Dépannage du transfert intra-fabric ACI - Transfert multipod

Contenu

Introduction

Ce document décrit les étapes à suivre pour comprendre et dépanner un scénario de transfert multipod ACI.

Informations générales

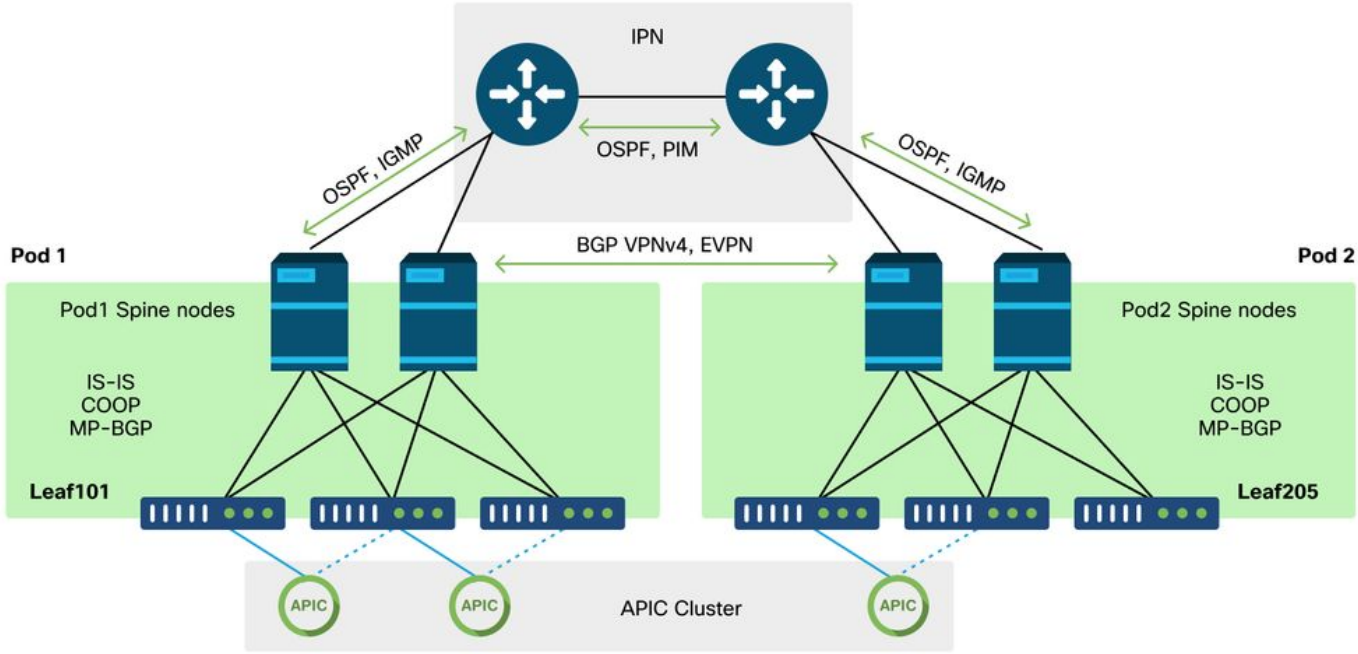
Le matériel de ce document a été extrait de la Dépannage de l'infrastructure axée sur les applications Cisco, deuxième édition , en particulier le **Transfert intra-fabric - Transfert multipod** chapitre.

Présentation du transfert multipod

Ce chapitre explique comment résoudre les scénarios dans lesquels la connectivité ne fonctionne pas correctement entre les pods dans un environnement à pods multiples

Avant d'examiner des exemples de dépannage spécifiques, il est important de prendre quelques instants pour comprendre les composants Multi-Pod à un niveau élevé.

Composants Multi-Pod



Comme un fabric ACI traditionnel, un fabric multipod est toujours considéré comme un fabric ACI unique et repose sur un cluster APIC unique pour la gestion.

Dans chaque pod individuel, l'ACI utilise les mêmes protocoles dans la superposition qu'un fabric

traditionnel. Cela inclut IS-IS pour l'échange d'informations TEP ainsi que la sélection de l'interface de multidiffusion sortante (OIF), COOP pour un référentiel de terminaux global et BGP VPNv4 pour la distribution de routeurs externes via le fabric.

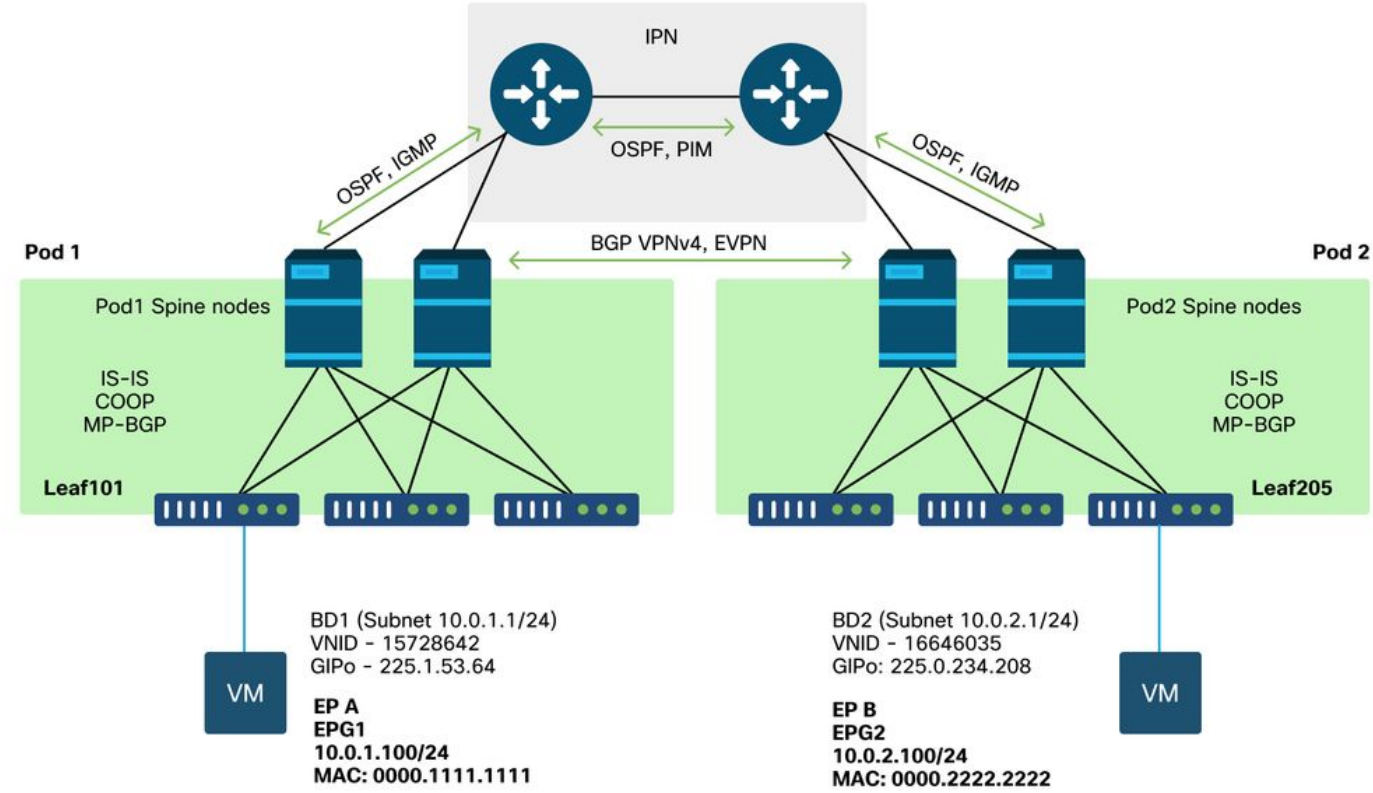
Multi-Pod s'appuie sur ces composants car il doit connecter chaque Pod ensemble.

- Pour échanger des informations de routage concernant les TEP dans le Pod distant, OSPF est utilisé pour annoncer le pool TEP récapitulatif via le réseau IP.
- Pour échanger des routes externes apprises d'un Pod à un autre, la famille d'adresses BGP VPNv4 est étendue entre les noeuds spine. Chaque Pod devient une grappe de réflecteurs de route distincte.
- Pour synchroniser les terminaux ainsi que d'autres informations stockées dans COOP sur les pods, la famille d'adresses EVPN BGP est étendue entre les noeuds spine.
- Enfin, afin de gérer l'inondation du trafic de diffusion, de monodiffusion inconnue et de multidiffusion (BUM) sur les pods, les noeuds spine de chaque pod agissent comme des hôtes IGMP et les routeurs IPN échangent des informations de routage de multidiffusion par le biais du PIM bidirectionnel.

Une grande partie des scénarios et des workflows de dépannage multipod sont similaires aux fabrics ACI à pod unique. Cette section Multi-Pod se concentrera principalement sur les différences entre le transfert de pod unique et de pod multiple.

Exemples de topologie pour plusieurs pods

Comme pour tout scénario de dépannage, il est important de commencer par comprendre quel est l'état attendu. Reportez-vous à cette topologie pour les exemples de ce chapitre.



Workflow général de dépannage du transfert multipod

Àun niveau élevé, lors du débogage d'un problème de transfert multipod, les étapes suivantes peuvent être évaluées :

1. Le flux est-il monodiffusion ou multidestination ? N'oubliez pas que, même si le flux doit être monodiffusé en état de fonctionnement, si ARP n'est pas résolu, il s'agit d'un flux multidestination.
2. Le flux est-il routé ou ponté ? Traditionnellement, un flux routé du point de vue de l'ACI est tout flux dont l'adresse MAC de destination est l'adresse MAC du routeur appartenant à une passerelle configurée sur l'ACI. En outre, si la diffusion ARP est désactivée, le leaf d'entrée est routé en fonction de l'adresse IP cible. Si l'adresse MAC de destination n'appartient pas à l'ACI, le commutateur effectue un transfert en fonction de l'adresse MAC ou suit le comportement de monodiffusion inconnue configuré sur le domaine de pont.
3. La feuille d'entrée est-elle en train de couler ? fTriage et ELAM sont les meilleurs outils pour le confirmer.

Si le flux est monodiffusion de couche 3 :

1. Le leaf d'entrée a-t-il un terminal qui apprend pour l'IP de destination dans le même VRF que l'EPG source ? Si c’est le cas, cela prévaudra toujours sur les routes apprises. Le terminal transmet directement à l'adresse du tunnel ou à l'interface de sortie où le terminal est appris.
2. S'il n'y a pas d'apprentissage de point de terminaison, le leaf d'entrée a-t-il une route pour la destination dont l'indicateur « Pervasive » est défini ? Cela indique que le sous-réseau de destination est configuré en tant que sous-réseau de domaine Bridge et que le tronçon suivant doit être le proxy spine dans le Pod local.
3. S’il n’existe pas de route omniprésente, toutes les routes apprises par le biais d’une L3Out constituent le dernier recours. Cette partie est identique au transfert L3Out sur un seul pod.

Si le flux est unicast de couche 2 :

1. Le leaf d'entrée a-t-il un terminal qui apprend l'adresse MAC de destination dans le même domaine de pont que l'EPG source ? Si c'est le cas, le leaf transfère vers l'IP du tunnel distant ou vers l'interface locale où le point d'extrémité est appris.
2. S'il n'y a pas d'apprentissage pour l'adresse MAC de destination dans le domaine de pont source, alors le noeud terminal transfère en fonction du comportement 'unknown-unicast' sur lequel le BD est défini. S'il est défini sur 'Flood', alors le leaf inondera le groupe de multidiffusion GIPo alloué pour le domaine de pont. Les pods locaux et distants doivent recevoir une copie diffusée. S'il est défini sur « Hardware Proxy », la trame est envoyée à la colonne vertébrale pour une recherche de proxy et transférée en fonction de l'entrée COOP de la colonne vertébrale.

Étant donné que les résultats de dépannage seraient considérablement différents pour la monodiffusion par rapport au mode de diffusion globale, les résultats de travail et les scénarios de monodiffusion seront pris en compte avant de passer au mode de diffusion globale.

Workflow de dépannage de monodiffusion multipod

En suivant la topologie, parcourez le flux de 10.0.2.100 sur leaf205 à 10.0.1.100 sur leaf101.

Remarque : avant de poursuivre, il est important de vérifier si le protocole ARP de la source est résolu pour la passerelle (pour un flux routé) ou l'adresse MAC de destination (pour un flux ponté)

1. Confirmez que la feuille d'entrée reçoit le paquet. Utilisez l'outil CLI ELAM présenté dans la section « Tools » (Outils) avec la sortie ereport disponible dans la version 4.2. L'application ELAM Assistant est également utilisée.

```
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
```

Notez que l'ELAM s'est déclenché, ce qui confirme la réception du paquet sur le commutateur d'entrée. Maintenant, regardez quelques champs dans le rapport puisque le résultat est étendu.

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

Le rapport contient beaucoup plus d'informations sur l'emplacement du paquet, mais l'application ELAM Assistant est actuellement plus utile pour interpréter ces données. Le résultat de l'assistant ELAM pour ce flux sera présenté plus loin dans ce chapitre.

2. La feuille d'entrée apprend-elle la destination en tant que point d'extrémité dans le VRF d'entrée ? Dans la négative, existe-t-il une route ?

```
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
+------------------------------------+--------------+----------------+-------------+----------
---+----------------------------+
```

L'absence de résultat dans la commande ci-dessus signifie que l'adresse IP de destination n'est pas apprise. Vérifiez ensuite la table de routage.

```
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
```

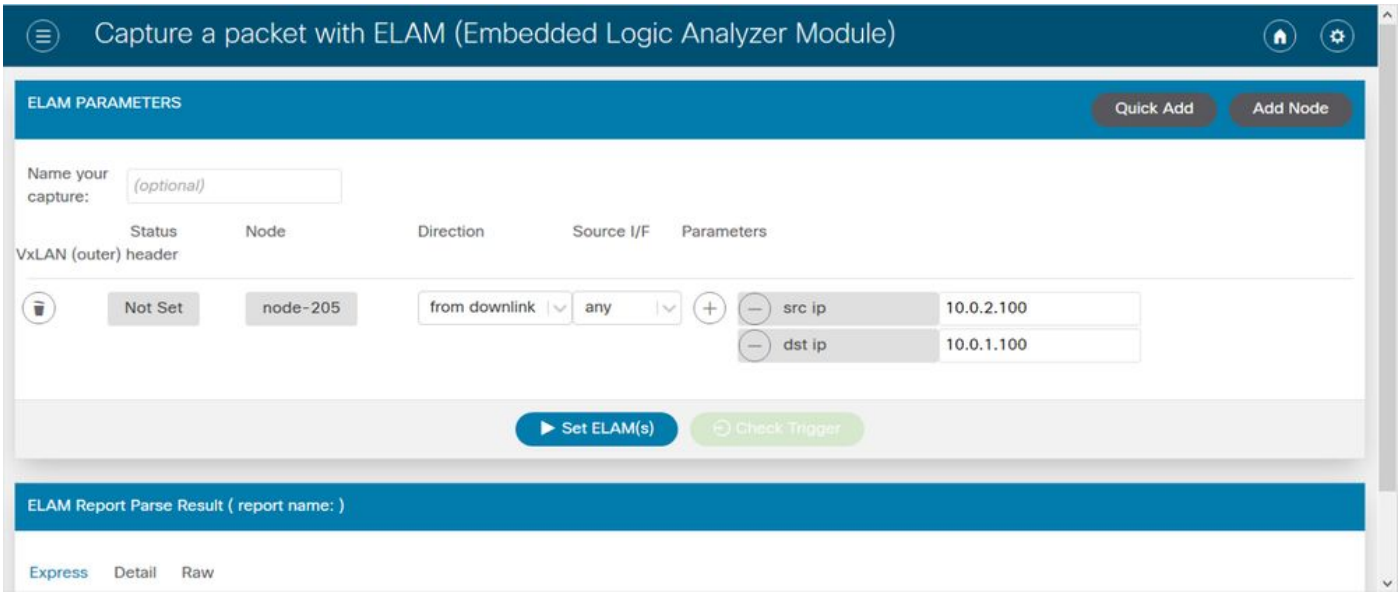
Dans le résultat ci-dessus, l'indicateur omniprésent est visible, ce qui indique qu'il s'agit d'une route de sous-réseau de domaine de pont. Le tronçon suivant doit être une adresse proxy anycast sur les spines.

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

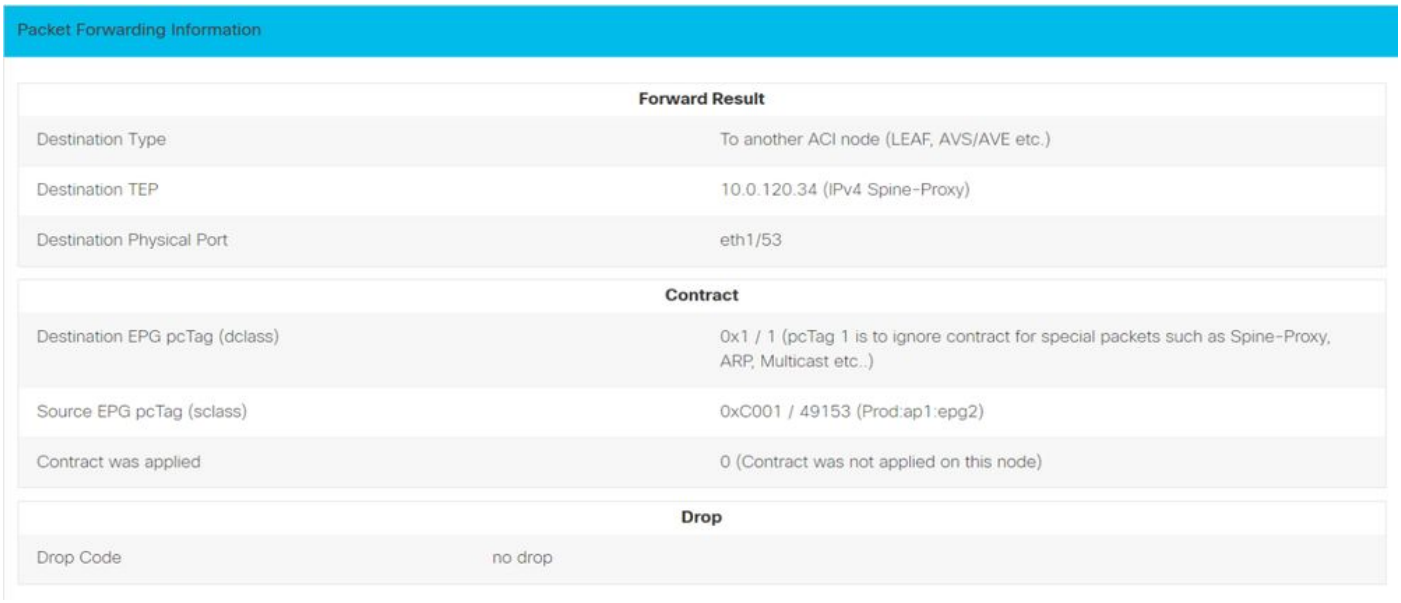
Notez que si le point d'extrémité est appris sur un tunnel ou une interface physique, cela aura la priorité, provoquant le transfert du paquet directement à cet endroit. Reportez-vous au chapitre « Transfert externe » de ce manuel pour plus de détails.

Utilisez l'assistant ELAM pour confirmer les décisions de transfert affichées dans les résultats ci-dessus.

Configuration de l'assistant ELAM



Vérification des décisions de transfert



Le résultat ci-dessus indique que le leaf d'entrée transfère le paquet à l'adresse proxy spine IPv4. C'est ce qui est attendu.

3. Confirmez sur la colonne vertébrale que l'adresse IP de destination est présente dans COOP afin que la requête proxy fonctionne.

Il y a plusieurs façons d'obtenir la sortie COOP sur la colonne vertébrale, par exemple, regardez-la avec une commande 'show coop internal info ip-db' :

```
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

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

Autres commandes à exécuter sur la colonne vertébrale :

Interroger COOP pour l'entrée L2 :

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

Interrogez COOP pour l'entrée L3 et obtenez l'entrée L2 parente :

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

Interroger COOP pour l'entrée L3 uniquement :

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

La chose utile à propos de la moquery multiple est qu'ils peuvent également être exécutés directement sur un APIC et l'utilisateur peut voir chaque colonne vertébrale qui a l'enregistrement dans coop.

4. Décision de transfert de proxy spine multipod

Si l'entrée COOP de la colonne vertébrale pointe vers un tunnel dans le Pod local, le transfert est basé sur le comportement ACI traditionnel.

Notez que le propriétaire d'un TEP peut être vérifié dans le fabric en exécutant à partir d'un APIC : **moquery -c ipv4Addr -f 'ipv4.Addr.addr=="<tunnel address>"'**

Dans le scénario de proxy, le tronçon suivant du tunnel est 10.0.0.34. Qui est le propriétaire de cette adresse IP ? :

```
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]
```

Cette adresse IP appartient aux deux noeuds spine dans la zone 1. Il s'agit d'une adresse IP spécifique appelée adresse proxy externe. De la même manière que l'ACI possède des adresses proxy détenues par les noeuds spine au sein d'un Pod (voir l'étape 2 de cette section), des adresses proxy sont également attribuées au Pod lui-même. Ce type d'interface peut être vérifié en exécutant :

```
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]
```

L'indicateur « external » indique qu'il s'agit d'un proxy externe TEP.

5. Vérifiez le protocole EVPN BGP sur la colonne vertébrale

L'enregistrement du point de terminaison de coopération doit être importé à partir de BGP EVPN sur la colonne vertébrale. La commande suivante peut être utilisée pour vérifier qu'il est dans EVPN (bien que s'il est déjà dans COOP avec un tronçon suivant du TEP externe du Pod distant, on peut supposer qu'il provient d'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
```

Notez que la commande ci-dessus peut également être exécutée pour une adresse MAC.

-192.168.1.254 est le plan de données TEP configuré lors de la configuration de plusieurs pods. Notez cependant que même s'il est annoncé dans BGP comme NH, le prochain saut réel sera le

TEP proxy externe.

-192.168.1.101 et .102 sont les noeuds spine du Pod 1 annonçant ce chemin.

6. Vérifiez COOP sur les spines dans le Pod de destination.

La même commande que précédemment peut être utilisée :

```
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

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

Vérifiez qui possède l'adresse du tunnel en exécutant la commande suivante sur un APIC :

```
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
```

La commande ci-dessus montre que le tunnel de COOP pointe vers leaf101. Cela signifie que leaf101 doit avoir l'apprentissage local pour le point d'extrémité de destination.

7. Vérifiez que le leaf de sortie possède l'apprentissage local.

Cela peut être fait via une commande « show endpoint » :

```
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
```

Notez que le terminal est appris. Le paquet doit être transféré en fonction du port-channel 5 avec l'étiquette VLAN 1075 définie.

Utilisation de fTriage pour vérifier le flux de bout en bout

Comme indiqué dans la section « Outils » de ce chapitre, fTriage peut être utilisé pour mapper un flux existant de bout en bout et comprendre ce que chaque commutateur du chemin fait du paquet. Cela est particulièrement utile dans les déploiements plus importants et plus complexes tels que Multi-Pod.

Notez que l'exécution complète de fTriage prendra un certain temps (15 minutes potentielles).

Lors de l'exécution de fTriage sur le flux exemple :

```
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"}}}
```

Le triage contient une grande quantité de données. Certains des champs les plus importants sont mis en surbrillance. Notez que le chemin du paquet était 'leaf205 (Pod 2) > spine3 (Pod 2) > spine2 (Pod 1) > leaf101 (Pod 1)'. Toutes les décisions de transfert et les recherches de contrat effectuées en cours de route sont également visibles.

Notez que s'il s'agissait d'un flux de couche 2, la syntaxe du fTriage devrait être définie sur quelque chose comme :

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

Requêtes proxy lorsque l'EP n'est pas dans COOP

Avant d'envisager des scénarios d'échec spécifiques, il y a un autre élément à discuter en rapport avec le transfert de monodiffusion sur Multi-Pod. Que se passe-t-il si le point de terminaison de destination est inconnu, si la requête est envoyée par proxy et si le point de terminaison n'est pas en COOP ?

Dans ce scénario, le paquet/la trame est envoyé à la colonne vertébrale et une demande de glanage est générée.

Lorsque le spine génère une requête de glane, le paquet d'origine est conservé dans la requête, mais le paquet reçoit l'ethertype 0xfff2 qui est un Ethertype personnalisé réservé aux glanes. Pour cette raison, il ne sera pas facile d'interpréter ces messages dans des outils de capture de paquets tels que Wireshark.

La destination de couche 3 externe est également définie sur 239.255.255.240, qui est un groupe de multidiffusion réservé spécifiquement pour les messages de glanage. Ceux-ci doivent être diffusés à travers le fabric et tout commutateur leaf de sortie dont le sous-réseau de destination de la requête de glane est déployé génère une requête ARP pour résoudre la destination. Ces ARP sont envoyés à partir de l'adresse IP de sous-réseau BD configurée (par conséquent, les requêtes proxy ne peuvent pas résoudre l'emplacement des points d'extrémité silencieux/inconnus si le

routage de monodiffusion est désactivé sur un domaine de pont).

La réception du message de glanage sur la feuille de sortie et la réponse ARP générée et reçue par la suite peuvent être vérifiées à l'aide de la commande suivante :

Vérification Glean ARP

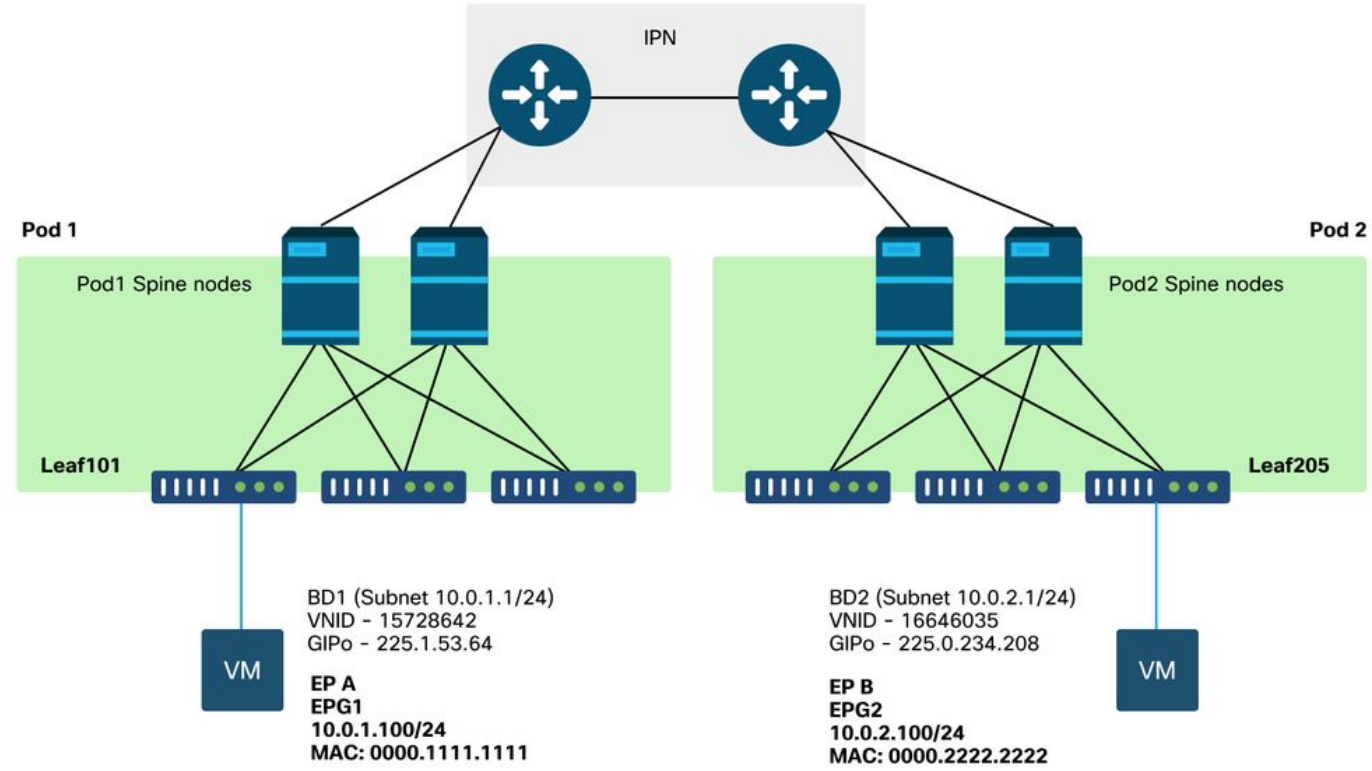
```
a-leaf205# show ip arp internal event-history event | grep -F -B 1 192.168.21.11
...
73) Event:E_DEBUG_DSF, length:127, at 316928 usecs after Wed May  1 08:31:53 2019
Updating epm ifidx: 1a01e000 vlan: 105 ip: 192.168.21.11, ifMode: 128 mac: 8c60.4f02.88fc   <<<
Endpoint is learned
75) Event:E_DEBUG_DSF,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

Pour référence, les messages de glanage envoyés à 239.255.255.240 sont la raison pour laquelle ce groupe doit être inclus dans la plage de groupes PIM bidirectionnels sur l'IPN.

Scénario de dépannage de plusieurs pods #1 (monodiffusion)

Dans la topologie suivante, EP B ne peut pas communiquer avec EP A.

Dépannage de la topologie



Notez que la plupart des problèmes observés pour le transfert multipod sont identiques aux problèmes observés dans un seul pod. Pour cette raison, les problèmes spécifiques à Multi-Pod sont concentrés sur.

Tout en suivant le workflow de dépannage de monodiffusion décrit précédemment, notez que la requête est transmise par proxy, mais que les noeuds spine dans la zone 2 n'ont pas l'adresse IP de destination dans COOP.

Motif: Point de terminaison manquant dans COOP

Comme indiqué précédemment, les entrées COOP pour les terminaux Pod distants sont renseignées à partir des informations EVPN BGP. Par conséquent, il est important de déterminer :

a.) La colonne vertébrale du Pod source (Pod 2) est-elle dans EVPN ?

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

b.) La colonne vertébrale du Pod distant (Pod 1) est-elle dans EVPN ?

```
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:
```

La colonne vertébrale du Pod 1 l'a et l'IP de tronçon suivant est 0.0.0.0 ; cela signifie qu'il a été exporté de COOP localement. Notez toutefois que la section « Annoncé aux homologues » n'inclut pas les noeuds spine de la zone 2.

c.) Le réseau EVPN BGP est-il activé entre les pods ?

```
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
```

Notez dans le résultat ci-dessus que les homologues EVPN BGP sont désactivés entre les pods. Toute valeur autre qu'une valeur numérique dans la colonne State/PfxRcd indique que la contiguïté n'est pas active. Les EP de la zone 1 ne sont pas acquis via EVPN et ne sont pas importés dans COOP.

Si ce problème se produit, vérifiez les points suivants :

1. Le protocole OSPF est-il activé entre les noeuds spine et les réseaux IP connectés ?
2. Les noeuds spine ont-ils des routes apprises via OSPF pour les IP spine distantes ?
3. Le chemin complet à travers l'IPN prend-il en charge le MTU jumbo ?
4. Toutes les contiguïtés de protocole sont-elles stables ?

Autres causes possibles

Si le point d'extrémité ne se trouve pas dans la base de données COOP d'un Pod et que le périphérique de destination est un hôte silencieux (non appris sur un commutateur leaf du fabric), vérifiez que le processus de nettoyage du fabric fonctionne correctement. Pour que cela fonctionne :

- Le routage monodiffusion doit être activé sur le BD.
- La destination doit se trouver dans un sous-réseau BD.
- Le réseau IP doit fournir un service de routage multidiffusion pour le groupe 239.255.255.240.

La partie multidiffusion est traitée plus en détail dans la section suivante.

Présentation de la diffusion multipod, de la monodiffusion inconnue et du transfert multidiffusion (BUM)

Dans l'ACI, le trafic est inondé via des groupes de multidiffusion superposés dans de nombreux scénarios différents. Par exemple, une inondation se produit pour :

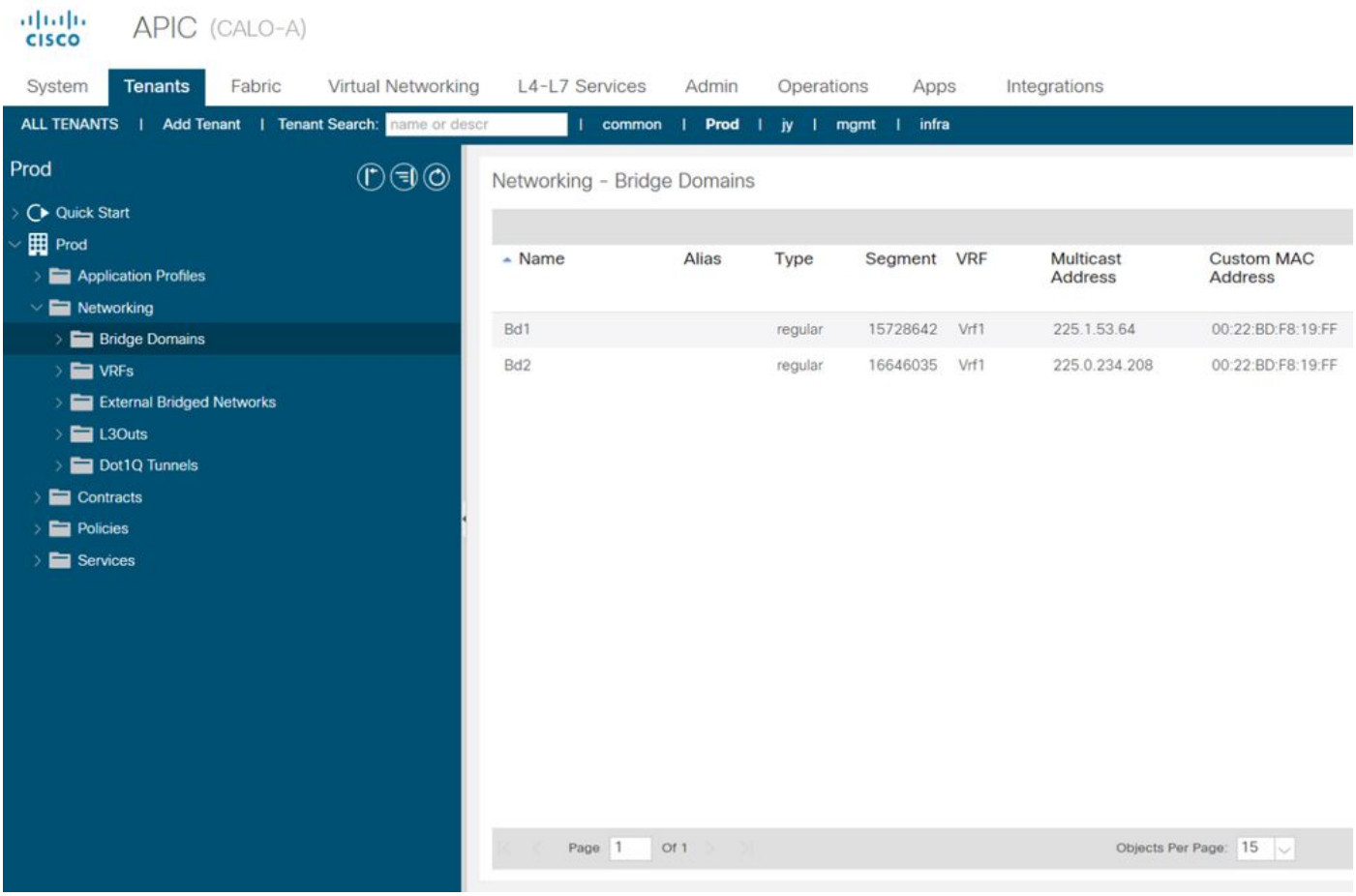
- Trafic de multidiffusion et de diffusion.

- Monodiffusion inconnue qui doit être diffusée.
- Messages de nettoyage ARP du fabric.
- Messages d'annonce EP.

De nombreuses fonctionnalités reposent sur le transfert BUM.

Dans l'ACI, une adresse de multidiffusion appelée adresse externe IP de groupe (ou GIPo) est attribuée à tous les domaines de pont. Tout le trafic qui doit être diffusé dans un domaine Bridge est diffusé sur ce GIPo.

BD GIPo dans GUI



L'objet peut être interrogé directement sur l'un des APIC.

BD GIPo dans Moquery

```
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
```

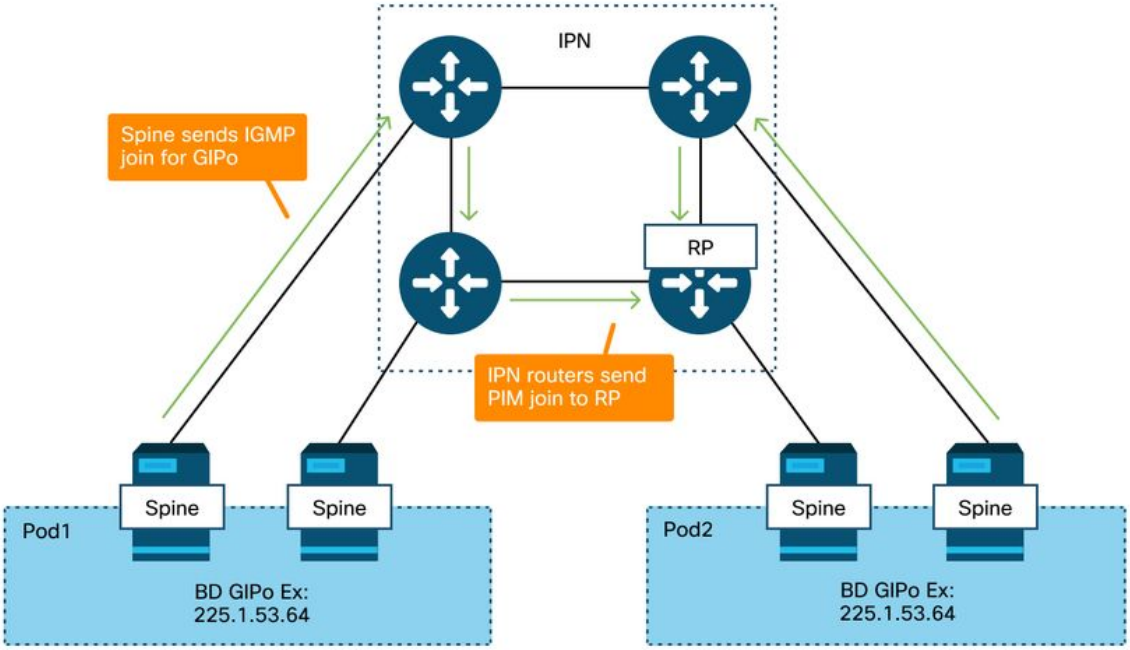
```
epClear                  : no
epMoveDetectMode         :
extMngdBy                :
hostBasedRouting         : no
intersiteBumTrafficAllow : no
intersiteL2Stretch       : no
ipLearning               : yes
ipv6McastAllow           : no
lcOwn                    : local
limitIpLearnToSubnets    : yes
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
```

Les informations ci-dessus sur l'inondation GIPo sont vraies, que Multi-Pod soit utilisé ou non. La partie supplémentaire de ce qui concerne Multi-Pod est le routage de multidiffusion sur l'IPN.

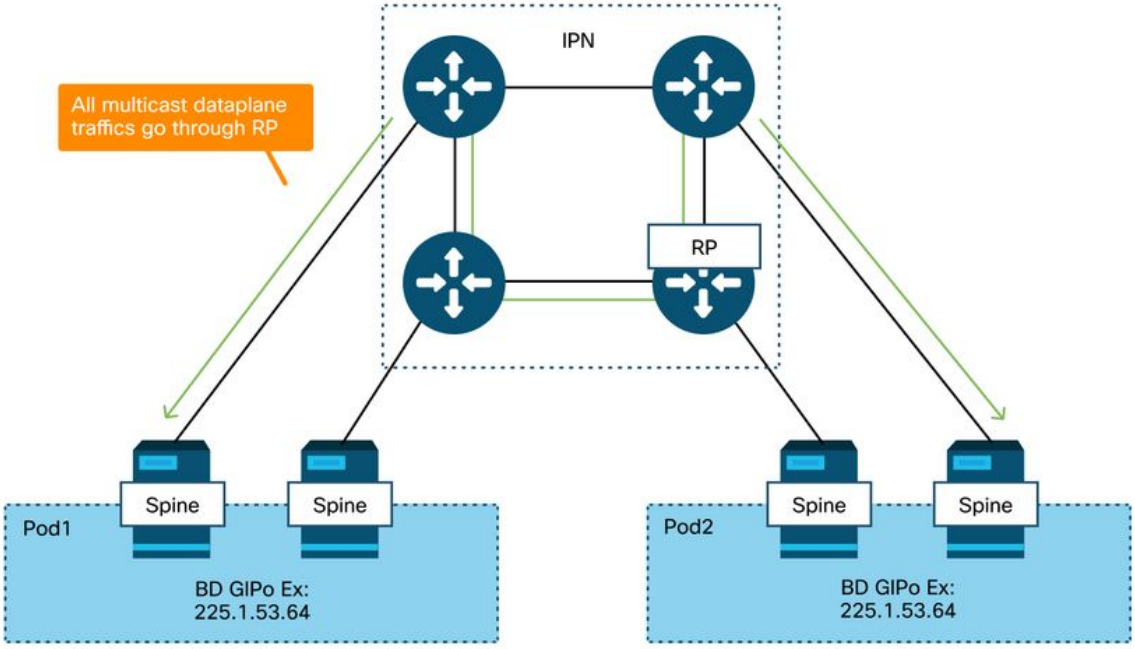
Le routage multidiffusion IPN implique les opérations suivantes :

- Les noeuds spine agissent comme des hôtes de multidiffusion (IGMP uniquement). Ils n'exécutent pas PIM.
- Si un BD est déployé dans un POD, alors un spine de ce POD enverra une jointure IGMP sur l'une de ses interfaces orientées IPN. Cette fonctionnalité est répartie sur tous les noeuds spine et l'interface IPN sur de nombreux groupes.
- Les IPN reçoivent ces jointures et envoient des jointures PIM vers le RP PIM bidirectionnel.
- PIM Bidir étant utilisé, il n'y a pas d'arbres (S, G). Seuls les arbres (*, G) sont utilisés dans le Bidir PIM.
- Tout le trafic du plan de données envoyé au GIPo passe par le RP.

Plan de contrôle de multidiffusion IPN



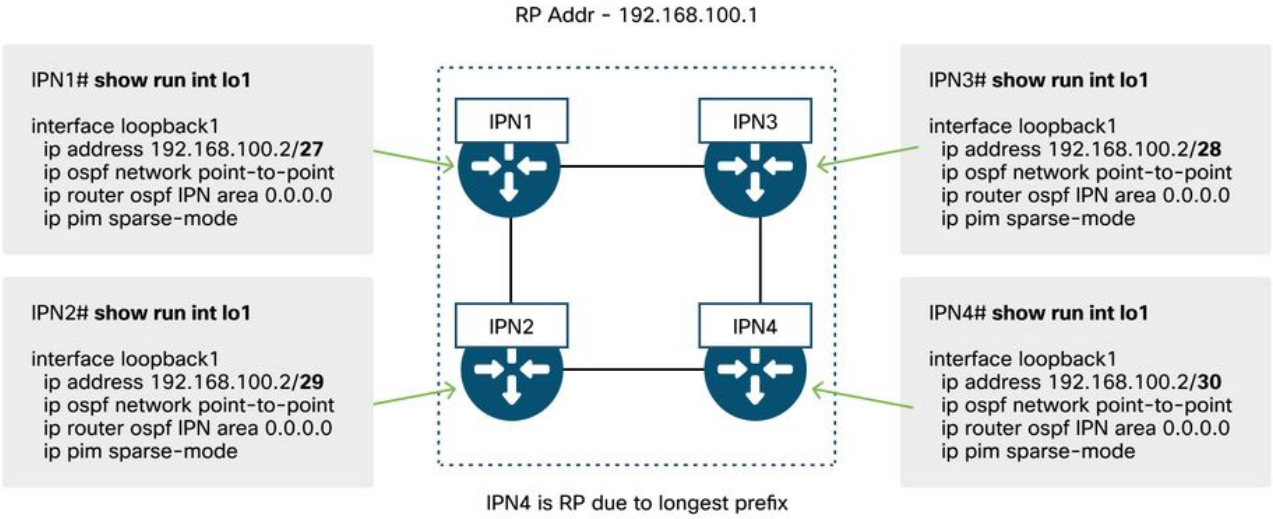
Plan de données multicast IPN



Le seul moyen de redondance RP avec PIM Bidir est d'utiliser Phantom. Ce sujet est traité en détail dans la section Découverte de plusieurs pods de ce livre. En résumé, notez qu'avec Phantom RP :

- Tous les réseaux IP doivent être configurés avec la même adresse RP.
- L'adresse RP exacte ne doit exister sur aucun périphérique.
- Plusieurs périphériques annoncent l'accessibilité au sous-réseau qui contient l'adresse IP du RP fantôme. Les sous-réseaux annoncés doivent varier en longueur de sous-réseau de sorte que tous les routeurs s'accordent sur le meilleur chemin annoncé pour le RP. Si ce chemin est perdu, alors la convergence dépend de l'IGP.

Configuration du RP fantôme



Workflow de dépannage de diffusion multipod, monodiffusion inconnue et multidiffusion (BUM)

1. Vérifiez d'abord si le flux est réellement traité comme multidestination par le fabric.

Le flux sera inondé dans le BD dans ces exemples courants :

- La trame est une diffusion ARP et la diffusion ARP est activée sur le BD.
- La trame est destinée à un groupe de multidiffusion. Notez que même si la surveillance IGMP est activée, le trafic est toujours diffusé dans le fabric sur le GIPo.
- Le trafic est destiné à un groupe de multidiffusion pour lequel l'ACI fournit des services de routage multidiffusion.
- Le flux est de couche 2 (flux ponté) et l'adresse MAC de destination est inconnue et le comportement de monodiffusion inconnu sur le BD est défini sur « Flood ».

La manière la plus simple de déterminer quelle décision de transmission sera prise est avec un ELAM.

2. Identifiez le GIPo BD.

Reportez-vous à la section précédente de ce chapitre qui traite de ce sujet. Les ELAM Spine peuvent également être exécutés via l'application ELAM Assistant pour vérifier que le trafic inondé est reçu.

3. Vérifiez les tables de routage de multidiffusion sur l'IPN pour ce GIPo.

Les résultats obtenus varient en fonction de la plate-forme IPN utilisée, mais à un niveau élevé :

- Tous les routeurs IPN doivent convenir du RP et le RPF pour ce GIPo doit pointer vers cette arborescence.
- Un routeur IPN connecté à chaque Pod doit obtenir une connexion IGMP pour le groupe.

Scénario de dépannage de plusieurs pods #2 (flux BUM)

Ce scénario couvrirait tout scénario impliquant que le protocole ARP n'est pas résolu dans les scénarios Multi-Pod ou BUM (monodiffusion inconnue, etc.).

Il y a plusieurs causes possibles communes ici.

Cause possible 1 : Plusieurs routeurs possèdent l'adresse PIM RP

Avec ce scénario, la feuille d'entrée inonde le trafic (vérifiez avec ELAM), le pod source reçoit et inonde le trafic, mais le pod distant ne l'obtient pas. Pour certains BD, l'inondation fonctionne, mais pour d'autres non.

Sur l'IPN, exécutez « show ip mroute <adresse GIPo> » pour le GIPo afin de voir que l'arborescence RPF pointe vers plusieurs routeurs différents.

Dans ce cas, vérifiez les points suivants :

- Vérifiez que l'adresse PIM RP réelle n'est configurée nulle part. Tout périphérique qui possède cette adresse RP réelle verrait une route /32 locale pour elle.
- Vérifiez que plusieurs routeurs IPN n'annoncent pas la même longueur de préfixe pour le RP dans le scénario de RP fantôme.

Cause possible 2 : Les routeurs IPN n'apprennent pas les routes pour l'adresse RP

De la même manière que pour la première cause possible, le trafic inondé ne quitte pas l'IPN. Le résultat de la commande « show ip route <rp address> » sur chaque routeur IPN indique uniquement la longueur de préfixe configurée localement plutôt que ce que les autres routeurs annoncent.

Le résultat est que chaque périphérique pense être le RP même si l'adresse IP réelle du RP n'est configurée nulle part.

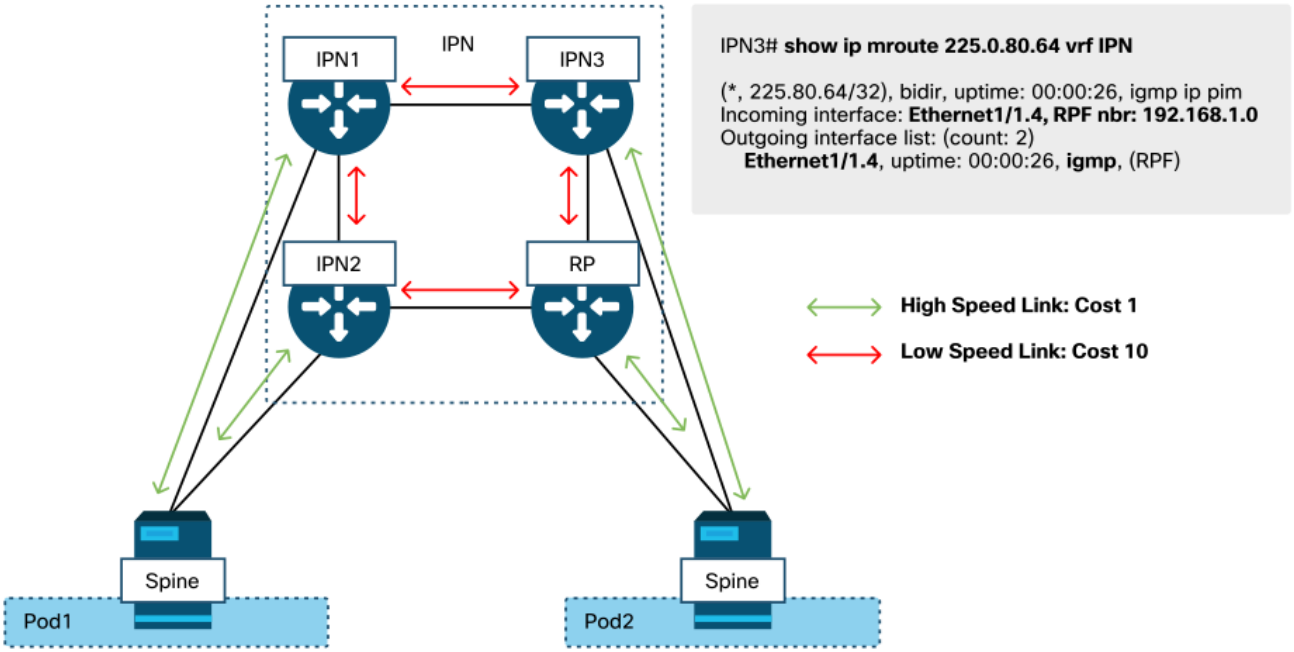
Si c'est le cas. vérifiez les points suivants :

- Vérifiez que les contiguïtés de routage sont actives entre les routeurs IPN. Vérifiez que la route se trouve dans la base de données de protocole réelle (telle que la base de données OSPF).
- Vérifiez que tous les bouclages qui sont censés être des RP candidats sont configurés en tant que types de réseau OSPF point à point. Si ce type de réseau n'est pas configuré, chaque routeur annoncera toujours une longueur de préfixe /32, quelle que soit la configuration réelle.

Cause possible 3 : Les routeurs IPN n'installent pas la route GIPo ou le RPF pointe vers l'ACI

Comme mentionné précédemment, l'ACI n'exécute pas PIM sur ses liaisons orientées IPN. Cela signifie que le meilleur chemin de l'IPN vers le RP ne doit jamais pointer vers l'ACI. Le scénario où cela pourrait se produire serait si plusieurs routeurs IPN sont connectés au même spine et qu'une meilleure métrique OSPF est vue à travers le spine qu'directement entre les routeurs IPN.

Interface RPF vers ACI



Pour résoudre ce problème :

- Assurez-vous que les contiguïtés de protocole de routage entre les routeurs IPN sont actives.
- Augmentez les métriques de coût OSPF pour les liaisons orientées IPN sur les noeuds spine à une valeur qui rendra cette métrique moins préférable aux liaisons IPN à IPN.

Autres références

Avant la version 4.0 du logiciel ACI, l'utilisation de la COS 6 par des périphériques externes présentait des difficultés. La plupart de ces problèmes ont été résolus par le biais d'améliorations de la version 4.0, mais pour plus d'informations, reportez-vous à la session Cisco Live « BRKACI-2934 - Troubleshooting Multi-Pod » et à la section « Quality of Service ».

À propos de cette traduction
Cisco a traduit ce document en traduction automatisée vérifiée par une personne dans le cadre d'un service mondial permettant à nos utilisateurs d'obtenir le contenu d'assistance dans leur propre langue.

Il convient cependant de noter que même la meilleure traduction automatisée ne sera pas aussi précise que celle fournie par un traducteur professionnel.