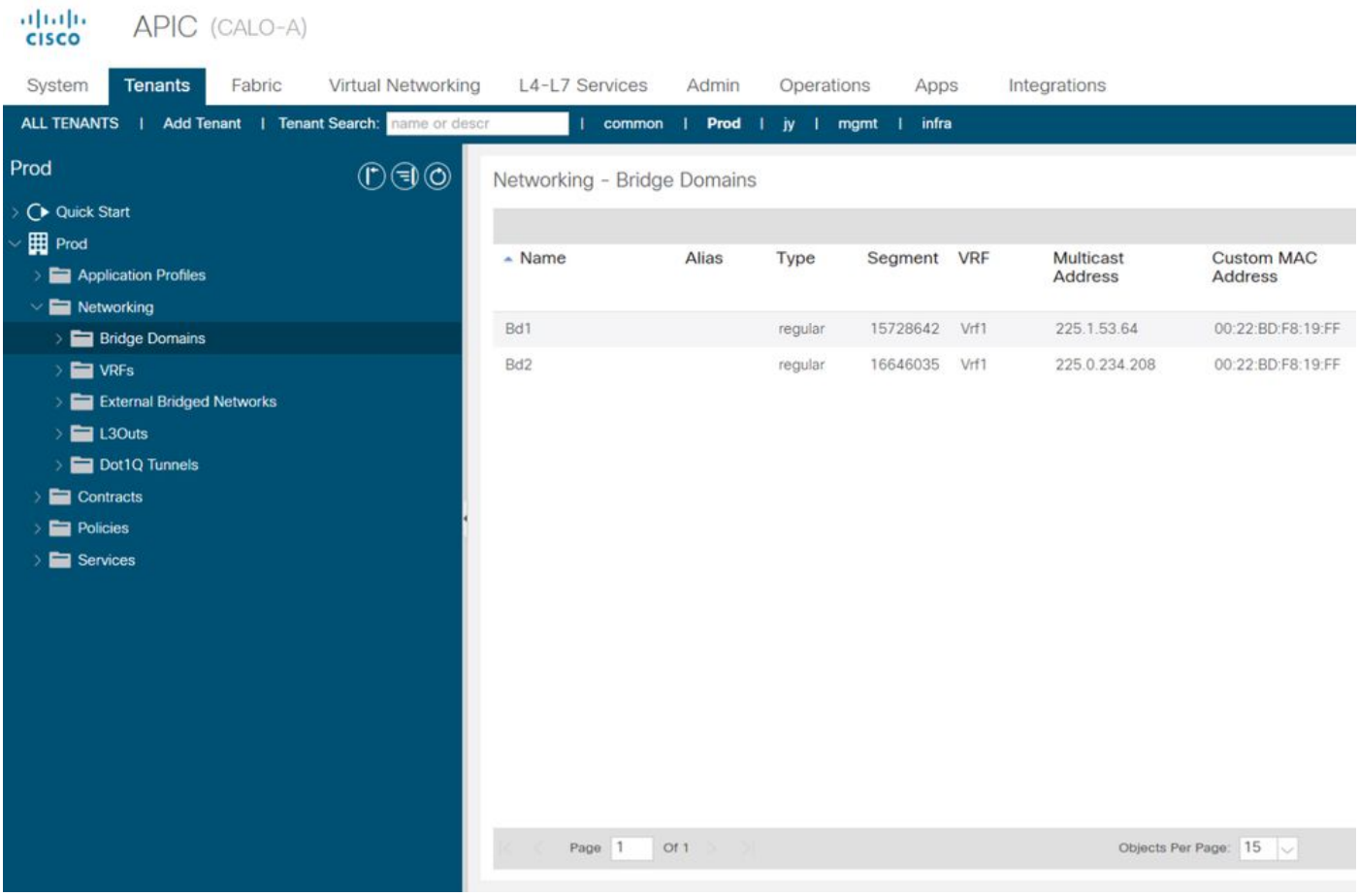
ACIでは、さまざまなシナリオでオーバーレイマルチキャストグループを介してトラフィックがフラッディングされます。たとえば、次の場合にフラッディングが発生します。

- マルチキャストおよびブロードキャストトラフィック。
- フラッディングが必要な不明なユニキャスト。
- ファブリックARP収集メッセージ。
- EPはメッセージをアナウンスします。

多くの機能がBUM転送に依存しています。

ACI内では、すべてのブリッジドメインにグループIP外部(GIPo)アドレスと呼ばれるマルチキャストアドレスが割り当てられます。ブリッジドメイン内でフラッディングする必要があるすべてのトラフィックは、このGIPoでフラッディングされます。

GUIでのBD GIPo



オブジェクトは、いずれかのAPICで直接クエリーできます。

MoqueryのBD GIPo

```
a-apic1# 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
```

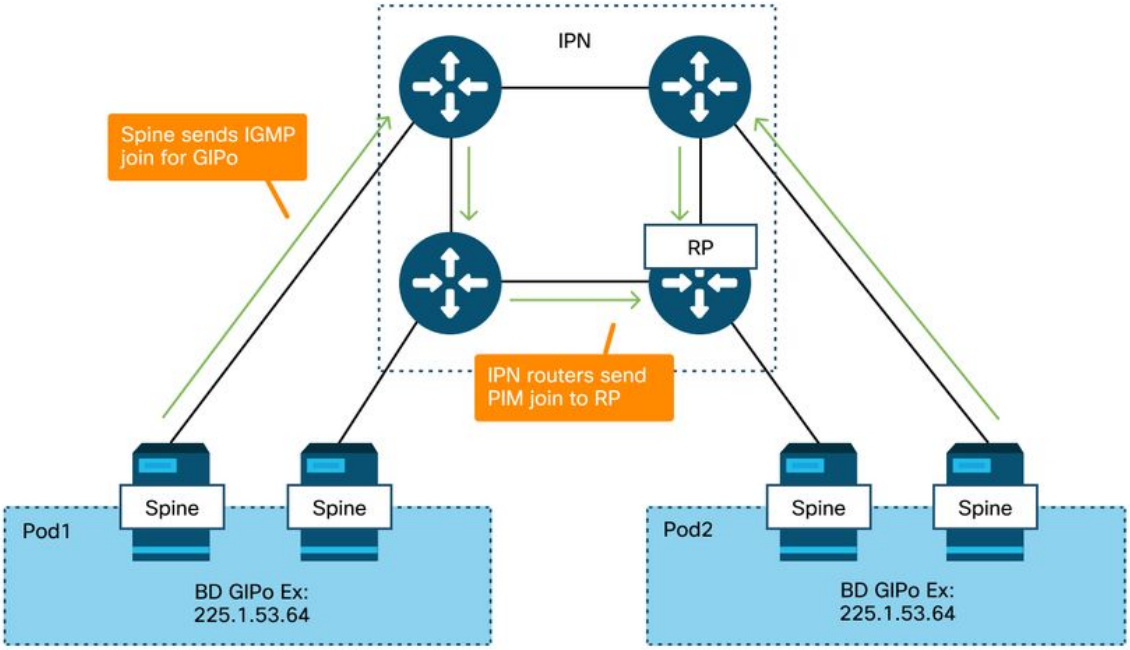
```
mtu                     : inherit
multiDstPktAct          : bd-flood
nameAlias               :
ownerKey                :
ownerTag                :
pcTag                   : 16387
rn                      : BD-Bd1
scope                   : 2392068
seg                     : 15728642
status                  :
type                    : regular
uid                     : 16011
unicastRoute            : yes
unkMacUcastAct          : proxy
unkMcastAct             : flood
v6unkMcastAct           : flood
vmac                    : not-applicable
```

GIPoフラッディングに関する上記の情報は、マルチポッドが使用されているかどうかにかかわらず当てはまります。このマルチポッドに関連するその他の部分は、IPNでのマルチキャストルーティングです。

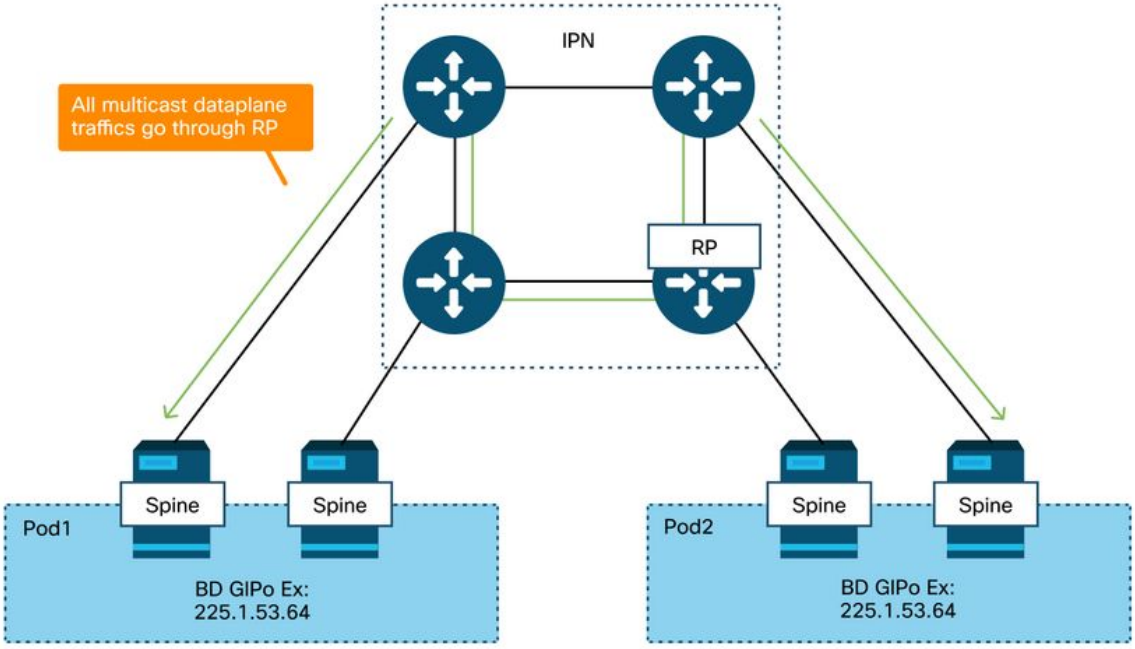
IPNマルチキャストルーティングには次の機能があります。

- スパインノードはマルチキャストホストとして機能します（IGMPのみ）。 PIMは実行されません。
- BDがポッドに展開されている場合、そのポッドから1つのスパインがIPNに面したインターフェイスの1つでIGMP参加を送信します。この機能は、すべてのスパインノードとIPNに面したインターフェイスを多数のグループに対してストライプします。
- IPNはこれらのJoinを受信し、双方向PIM RPに向けてPIM Joinを送信します。
- PIM Bidirが使用されているため、(S,G)ツリーはありません。PIM Bidirでは(*,G)ツリーのみが使用されます。
- GIPoに送信されるすべてのデータプレーントラフィックはRPを通過します。

IPNマルチキャストコントロールプレーン



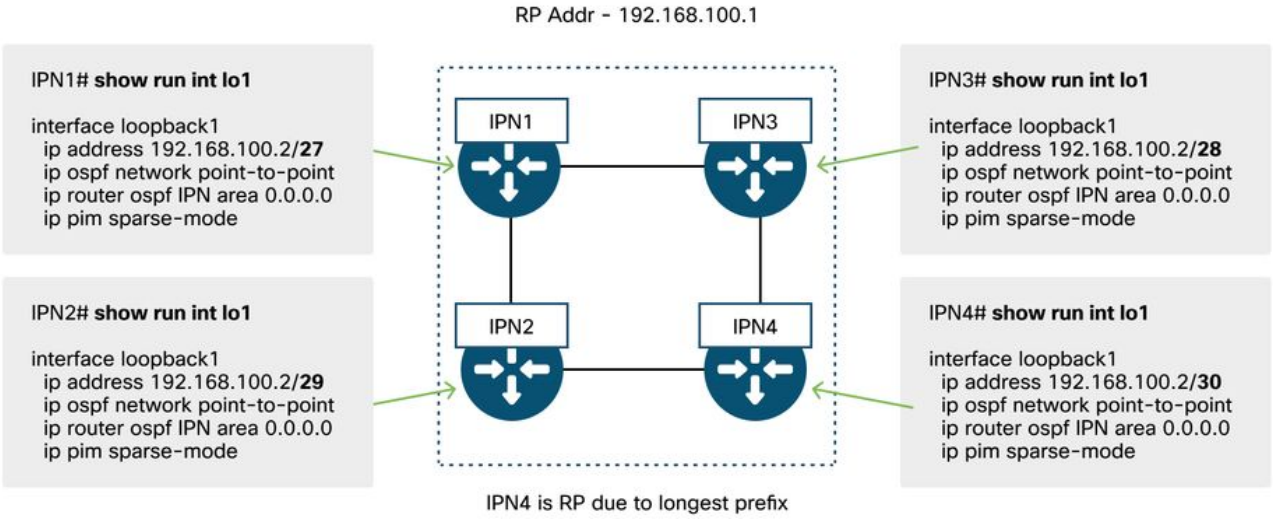
IPNマルチキャストデータプレーン



PIM BidirによるRP冗長性の唯一の手段は、Phantomを使用することです。これについては、本書の「マルチポッドディスカバリ」の部分で詳しく説明します。要約すると、ファントムRPでは次の点に注意してください。

- すべてのIPNは同じRPアドレスで設定する必要があります。
- 正確なRPアドレスはどのデバイスにも存在してはなりません。
- 複数のデバイスが、ファントムRP IPアドレスを含むサブネットへの到達可能性をアドバタイズします。アドバタイズされたサブネットはサブネット長が異なるため、すべてのルータがRPに対してベストパスをアドバタイズしているルータを決定します。このパスが失われると、コンバージェンスはIGPに依存します。

ファントムRPの設定



マルチポッドブロードキャスト、不明なユニキャスト、およびマルチキャスト(BUM)のトラブルシューティングワークフロー

1.最初に、フローがファブリックによって本当にマルチデスティネーションとして処理されているかどうかを確認します。

次の一般的な例では、BDにフローがフラッディングされます。

- フレームはARPブロードキャストであり、BDでARPフラッディングが有効になっています。
- フレームはマルチキャストグループ宛てです。IGMPスヌーピングが有効になっていても、トラフィックは常にGIPoのファブリックにフラッディングされることに注意してください。
- トラフィックの宛先は、ACIがマルチキャストルーティングサービスを提供しているマルチキャストグループです。
- フローはレイヤ2（ブリッジドフロー）であり、宛先MACアドレスは不明で、BD上の不明なユニキャスト動作は「フラッド」に設定されます。

どの転送が決定されるかを決定する最も簡単な方法は、ELAMを使用することです。

2. BD GIPoを特定します。

これについては、この章の前半のセクションを参照してください。スパインELAMは、ELAMアシスタントアプリケーションを介して実行し、フラッディングされたトラフィックが受信されていることを確認することもできます。

3.そのGIPoのIPNのマルチキャストルーティングテーブルを確認します。

これを実行するための出力は、使用しているIPNプラットフォームによって異なりますが、高いレベルでは次のようになります。

- すべてのIPNルータはRPで合意する必要があり、このGIPoのRPFはこのツリーを指す必要があります。
- 各ポッドに接続された1台のIPNルータが、グループのIGMP加入を取得する必要があります。

マルチポッドトラブルシューティングシナリオ#2（BUMフロー）

このシナリオでは、マルチポッドまたはBUMシナリオ（不明なユニキャストなど）でARPが解決されないシナリオを扱います。

ここでは、いくつかの一般的な原因が考えられます。

考えられる原因 1：複数のルータがPIM RPアドレスを所有している

このシナリオでは、入力リーフがトラフィックをフラッディングし（ELAMで確認）、送信元ポッドがトラフィックを受信してフラッディングしますが、リモートのポッドはトラフィックを取得しません。一部のBDではフラッディングが機能しますが、他のBDでは機能しません。

IPNで、GIPoに対して「show ip mroute <GIPo address>」を実行し、RPFツリーが複数の異なるルータを指していることを確認します。

その場合は、次の点を確認してください。

- 実際のPIM RPアドレスが設定されていないかどうかを確認します。実際のRPアドレスを所有するデバイスは、そのRPアドレスに対するローカル/32ルートを認識します。
- ファントムRPシナリオでは、複数のIPNルータがRPに同じプレフィクス長をアドバタイズしていないことを確認します。

考えられる原因 2：IPNルータがRPアドレスのルートを学習していない

最初の考えられる原因と同じように、ここでフラッディングされたトラフィックはIPNから出ることができません。各IPNルータでの「show ip route <rp address>」の出力には、他のルータがアドバタイズしているプレフィクス長ではなく、ローカルに設定されたプレフィクス長のみが表示されます。

この結果、実際のRP IPアドレスがどこにも設定されていなくても、各デバイスが自身をRPと見なします。

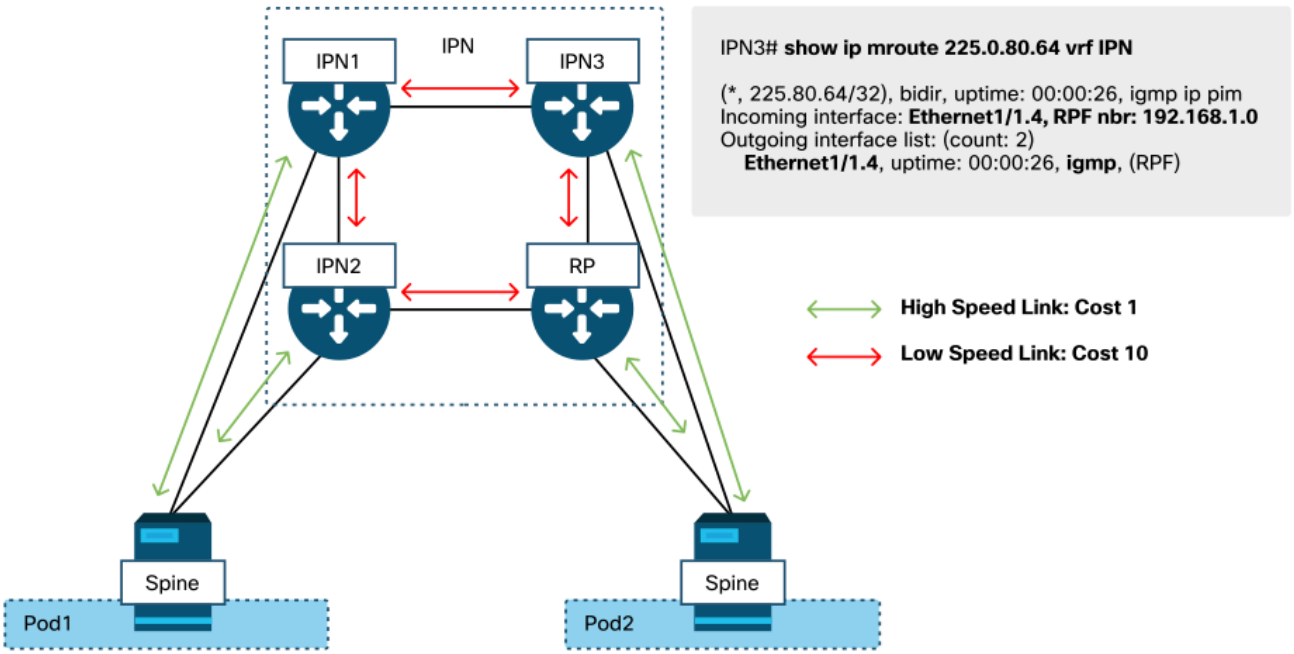
この場合、次の項目をチェックします。

- IPNルータ間のルーティング隣接関係がアップしていることを確認します。ルートが実際のプロトコルデータベース（OSPFデータベースなど）にあることを確認します。
- RP候補となるすべてのループバックがOSPFポイントツーポイントネットワークタイプとして設定されていることを確認します。このネットワークタイプが設定されていない場合、各ルータは実際の設定に関係なく、常に/32プレフィクス長をアドバタイズします。

考えられる原因 3：IPNルータがGIPoルートまたはRPFポイントをACIにインストールしていない

前述したように、ACIはIPNに面したリンクではPIMを実行しません。これは、RPに向かうIPNのベストパスがACIを指してはならないことを意味します。この問題が発生する可能性があるのは、複数のIPNルータが同じスパインに接続されており、IPNルータ間で直接接続するよりもスパインを通じてOSPFメトリックが向上する場合です。

ACIへのRPFインターフェイス



この問題を解決するには：

- IPNルータ間のルーティングプロトコルの隣接関係がアップしていることを確認します。
- スパインノード上のIPN側のリンクのOSPFコストメトリックを、そのメトリックがIPN間リンクよりも望ましくない値に増やします。

その他の参考資料

ACIソフトウェア4.0より前は、外部デバイスによるCOS 6の使用に関する課題がありました。これらの問題のほとんどは4.0の機能拡張によって解決されていますが、詳細については、CiscoLiveセッション「BRKACI-2934 – マルチポッドのトラブルシューティング」および「サービス品質」のセクションを参照してください。

翻訳について
シスコは世界中のユーザにそれぞれの言語でサポート コンテンツを提供するために、機械と人による翻訳を組み合わせて、本ドキュメントを翻訳しています。ただし、最高度の機械翻訳であっても、専門家による翻訳のような正確性は確保されません。シスコは、これら翻訳の正確性について法的責任を負いません。原典である英語版（リンクからアクセス可能）もあわせて参照することを推奨します。