ACI Forwarding Behavior Deep Dive

Cisco Systems Inc.

2020年 4月

Agenda

1. [復習] Endpoint Learning 基礎 : Endpoint (Local / Remote), COOP DB
2. ACI 内部転送動作
3. ACI における VLAN と VXLAN
4. ACI 転送関連パラメータ : Pervasive Gateway, Pervasive Route, 転送スコープ, BDパラメータ, Spine Proxy
5. Packet Walk
6. まとめ

参考資料

- ACI Fabric Endpoint Learning White Paper

  https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-739989.html

- **Cisco Live! BRKACI-3545 “ACI Forwarding Behavior”**

  https://www.ciscolive.com/c/dam/r/ciscolive/emea/docs/2020/pdf/BRKACI-3545.pdf

以下、一般的なACIの導入・運用管理においては必要とはしないであろうものの、興味のある方向け (※本資料でも扱い範囲外)

- EX Hardware: ACI Packet Forwarding Deep Dive

  https://www.cisco.com/c/en/us/support/docs/cloud-systems-management/application-policy-infrastructure-controller-apic/213346-ex-hardware-aci-packet-forwarding-deep.html (英語)

  https://www.cisco.com/c/ja_jp/support/docs/cloud-systems-management/application-policy-infrastructure-controller-apic/213346-ex-hardware-aci-packet-forwarding-deep.html (日本語)

略語表記

略語	正式名称
ACI	Application Centric Infrastructure
APIC	Application Policy Infrastructure Controller
EP	Endpoint
EPG	Endpoint Group
BD	Bridge Domain
VRF	Virtual Routing and Forwarding
COOP	Council of Oracle Protocol
VXLAN	Virtual Extensible LAN
VNID	Virtual Network Identifier
TEP	Tunnel Endpoint
pcTag	Policy Control Tag
sclass	Source Class (=pcTag)

表記	意味
dXXXo	Outer Destination XXX (MAC/IP)
sXXXo	Outer Source XXX (MAC/IP)
dXXXi	Inner Destination XXX (MAC/IP)
sXXXi	Inner Source XXX (MAC/IP)
GIPo	Outer Multicast Group IP

[復習] Endpoint 基礎

※詳細は “ACI Design : Endpoint Learning Deep Dive” 参照

Endpoint とは ACI 内部の接続端末情報

従来のネットワーク機器におけるMACアドレスとIPアドレスの学習動作とは異なり、ACIではホストルートとしてMACアドレステーブルとARPテーブルではなくEndpointテーブルを利用する(外部接続を除く)。ホストルート以外はRIBが合わせて参照される。

Table	含まれる情報	概要
RIB	Routing Table (内部ホストルートを除く)	ACIにおいてもVRF毎にRouting Tableが構成され、必要に応じてAPICからLeafに対してプログラムされる
Endpoint	MACアドレス・IPアドレス	1つのエントリーにMACアドレスは1つ、IPアドレスは なし or 1つ or 複数 IPアドレスはホストルート /32 (IPv4) or /128 (IPv6) でのみ学習する
ARP	L3out 隣接ノードについては、ARPに基づいて IPアドレスに紐づく対向ピアのMACアドレスを学習する	

ACI内部ではARP(= Control-plane)のみに依存せず、Leafスイッチに届いたパケットの送信元MACアドレス・IPアドレスから "も" Endpointを学習する(= Data-plane Learning)。

※DownlinkからのパケットではLocal Endpoint、UplinkからのパケットではRemote Endpointとして学習する

Local Endpoint と Remote Endpoint

各Leafスイッチは自身のDownlink側から学習したEndpoint情報をLocal Endpoint、Fabric Port側から学習したEndpoint情報をRemote Endpointとして区別して扱う。

Endpoint種別	学習対象	利用目的
Local Endpoint	1エントリに必ず1つのMACアドレスと、必要に応じて1つ以上のIPアドレスの組合せ	自身の配下に存在するEndpoint情報の把握と、COOPを通じたSpineへの通知
Remote Endpoint	VRFのVNIDに紐づく場合はIPアドレスのみ、BDのVNIDに紐づく場合はMACアドレスのみ	キャッシュとポリシー適用の最適化のために利用 ※COOPによる通知対象ではない

Local Endpoint、Remote Endpoint いずれについてもACIはデフォルトではData-planeでの学習も行う(送信元MAC/IPアドレスに基づく学習)。※ARP/ND学習も利用する

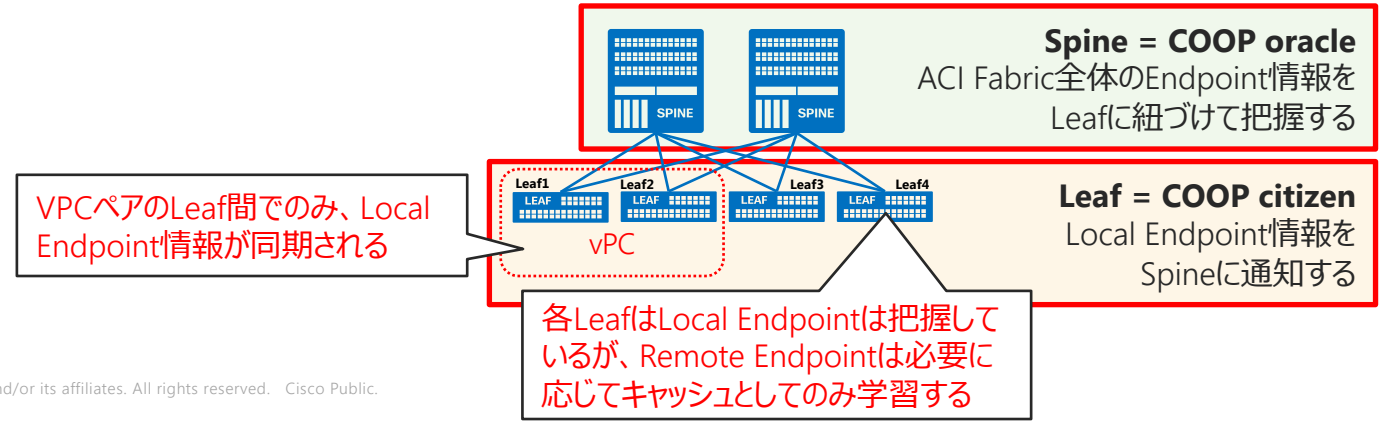
※VPC構成の場合はVPCノードはもちろん、Orphanノードについても VPCピア間で同期されるため、これらのノードはVPCペア同士でLocal Endpointとして扱われる。

COOP (Council of Oracles Protocol) DB

ACI Fabric では Leaf スイッチと Spine スイッチは明確に役割が分けられている。

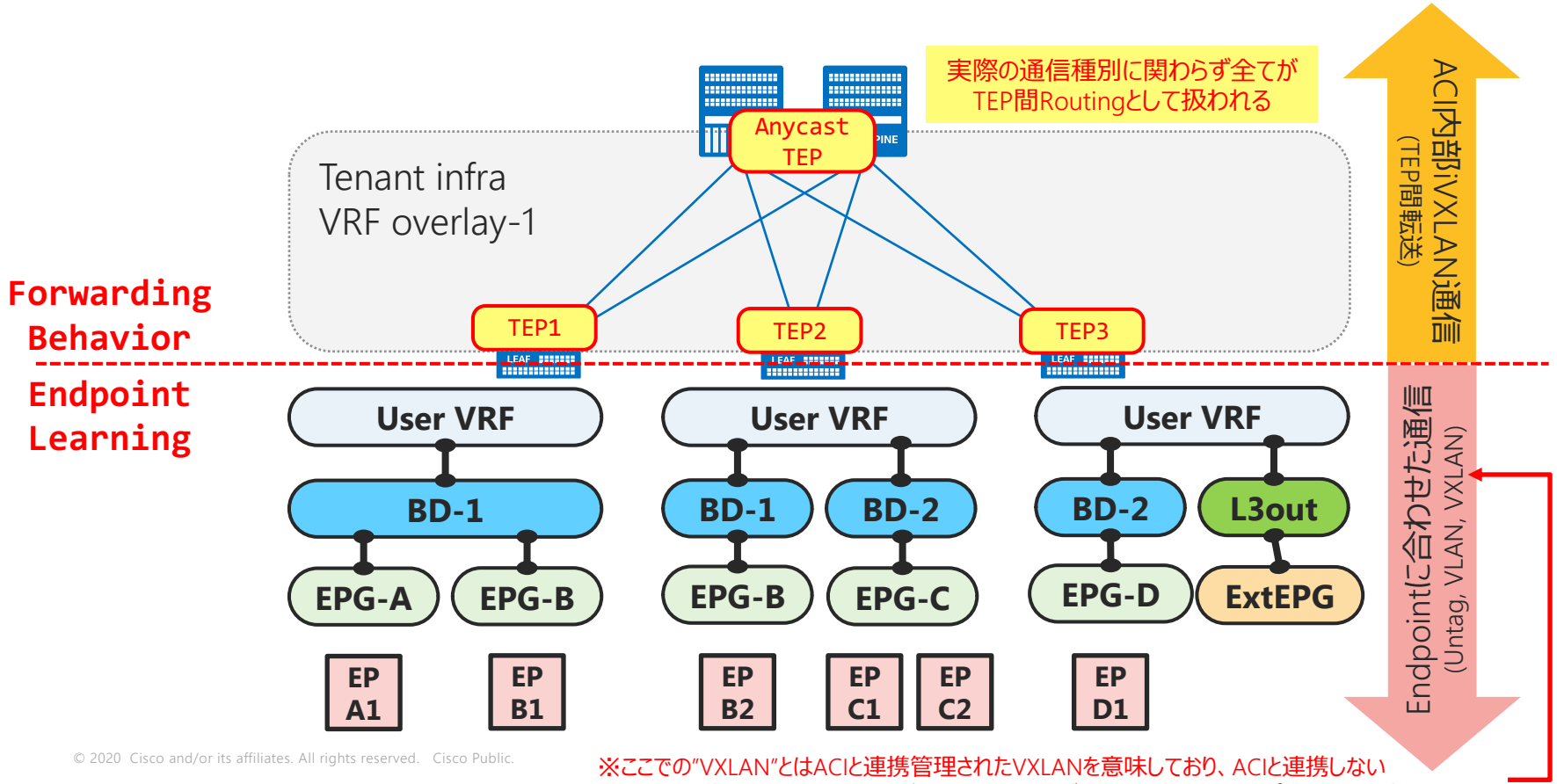
- **Leaf = COOP citizen** : 学習した Local Endpoint 情報を COOP を通じて Spine スイッチに通知する(Zero Message Queue : ZMQ)
- **Spine = COOP oracle** : COOP から COOP DB を構成し相互に同期する

※VPCを構成しているLeaf同士の間でのみ、Orphan port を含む Local Endpoint情報が同期される。



ACI 内部転送動作

ACI 内部転送 = TEP間 iVXLAN Overlay



※ここでの"VXLAN"とはACIと連携管理されたVXLANを意味しており、ACIと連携しない
Host側で付加されたVXLAN等は単に VLAN Tag が付加された IP(UDP)パケット として扱われます。

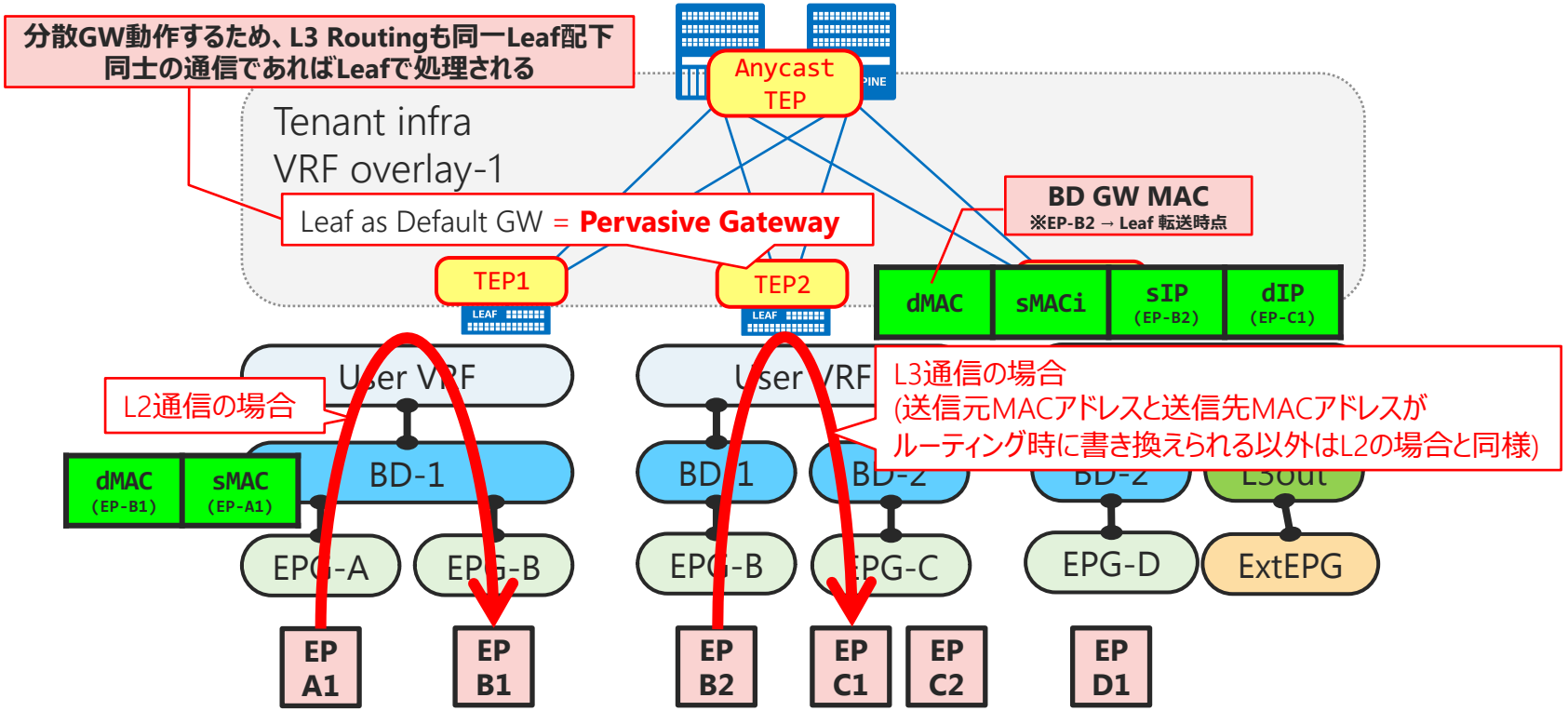
ACI 内部転送パターン

1. **同一 Leaf 配下の Endpoint 間での通信**
2. **別 Leaf 配下の Endpoint との通信**

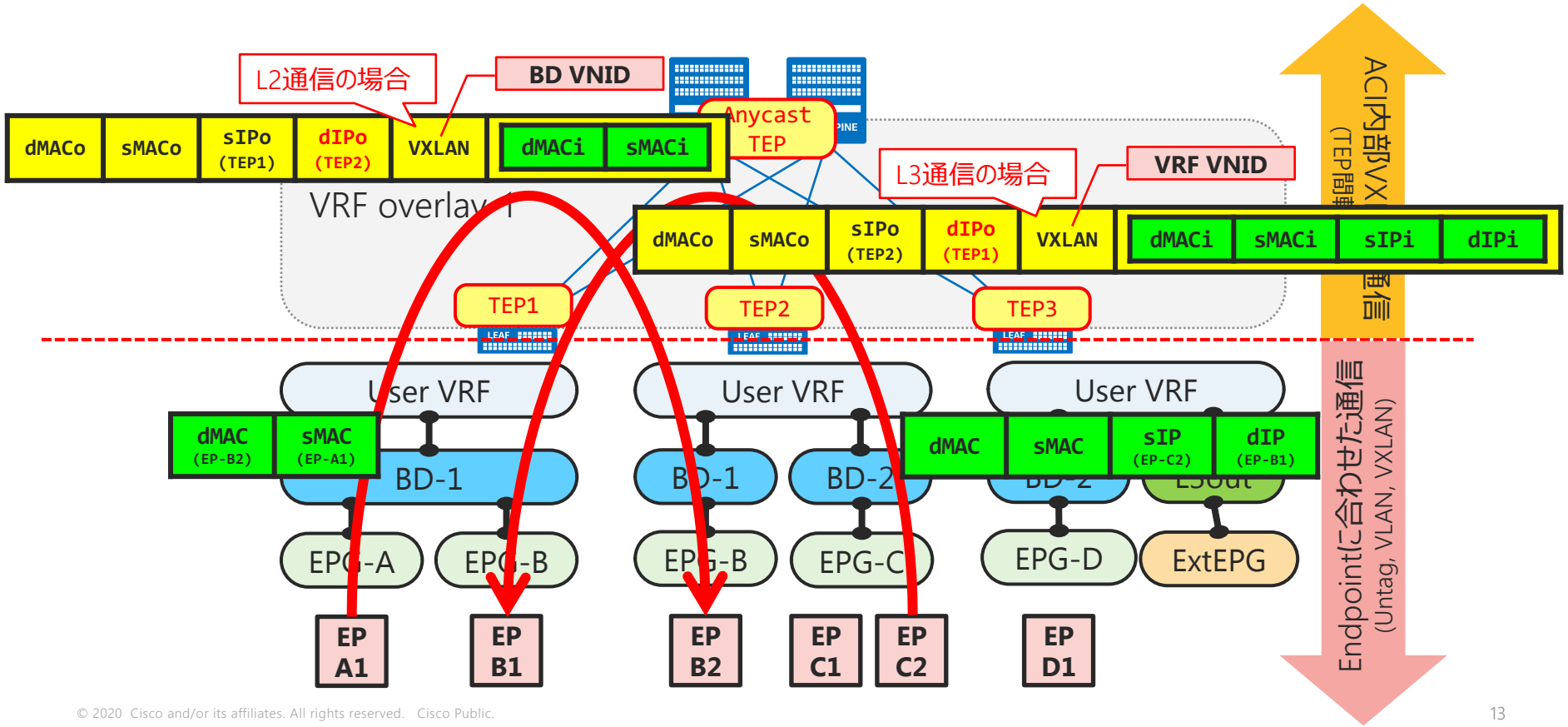
   (送信元 Leaf が宛先 Endpoint が存在する送信先 Leaf を把握している場合)
3. **別 Leaf 配下の Endpoint との通信**

   (送信元 Leaf が宛先 Endpoint が存在する送信先 Leaf を把握していない場合)

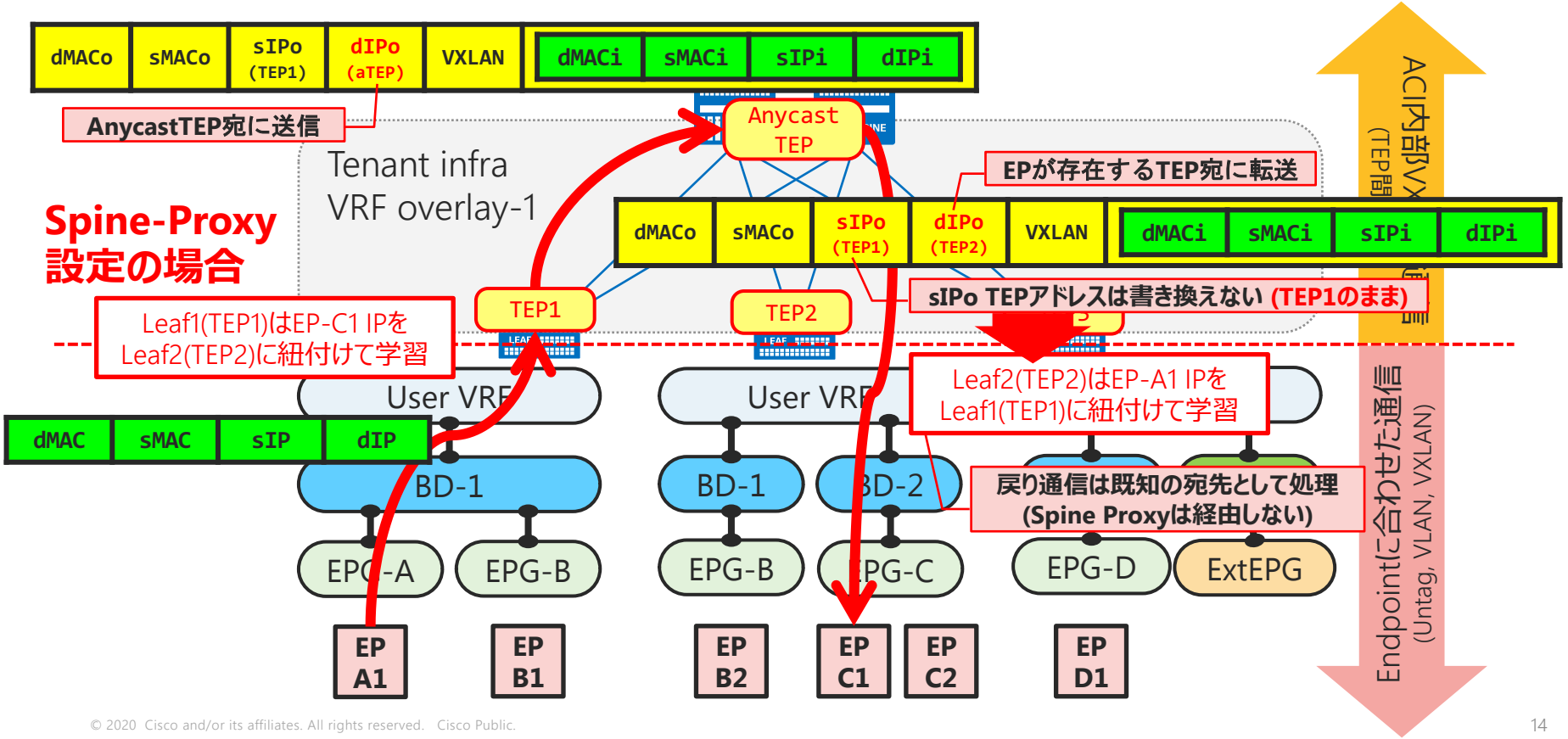
ACI 転送パターン(1) : 同一Leaf配下



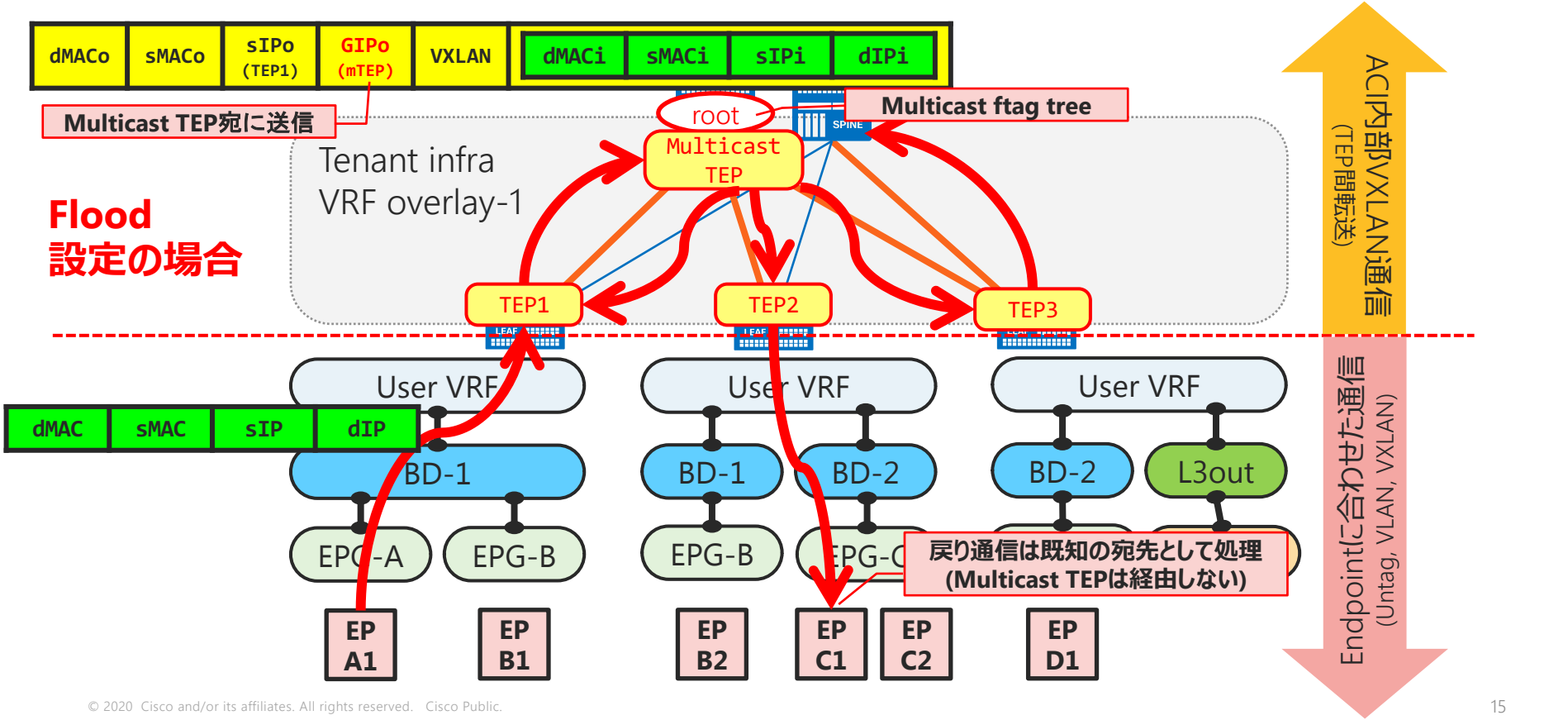
ACI 転送パターン(2) : 別Leaf配下/既知の宛先



ACI 転送パターン(3a) : 別Leaf配下/未知の宛先

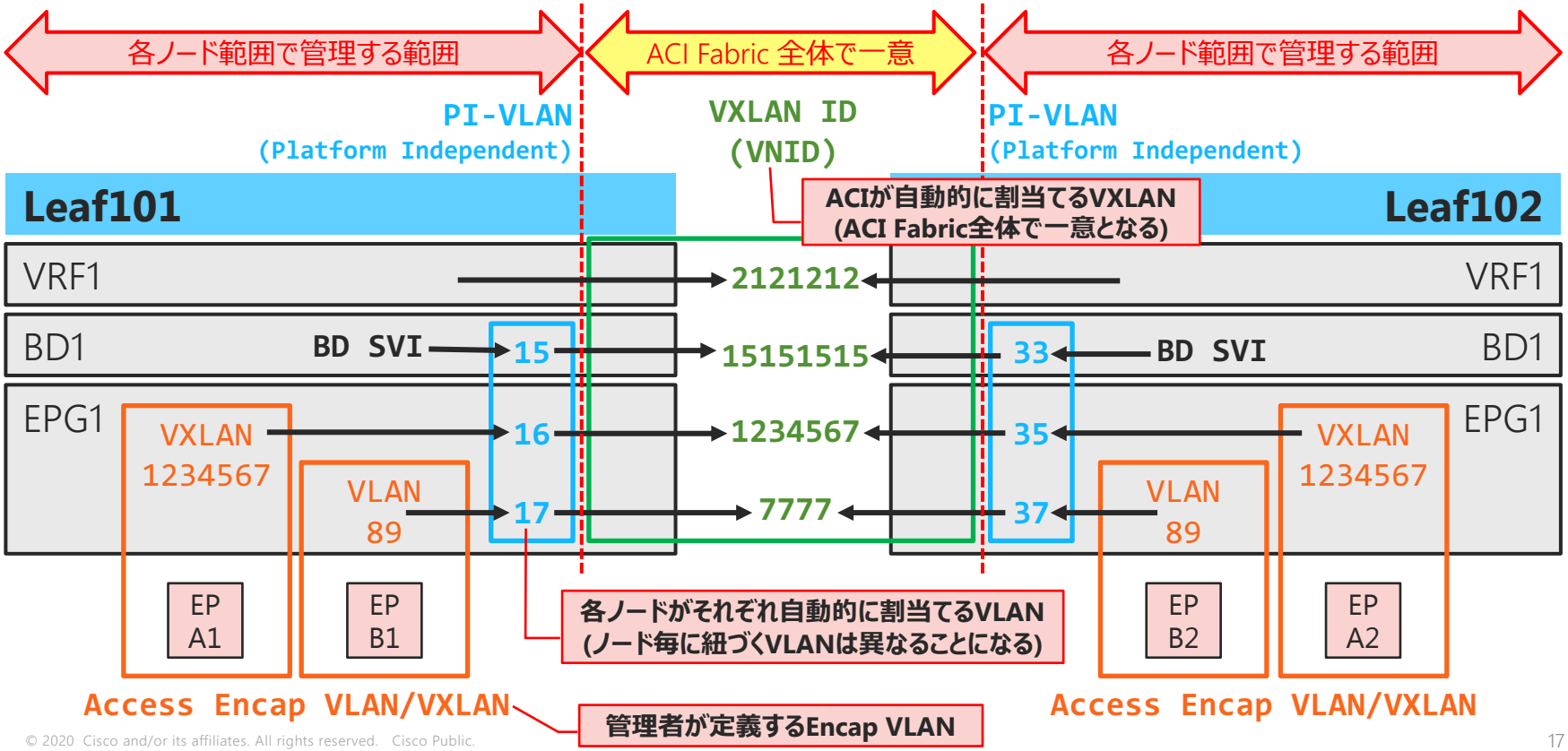


ACI 転送パターン(3b) : 別Leaf配下/未知の宛先



ACI における VLAN と VXLAN

ACIにおける VLAN と VXLAN の扱い



※ここでの"VXLAN"とはACIと連携管理されたVXLANを意味しており、ACIと連携しないHost側で付加されたVXLAN等は単に VLAN Tag が付加された IP(UDP)パケット として扱われます。

Endpoint Table と VLAN Table における VLAN

EndpointからVLAN(≒EPG)のPI-VLANを確認する(mac or ip)

```
Pod1-Leaf1# show endpoint ip 192.168.1.221
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
      Domain                          VLAN            IP Address        IP Info
+------------------------------------+---------------+-----------------+--------------+-------------+
39                                         vlan-511    0050.5680.e818 L                    eth1/34
tsetaka_demo1:VRF1                         vlan-511     192.168.1.221 L                    eth1/34
```

PI-VLAN

VLAN(≒EPG)に紐づくPI-VLAN

Encap VLAN

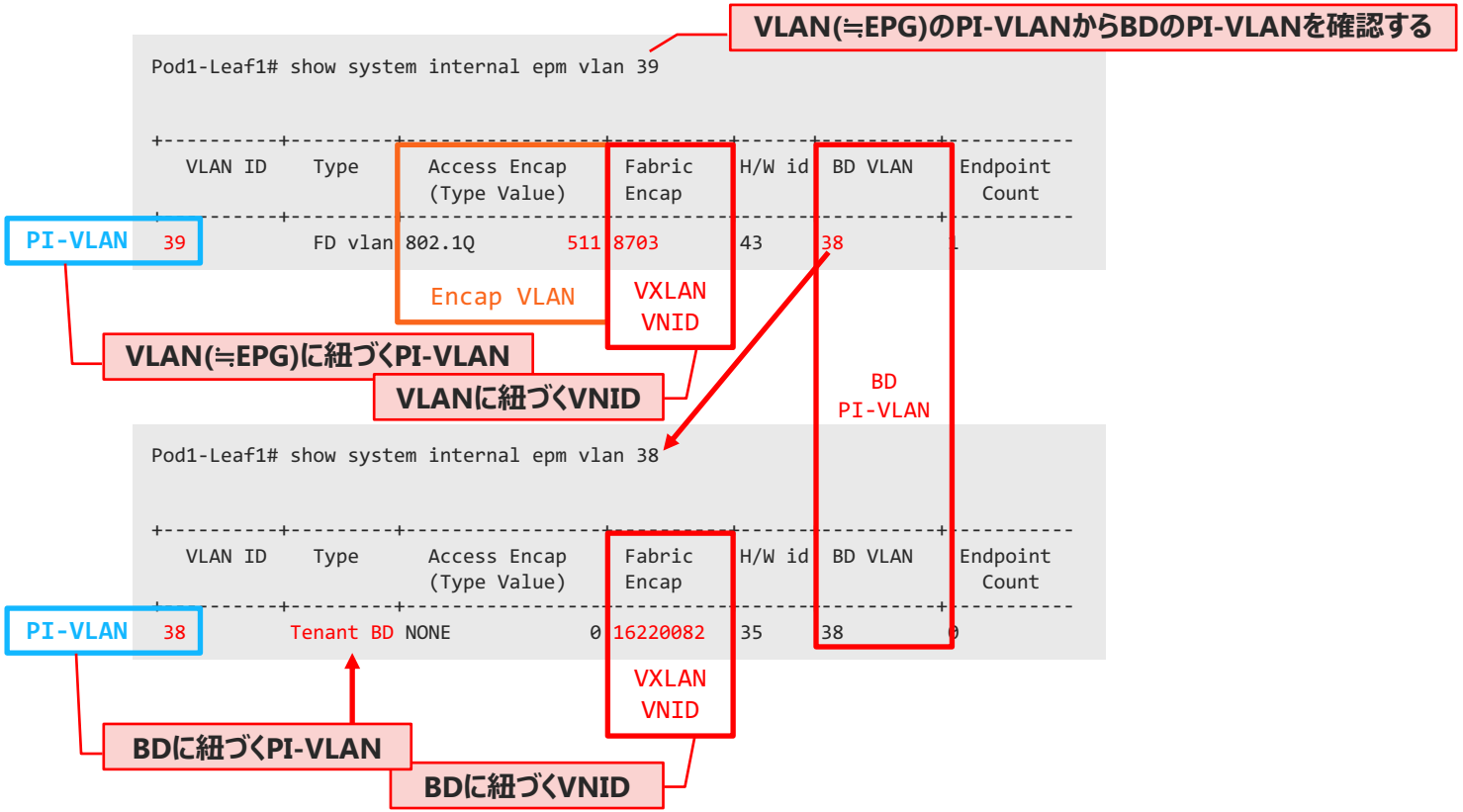
```
Pod1-Leaf1# show vlan id 39 extended

 VLAN Name                             Encap            Ports
 ---- -------------------------------- ---------------- -------------------------
```

PI-VLAN 39 tsetaka_demo1:DEMO_APP1:EPG1 vlan-511 Eth1/34

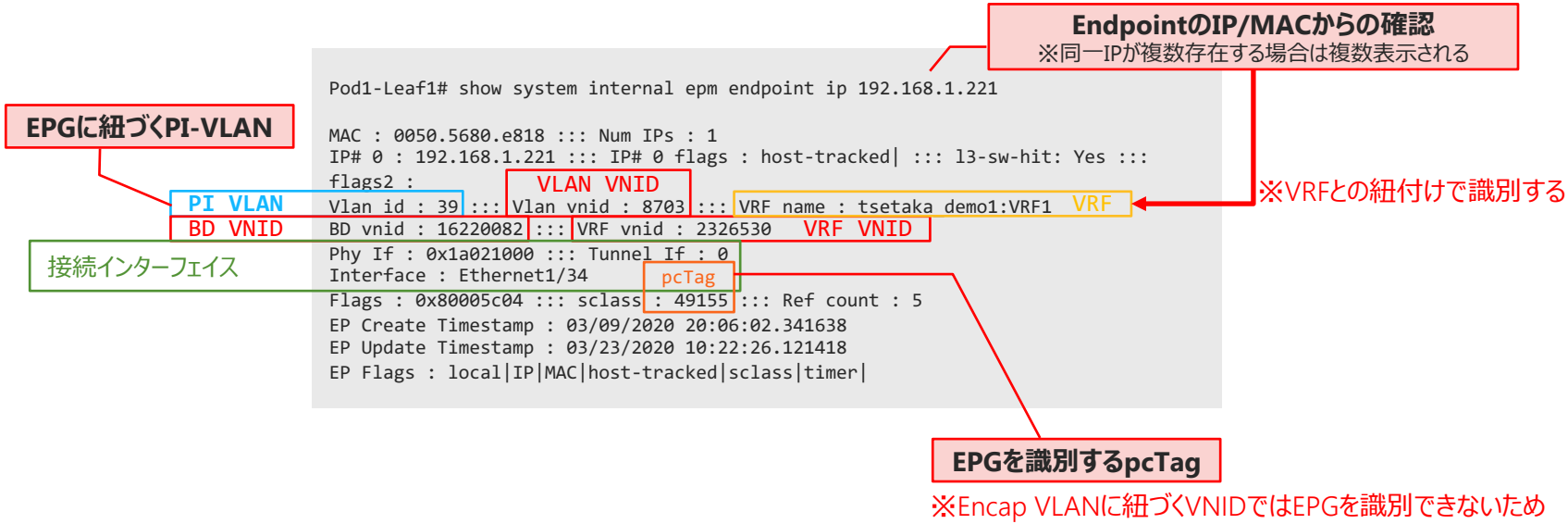
extendオプションを付加すると表示される

※Global Scopeの場合、同一Leaf範囲内で1つのEncap VLANは1つのEPGに対してのみ紐づく。Local Scope設定とした場合、ポート毎にEncap VLANとEPGの紐付けが可能となるため、同一Leaf範囲の同一Encap VLANであっても、同一EPGに紐付かない場合がある。

EPG/BDに紐づく PI-VLAN と VXLAN VNID



Endpointに紐づく情報の一括確認




```
Pod1-Leaf1# show system internal epm endpoint ip 192.168.1.221

MAC : 0050.5680.e818 ::: Num IPs : 1
IP# 0 : 192.168.1.221 ::: IP# 0 flags : host-tracked| ::: l3-sw-hit: Yes :::
flags2 :
Vlan id : 39 ::: Vlan vnid : 8703 ::: VRF name : tsetaka demo1:VRF1
BD vnid : 16220082 ::: VRF vnid : 2326530
Phy If : 0x1a021000 ::: Tunnel If : 0
Interface : Ethernet1/34
Flags : 0x80005c04 ::: sclass : 49155 ::: Ref count : 5
EP Create Timestamp : 03/09/2020 20:06:02.341638
EP Update Timestamp : 03/23/2020 10:22:26.121418
EP Flags : local|IP|MAC|host-tracked|sclass|timer|
```

Tunnelインターフェイスの宛先ノードの確認 (1)

```
Pod1-Leaf1# show endpoint mac 0050.5680.a836
+----------------------------------+---------------+-----------------+--------------+-------------+
      VLAN/                           Encap           MAC Address       MAC Info/       Interface
      Domain                          VLAN            IP Address        IP Info
+----------------------------------+---------------+-----------------+--------------+-------------+
58/tsetaka_demo1:VRF2                vxlan-16646017    0050.5680.a836                    tunnel5
```

Remote Endpoint

```
Pod1-Leaf1# show interface tunnel 5
Tunnel5 is up
    MTU 9000 bytes, BW 0 Kbit
    Transport protocol is in VRF "overlay-1"
    Tunnel protocol/transport is ivxlan
    Tunnel source 10.0.120.64/32 (lo0)
    Tunnel destination 10.0.88.65
    Last clearing of "show interface" counters never
    Tx
    0 packets output, 1 minute output rate 0 packets/sec
    Rx
    0 packets input, 1 minute input rate 0 packets/sec
```

宛先TEPアドレス

Tunnelインターフェイスの宛先ノードの確認 (2)

```
Pod1-Leaf1# acidiag fnvread
      ID   Pod ID                 Name     Serial Number         IP Address    Role        State   LastUpdMsgId
-------------------------------------------------------------------------------------------------------------
     101        1           Pod1-Leaf1      FDO21470WDK     10.0.120.64/32    leaf         active   0
     102        1           Pod1-Leaf2      FDO21470WAD     10.0.120.66/32    leaf         active   0
     103        1           Pod1-Leaf3      FDO20510MBQ     10.0.120.67/32    leaf         active   0
     104        1           Pod1-Leaf4      FDO2053177E     10.0.120.68/32    leaf         active   0
     115        1          Pod1-Leaf15      FDO212809T7      10.0.80.64/32    leaf         active   0
     151        1             Pod1-RL1      FDO212809T2     10.70.0.144/32    leaf         active   0
    1001        1          Pod1-Spine1      FDO21372ZXC     10.0.120.65/32   spine         active   0
    1002        1          Pod1-Spine2      SAL1811NN46     10.0.120.69/32   spine         active   0
```

個別ノードのTEPアドレス

```
apic1# moquery -c vpcDom | egrep 'virtualIp|dn|#'
# vpc.Dom
dn                  : topology/pod-1/node-101/sys/vpc/inst/dom-10
virtualIp           : 10.0.88.64/32
# vpc.Dom
dn                  : topology/pod-1/node-102/sys/vpc/inst/dom-10
virtualIp           : 10.0.88.64/32
# vpc.Dom
dn                  : topology/pod-1/node-103/sys/vpc/inst/dom-20
virtualIp           : 10.0.88.65/32
# vpc.Dom
dn                  : topology/pod-1/node-104/sys/vpc/inst/dom-20
virtualIp           : 10.0.88.65/32
```

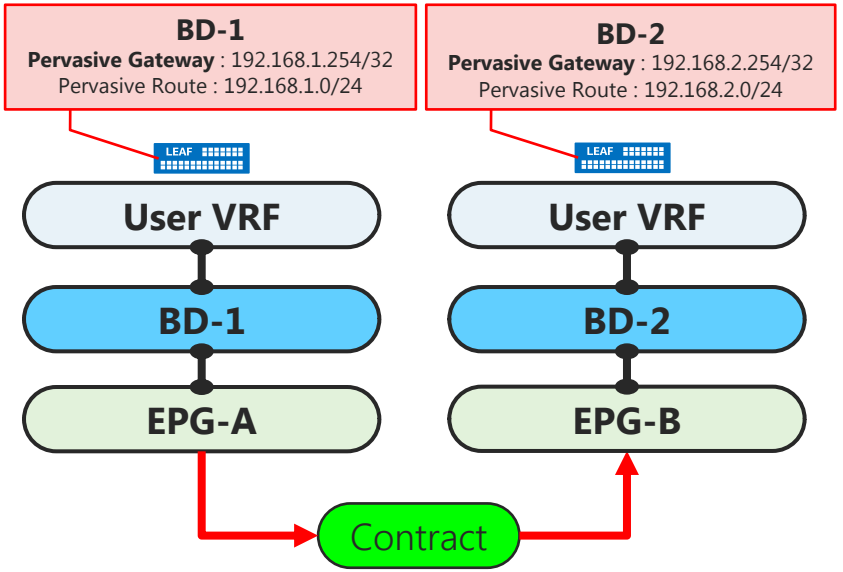
VPCに紐づくTEPアドレス

ACI転送関連パラメータ

Pervasive Gateway
Pervasive Route

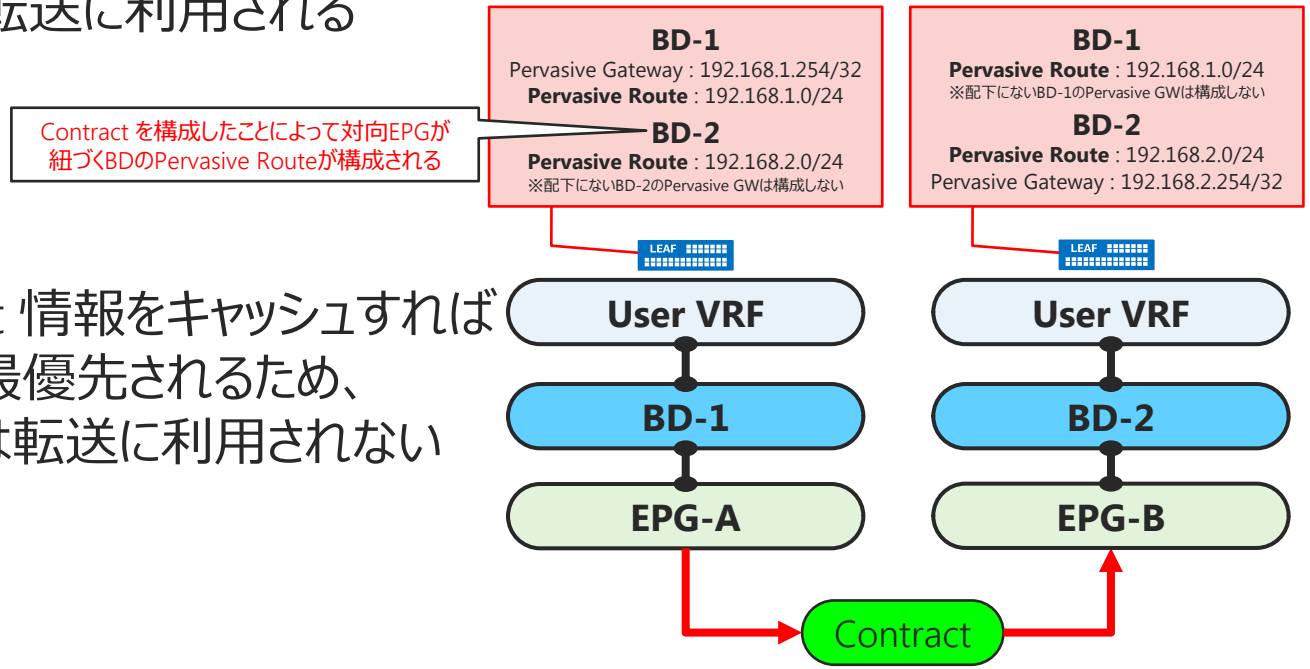
Pervasive Gateway (BD SVI)

- ACI Fabric に接続した Endpoint に対する Default Gateway
- BDに紐づくSubnet Prefixを示す Pervasive Route と組み合わせて構成される
- BDに紐づくEndpointが存在するLeafで
  BD PI-VLAN に紐づく SVI として
  構成される
- 複数Subnetを同一BDに構成した場合、
  Primary IP となる1つを除いて
  Secondary IP として構成される
  ※どのPervasive GatewayをPrimaryとするかは
  指定することが可能



Pervasive Route

- VRFに紐付いて ACI Fabric 内のBDに紐づく Prefix 範囲であることを示す
- Spine Proxy への転送に利用される経路情報



- Remote Endpoint 情報をキャッシュすればホストルートとして最優先されるため、Pervasive Routeは転送に利用されない

Pervasive Gateway と Pervasive Route

BD-SVI

BDに紐づくPI-VLAN

BDが紐づくVRF

```
Pod1-Leaf1# show ip interface vlan 38
IP Interface Status for VRF "tsetaka_demo1:VRF1"
vlan38, Interface status: protocol-up/link-up/admin-up, iod: 105, mode: pervasive
  IP address: 192.168.1.254, IP subnet: 192.168.1.0/24
  IP broadcast address: 255.255.255.255
  IP primary address route-preference: 0, tag: 0
```

BDに構成されたPervasive Gateway

VRF RIB

```
Pod1-Leaf1# show ip route vrf tsetaka_demo1:VRF1
IP Route Table for VRF "tsetaka_demo1:VRF1"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

192.168.1.0/24, ubest/mbest: 1/0, attached, direct, pervasive
    *via 10.0.8.66%overlay-1, [1/0], 01w06d, static, tag 4294967294
192.168.1.254/32, ubest/mbest: 1/0, attached, pervasive
    *via 192.168.1.254, vlan38, [0/0], 01w06d, local, local
192.168.2.0/24, ubest/mbest: 1/0, attached, direct, pervasive
    *via 10.0.8.66%overlay-1, [1/0], 01w06d, static, tag 4294967294
```

BDに構成された Pervasive Gateway (BD SVI)

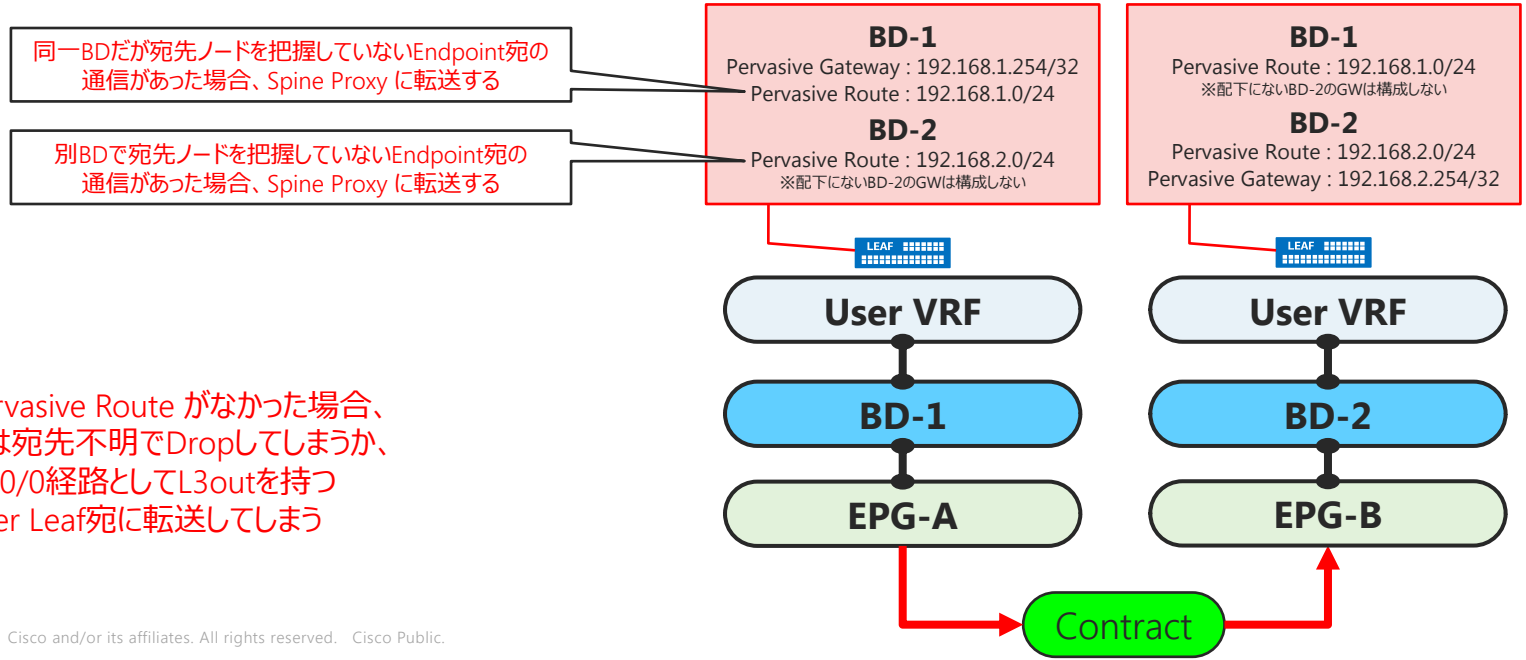
Pervasive Route (VRF範囲に紐づくBD Subnet)

Pervasive Route (Contractを結んだ相手EPGのBD Subnet)

※このLeafにはBDが存在しない = BD SVIは構成されない

Pervasive Route と Spine Proxy

- Spine Proxyに転送する判断をするためには、Pervasive Route 情報が必要
- Remote Endpoint 学習後はホストルートなので優先される (TEP間直接転送)



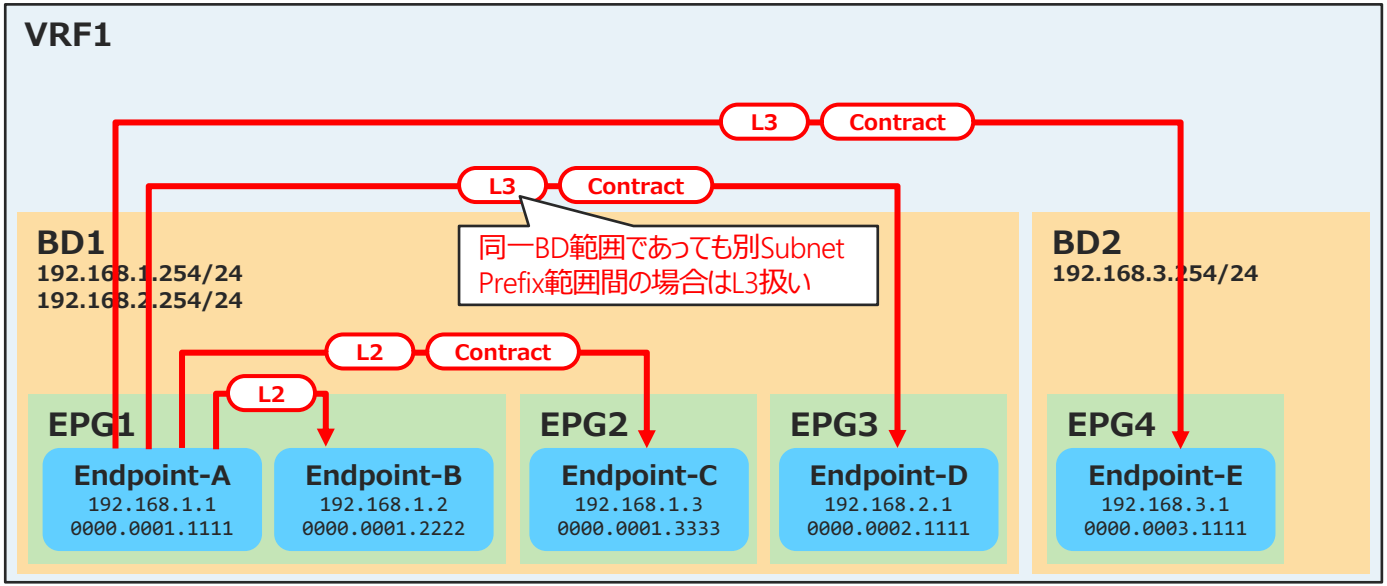
※Pervasive Route がなかった場合、
Leafは宛先不明でDropしてしまうか、
0.0.0.0/0経路としてL3outを持つ
Border Leaf宛に転送してしまう

転送スコープ(BD/VRF)

L2転送 と L3転送

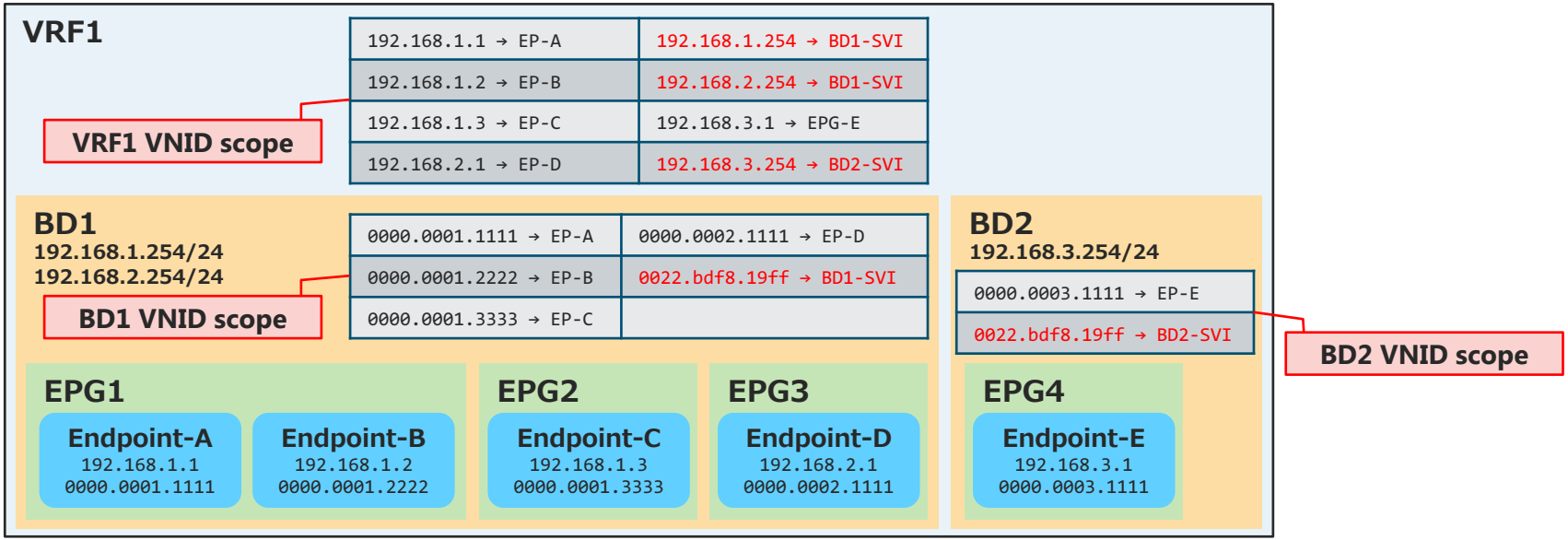
宛先によってBDスコープ(L2)が用いられるか、VRFスコープ(L3)が用いられるかが決まる。

※EPG間通信(もしくはIntra-EPG Contract構成の場合のEPG内通信)の場合は、BD/VRFスコープに基づく転送処理に加えてContractのチェックが行われて最終的な通信可否が判定される。



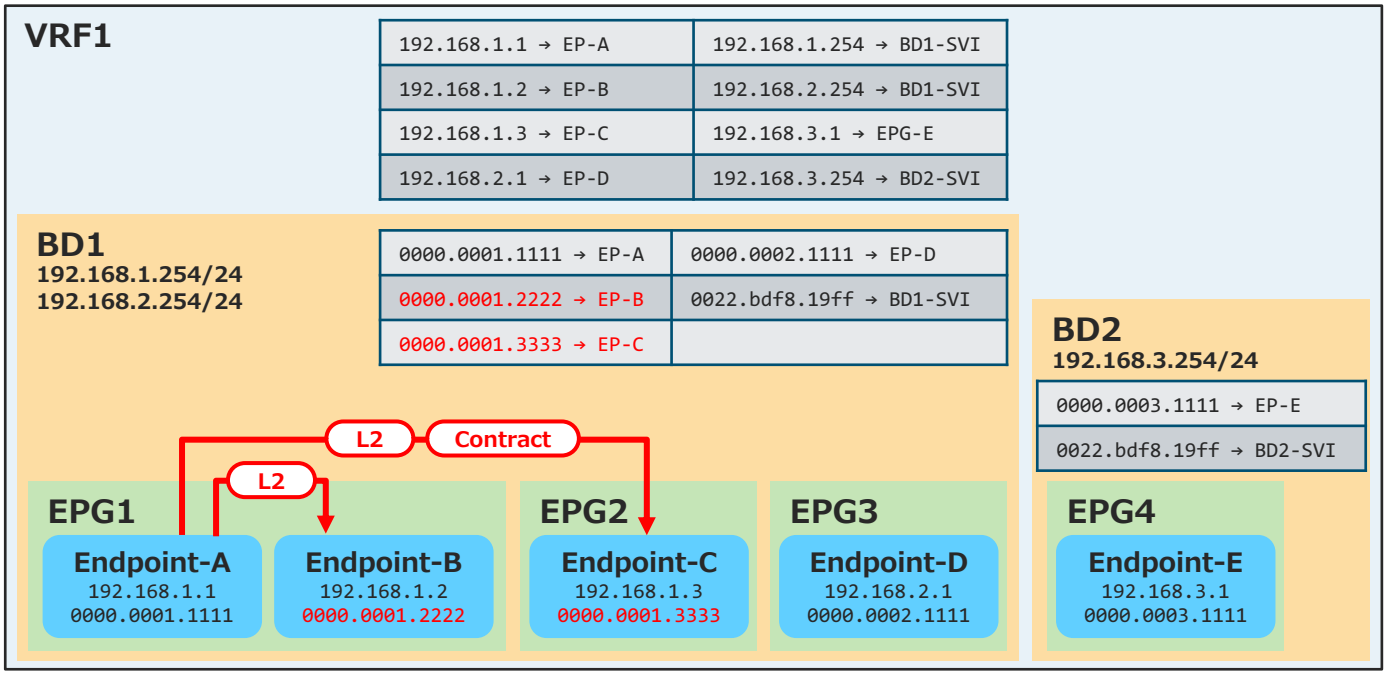
BDスコープ と VRFスコープ

VRFおよびBD毎に用意されるVNID毎に転送スコープが構成され、L2通信の場合はBDのスコープ(MAC)、L3通信の場合はVRFのスコープ(IP)に基づいて転送処理される。



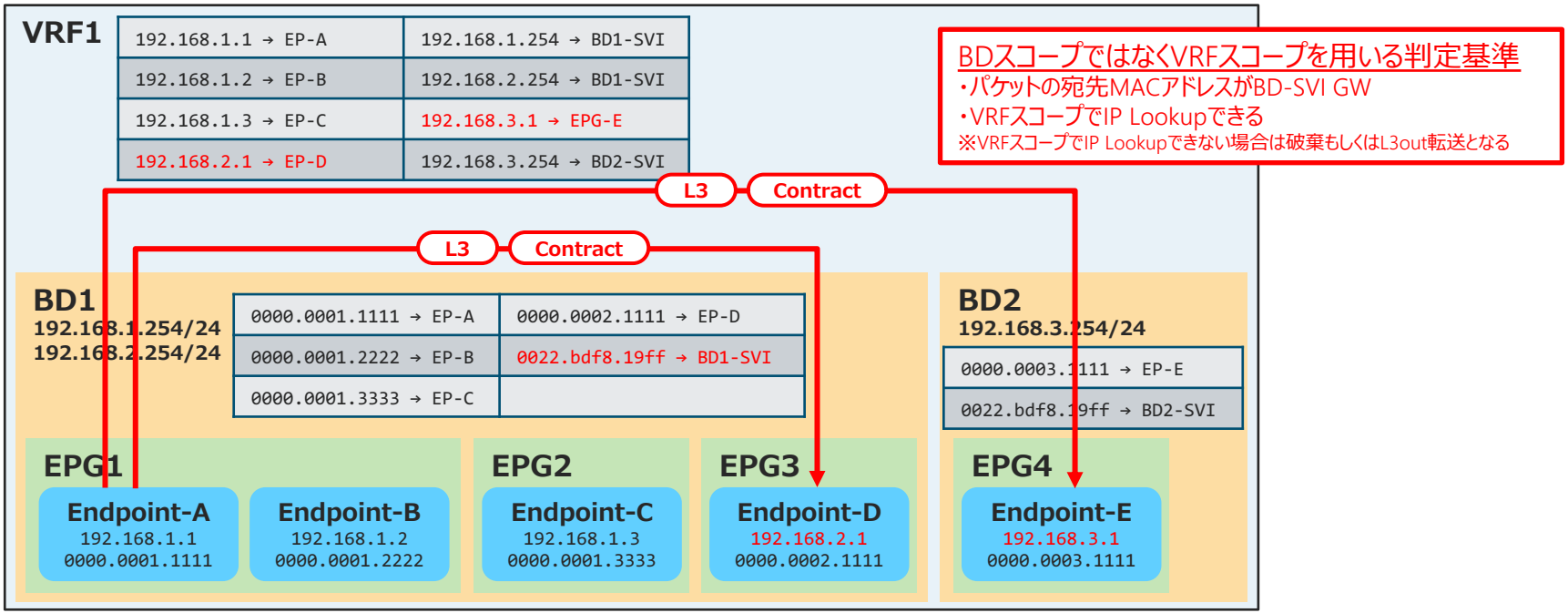
L2 : BDスコープ

EP-A から EP-B や EP-C に通信する場合には、BDスコープに基づいて宛先確認が行われ、転送処理される(EPG間通信の場合はContract判定も別途行われる)。



L3：VRFスコープ

EP-A から EP-D や EP-E に通信する場合には、VRFスコープに基づいて宛先確認が行われ、転送処理される(EPG間通信の場合はContract判定も別途行われる)。



BDパラメータ

転送動作に関連する BD パラメータ設定

BD > Policy > General	
L2 Unknown Unicast	**Flood**
	Hardware Proxy (Default)
L3 Unknown Multicast Flooding ***IPv6 L3 Unknown Multicast***	**Flood** (Default)
	Optimized Flood
Multi Destination Flooding	**Flood in BD** (Default)
	Drop
	Flood in Encapsulation
ARP Flooding	GUIでBDを作成した場合はデフォルト有効(ACI 4.2～)、API/CLI作成の場合はデフォルト無効(互換性維持のため) ([L2 Unknown Unicast] : Flood 時は必ず有効となる)

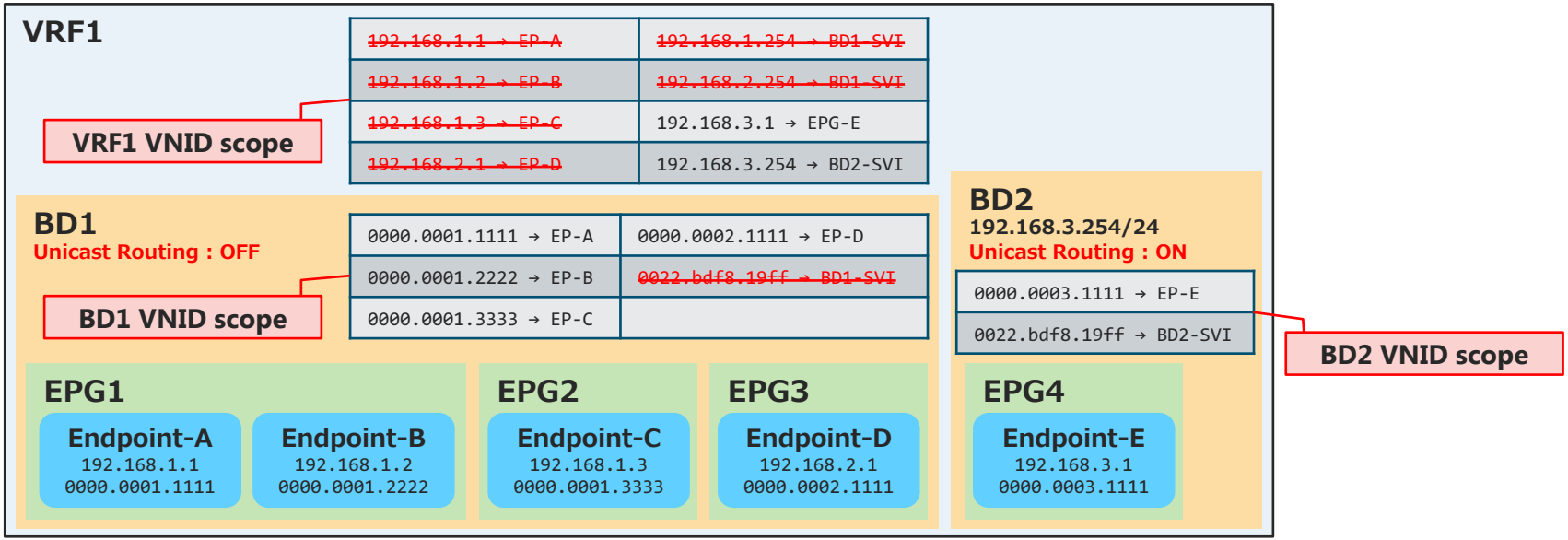
※詳細後述

BD > Policy > L3 Configurations	
Unicast Routing	・有効の場合、BD-SVI に設定した IPアドレスが BD範囲の Default GW として利用される ・無効の場合、BD範囲で IP学習 が無効化され、BD-SVIが無効となるためBDはL2のみの動作となる
Custom MAC Address	・BD-SVI に紐づく MACアドレス (デフォルトでは、全BD共通で **"0022.bdf8.19ff"** が用いられる) ・既存環境からGWを移行する際などで、必要に応じて MACアドレスを引き継ぐために適切に設定する

Unicast Routing : OFF

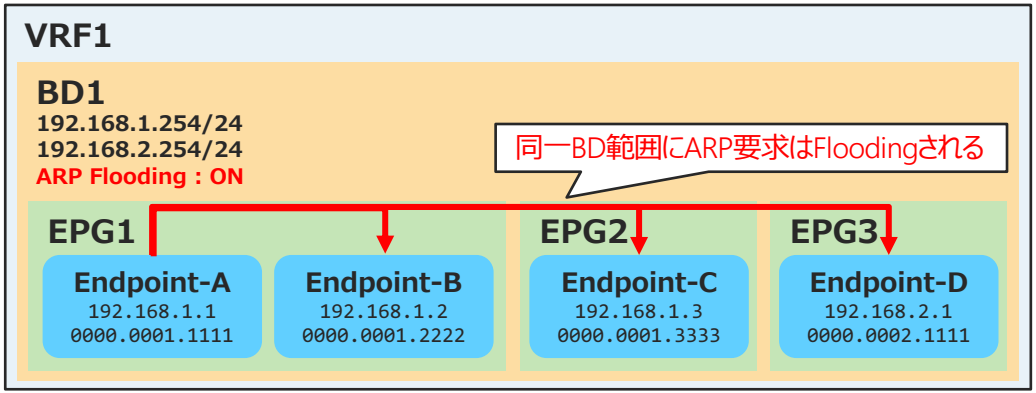
- BDスコープからはBD-SVIに紐づくMACアドレスのエントリが削除される
- VRFスコープからは当該BDに紐づく全てのIPアドレスのエントリが削除される

→ MACアドレスに基づくL2範囲のEndpoint間でのみ通信ができる状態となる



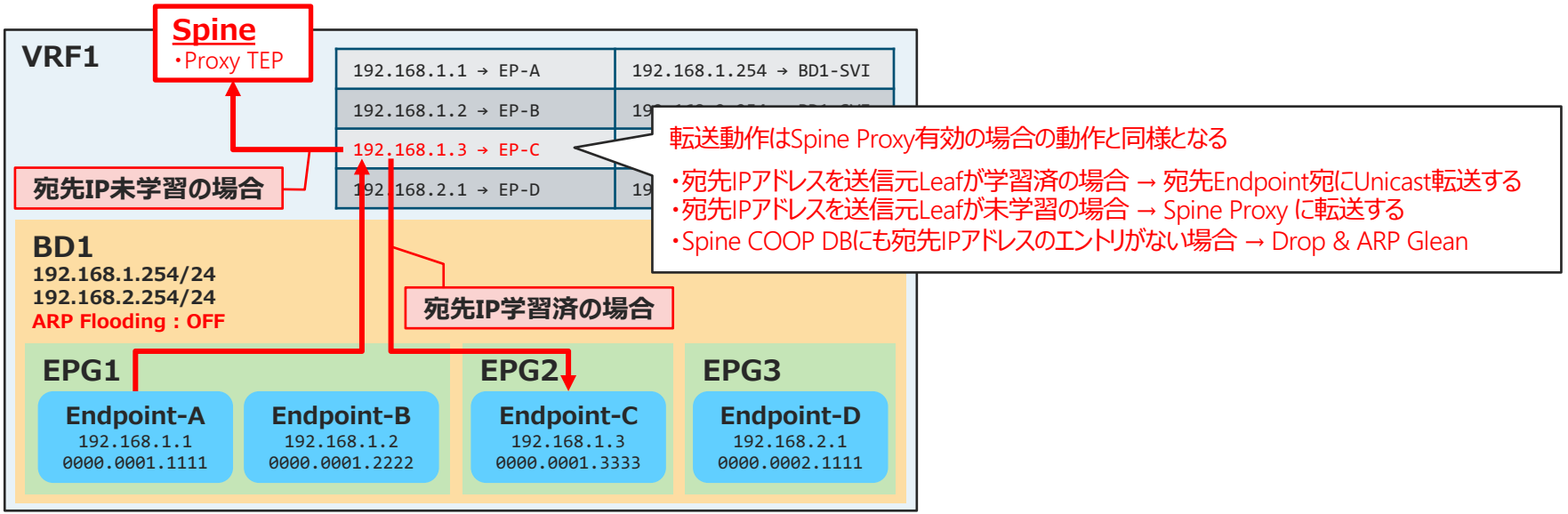
ARP Flooding : ON

- ARP Flooding 有効 = 従来のスイッチと同等の動作となる
- 宛先MACアドレスが ffff.ffff.ffff の場合、BD範囲全体に Broadcast Flooding する
  (Leaf範囲だけではなくACI FabricのBD範囲全体にFloodingされる)
- Unicast Routing 無効の場合は、自動的に ARP Flooding は有効になる ※非明示的
  (ARP Flooding : OFF ≒ L3 Unicast となり Unicast Routing が必要なため)



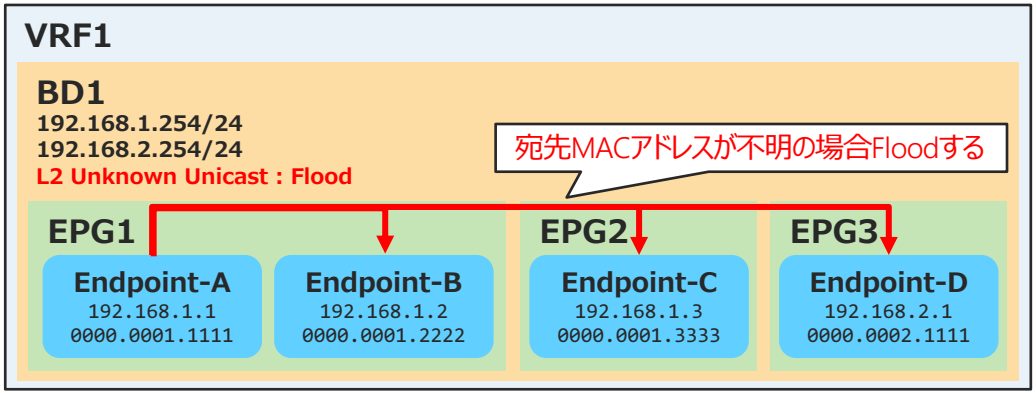
ARP Flooding : OFF

- ARP Flooding 無効 = ARP要求を L3 Unicast として処理する
(宛先IPアドレスを参照して転送する)
- ARPは Contract によってはフィルタされない



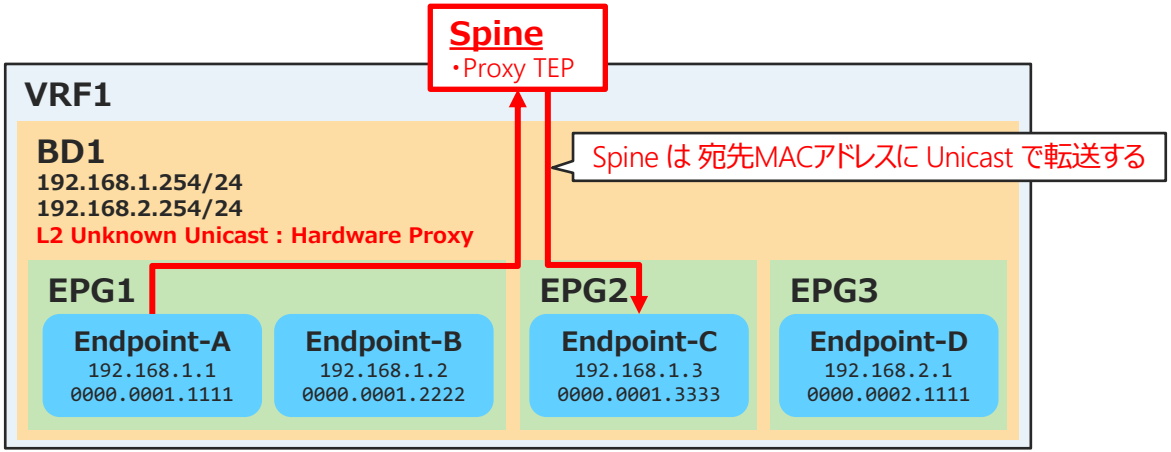
L2 Unknown Unicast : Flood

- L2 Unknown Unicast : Flood = 従来のスイッチと同等の動作となる
  ※L2 Unknown Unicast : Flood 設定に変更した場合、自動的に ARP Flooding も有効になる
- Unicast Routing 無効の場合には、Flood 設定とすることを推奨
- L2サイレントホストが存在する場合は L2 Unknown Unicast : Flood 設定が推奨
  ※L3の場合はARP Glean機能で対処できる



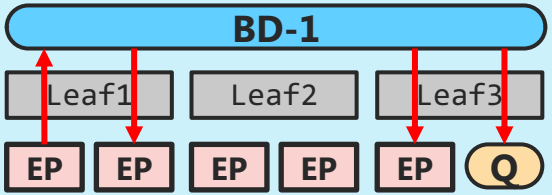
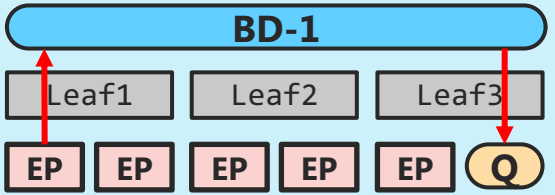
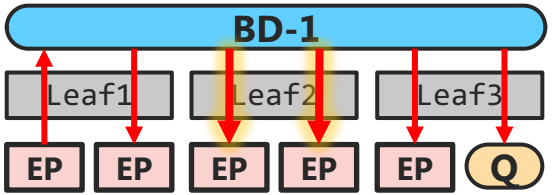
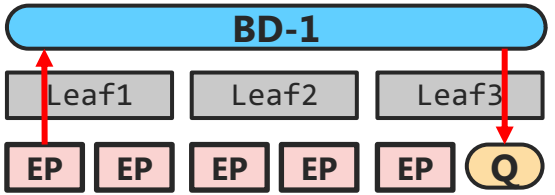
L2 Unknown Unicast : Hardware Proxy

- L2 Unknown Unicast : Hardware Proxy = Spine Proxy による宛先解決
- 送信元Leafが宛先MACアドレスを未学習の場合にSpine Proxyに転送する
- Spine の COOP DB に 宛先MACアドレスのエントリが存在しない場合はDropする



L3 Unknown Multicast Flooding

- IGMP Snooping によって確認されていない未知の IP Multicast グループに対する Flooding 処理方法の指定

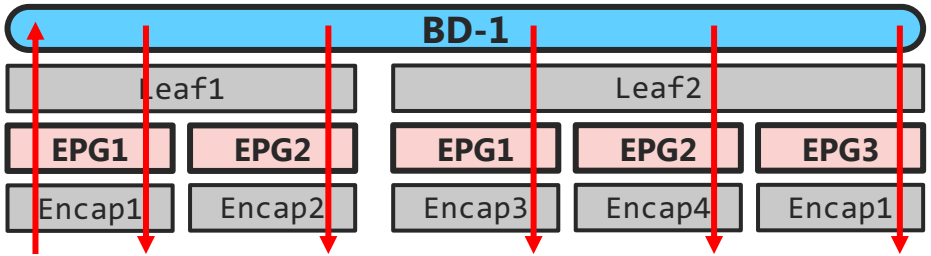
L3 Unknown Multicast Flooding	Flood	Optimized Flood
1st Gen Leaf	受信LeafおよびQuerierのあるLeafにのみFloodする 	Querierポートに対してのみFloodする 
2nd Gen Leaf	全LeafのBD範囲全ポートにFloodする 	Querierポートに対してのみFloodする 

Multi Destination Flooding

- L2 Multicast, Broadcast, Link-local に対する Flooding 動作の指定
  (OSPF, BGP, EIGRP, CDP, LACP, LLDP, ISIS, IGMP, PIM, ST-BPDU, ARP/GARP, RARP, NDを除く)

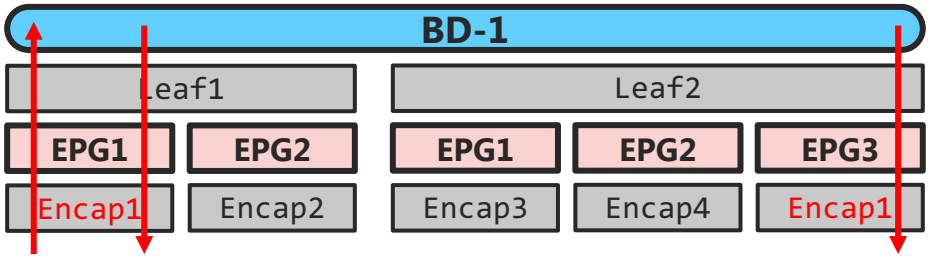
Flood in BD

※EPG, VLANに関係なく同一BD範囲にFlood

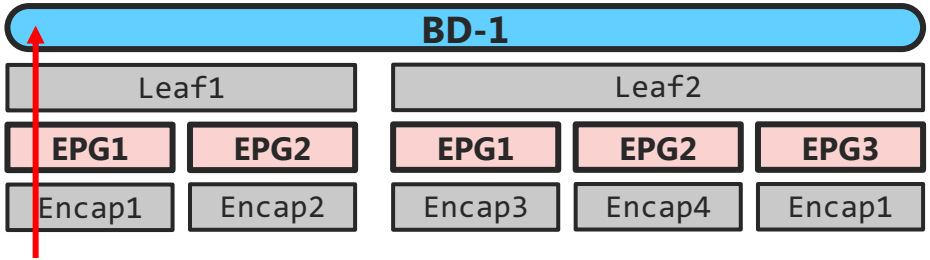


Flood in Encapsulation

※EPGに関係なく同一BDかつ同一VLAN範囲にFlood (詳細次項参照)



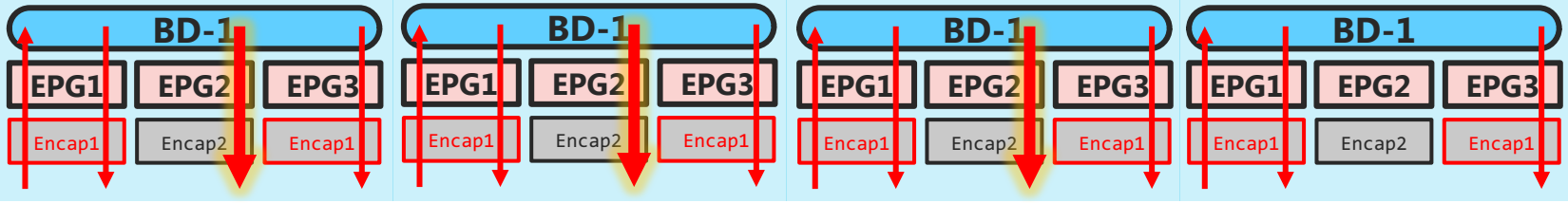
Drop



Flood in Encapsulation 組み合わせ動作

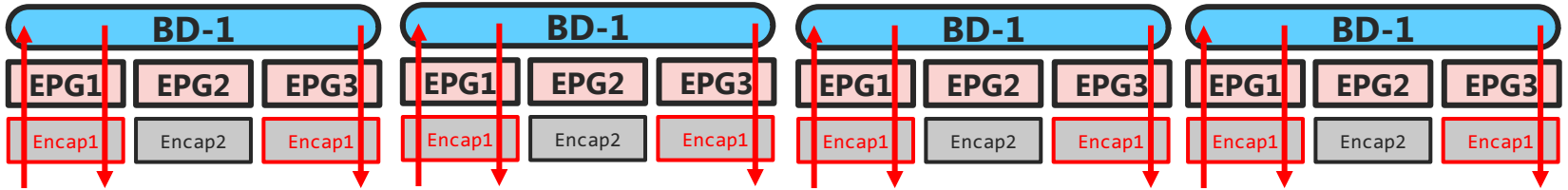
- **BD** → [Multi Destination Flooding] : Flood in Encapsulation
- **EPG** → [Flood on Encapsulation] : Enabled

ACI Version	[L2 Unknown Unicast] : Flood	[ARP Flooding] : ☑Enabled	[Unknown L3 Multicast Flood] : Flood	その他 Multi-Destination 通信
Gen1 *Gen2 ～3.0*	※Flood in Encap 無効	※Flood in Encap 無効	※Flood in Encap 無効	



ACI 3.1以降 & Gen2 Leaf 利用の場合、全てのMulti-Destination Flood動作で “Flood in Encap” 設定が有効となる

Gen2 3.1～



Spine Proxy

Anycast TEP

- 全Spineが共有して保持する Spine Proxy 用 TEPアドレス

```
Pod1-Leaf1# show ip route vrf tsetaka_demo1:VRF1
IP Route Table for VRF "tsetaka_demo1:VRF1"

192.168.1.0/24, ubest/mbest: 1/0, attached, direct, pervasive
    *via 10.0.8.66%overlay-1, [1/0], 01w06d, static, tag 4294967294
192.168.1.254/32, ubest/mbest: 1/0, attached, pervasive
    *via 192.168.1.254, vlan38, [0/0], 01w06d, local, local
192.168.2.0/24, ubest/mbest: 1/0, attached, direct, pervasive
    *via 10.0.8.66%overlay-1, [1/0], 01w06d, static, tag 4294967294
```

Pervasive Route の Next Hop は必ず Spine Proxy TEP 宛となる

```
Pod1-Leaf1# show isis dteps vrf overlay-1

IS-IS Dynamic Tunnel End Point (DTEP) database:
DTEP-Address       Role    Encapsulation   Type
10.0.80.64         LEAF    N/A             PHYSICAL
10.0.120.65        SPINE   N/A             PHYSICAL
10.0.8.65          SPINE   N/A             PHYSICAL,PROXY-ACAST-MAC
10.0.8.66          SPINE   N/A             PHYSICAL,PROXY-ACAST-V4
10.0.8.64          SPINE   N/A             PHYSICAL,PROXY-ACAST-V6
10.0.88.64         LEAF    N/A             PHYSICAL
10.0.120.66        LEAF    N/A             PHYSICAL
10.0.88.65         LEAF    N/A             PHYSICAL
```

[L2 Unknown Unicast] : Hardware Proxy 設定時の転送先

L2通信用Spine Proxy TEPアドレス

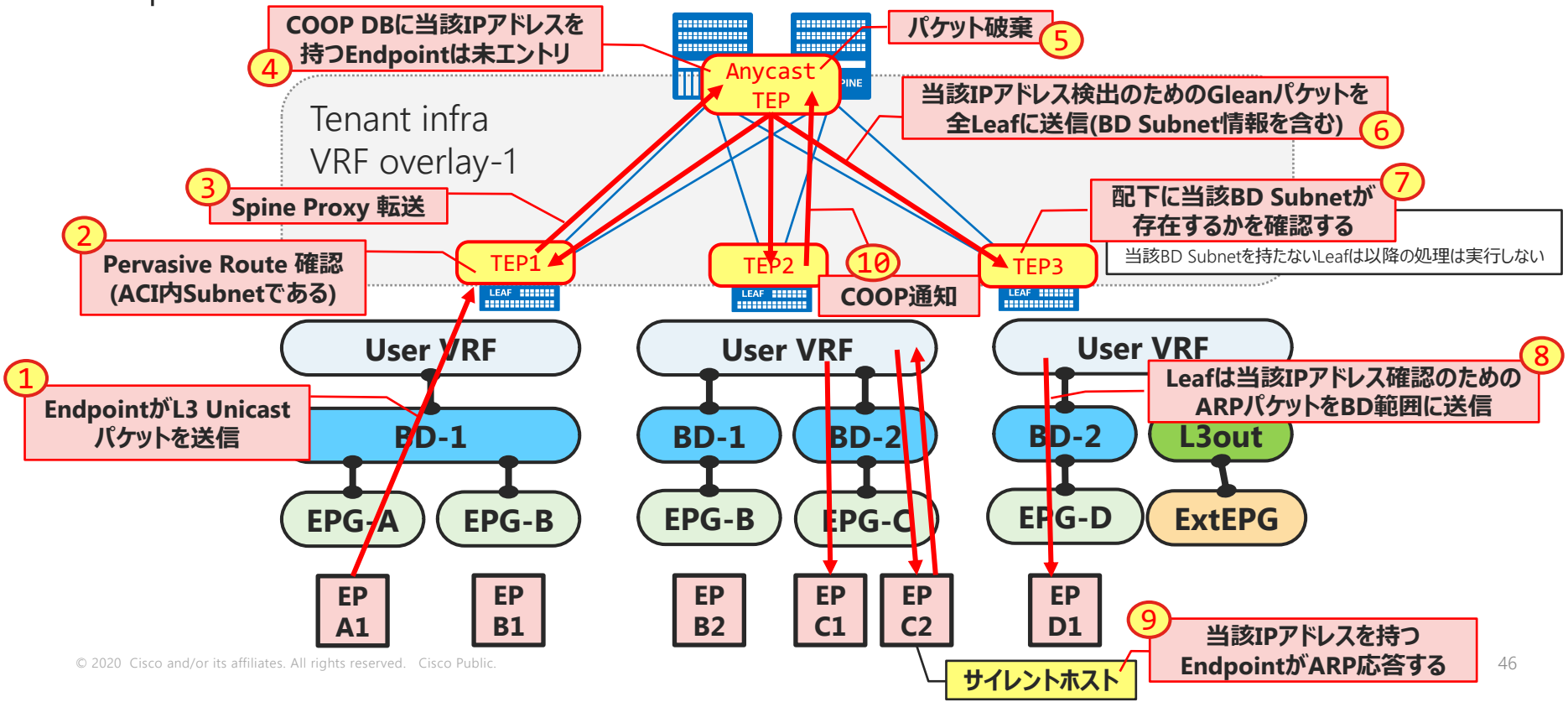
L3(IPv4)通信用Spine Proxy TEPアドレス

L3(IPv6)通信用Spine Proxy TEPアドレス

宛先TEP不明時の転送先 および [ARP Flooding] : OFF 時のARP要求転送先

ARP Glean

- Spine COOP 未エントリ L3 サイレントホスト検出のための仕組み



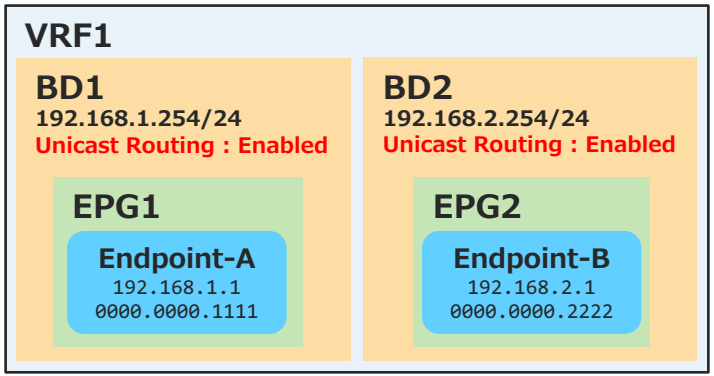
Packet Walk

Packet Walk : 論理トポロジー・物理トポロジー

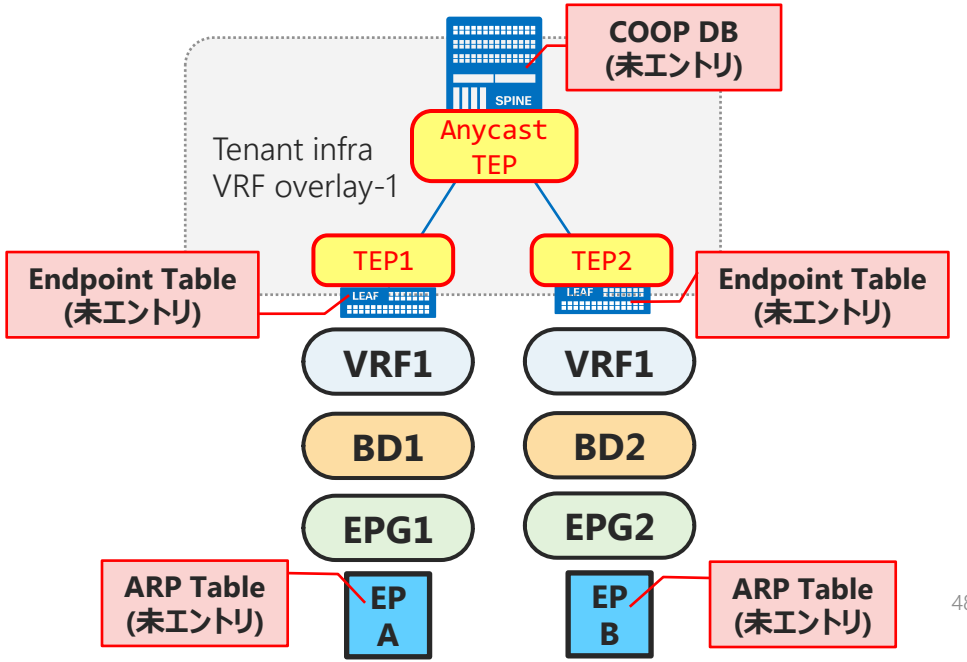
ACI Fabric (COOP, Endpoint Table), Endpoint (ARP)等は完全に未学習ステータス

- Endpoint-A から Endpoint-B に ICMP (Ping要求) を送信し ICMP (Ping応答) が返されるまでの流れ

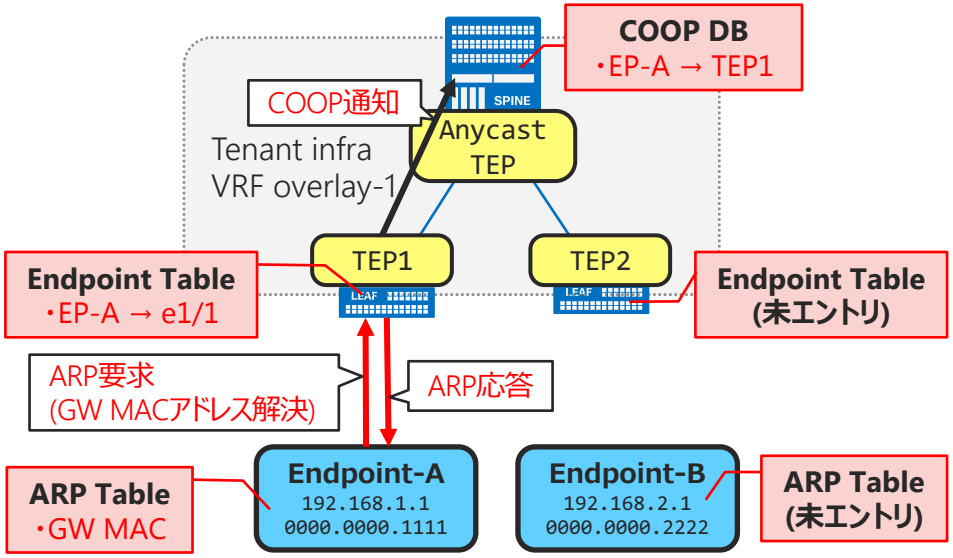
論理トポロジー



物理トポロジー

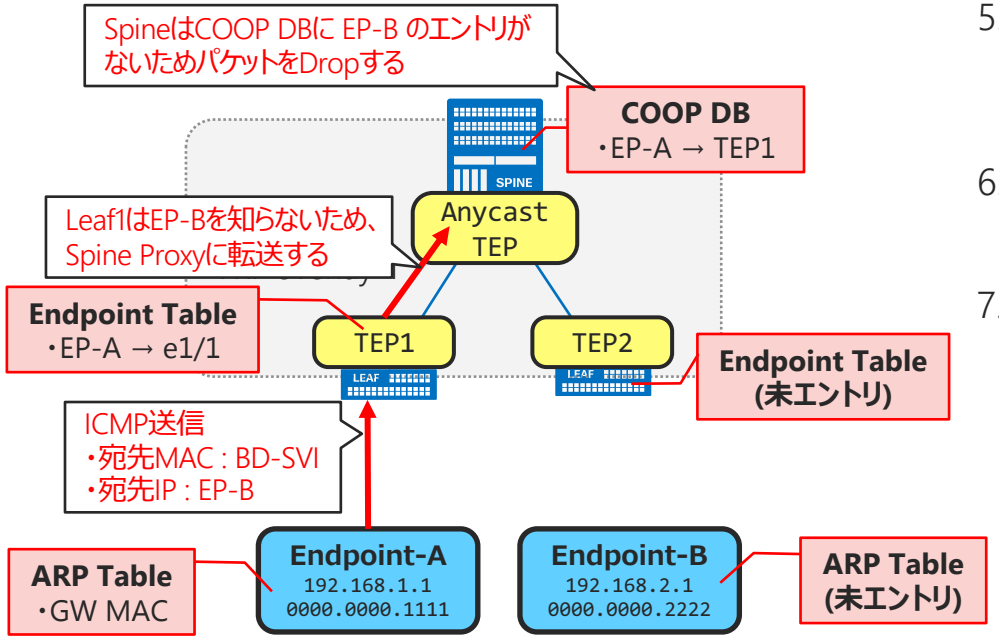


Packet Walk (1) EP-1 ARPによるGW MAC解決



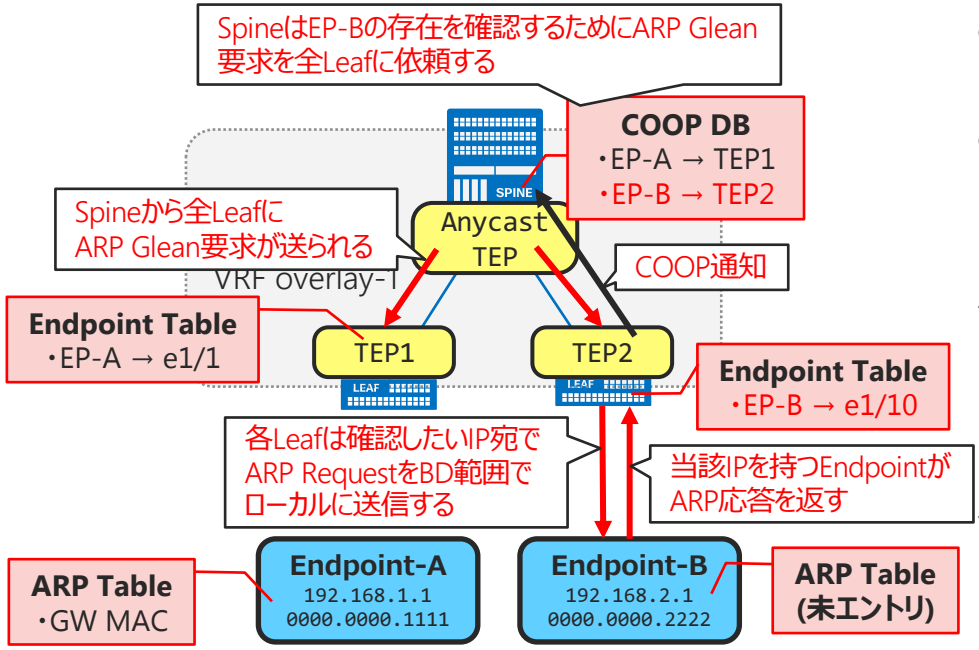
1. EP-Aは別Subnetに存在するEP-Bへ通信を行うために必要となるGWのMACアドレスを確認するためにARP要求をGWアドレス宛に送信する
2. Leaf1はEP-Aの存在を学習(送信元MACアドレス : 0000.0000.1111、送信元IPアドレス : 192.168.1.1)しEndpoint TableにLocal EndpointとしてEP-Aをエントリ、ARPに応答する
3. EP-AはDefault GWとなるBD-SVIに紐づくMACアドレスを学習する
4. Leaf1はSpineに対してEP-Aの情報をCOOPで通知し、SpineはEP-AをLeaf1(TEP1)に紐づけてCOOP DBにエントリする

Packet Walk (2) EP-B宛ICMPの送信



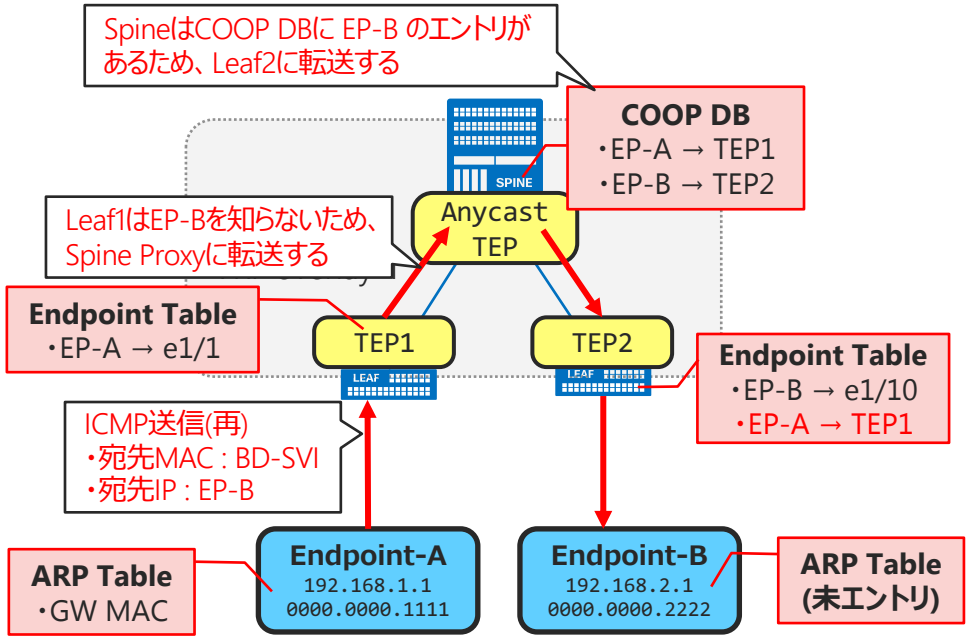
5. EP-AはGW MAC (BD-SVI)を送信先MACアドレスとして宛先EP-B IPアドレス(192.168.2.1)宛のICMPパケットを送信する
6. Leaf1はEP-Bを把握していないため、Spine Proxy宛に転送する
7. SpineはEP-Bの情報がCOOP DBに存在しないため、パケットをドロップ(破棄)する

Packet Walk (3) ARP Gleanによる存在確認



8. SpineはEP-B(192.168.2.1)の存在を確認するためにARP Glean要求を全Leafに対して依頼する
9. 当該BDを持つLeafはARP Glean要求に対して、当該BD範囲に対して対象IP(192.168.2.1)宛のARP要求をローカルに送信する
10. 対象IPを持つEndpointがARP応答を返すことで、Leaf2はEP-Bの存在を学習 (送信元MACアドレス : 0000.0000.2222、送信元IPアドレス : 192.168.2.1) し、Endpoint TableにEP-Bをエントリする
11. Leaf2はSpineに対してEP-Bの情報をCOOPで通知し、SpineはEP-BをLeaf2に紐づけてCOOP DBにエントリする

Packet Walk (4) EP-B宛ICMPの再送信



12. EP-Aは再度GW MAC (BD-SVI)を送信先MACアドレスとして宛先EP-B IPアドレス(192.168.2.1)宛のICMPパケットを送信する

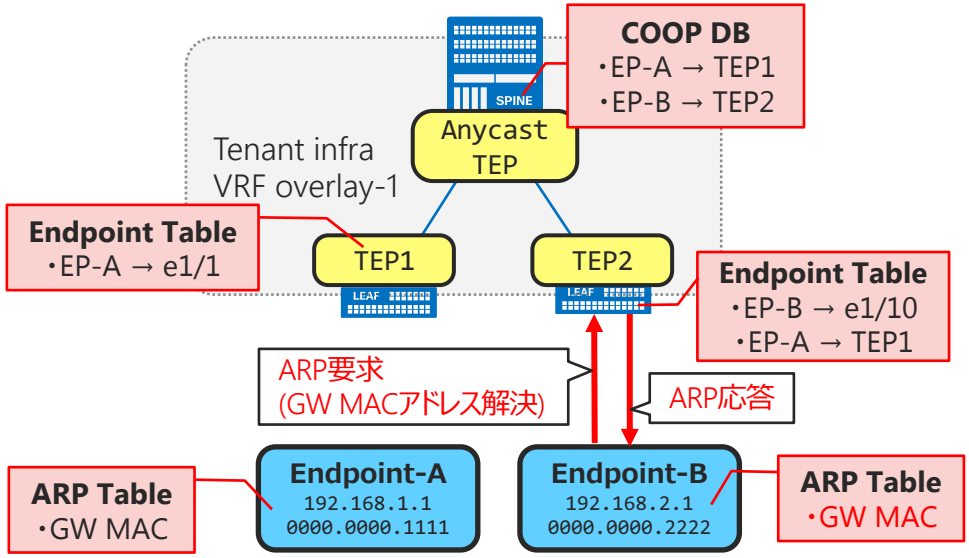
13. Leaf1はEP-Bを把握していないため、Spine Proxy宛に転送する

14. SpineはCOOP DBにエントリされているEP-Bの情報を確認し、EP-Bが紐付けられているLeaf2(TEP2)宛に転送する

15. Leaf2はEP-AをRemote EndpointとしてTEP1に紐づけてキャッシュとして学習する
(Spine Proxyは送信元TEPアドレスを書き換えないため)

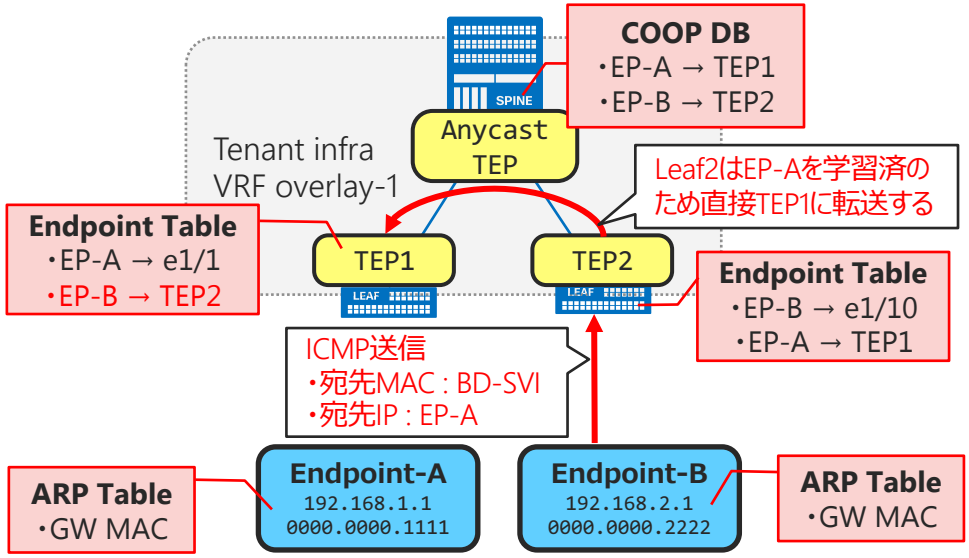
16. EP-BにEP-AからのICMPパケットが届く

Packet Walk (5) EP-B ARPによるGW MAC解決



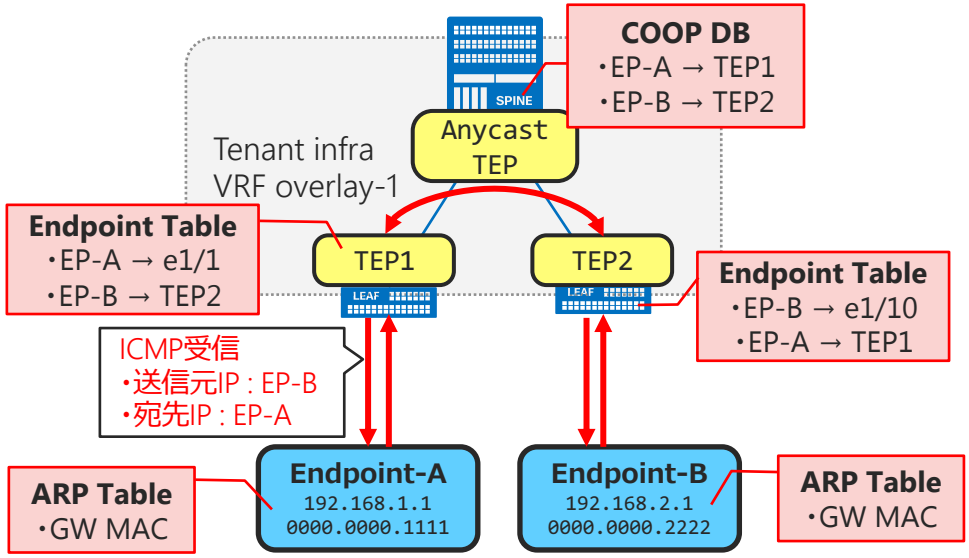
17. EP-BはEP-AからのICMP要求に応答するために、まずEP-A宛の通信に必要なGWのMACアドレスを確認するためにARP要求をGWアドレス宛に送信する
18. Leaf2はARP応答を返す(EP-Bは既に学習済)
19. EP-BはDefault GWとなるBD-SVIに紐づくMACアドレスを学習する

Packet Walk (6) EP-A宛ICMP送信(応答)



20. EP-BはEP-AからのICMP要求に応答するパケットを送信する
21. Leaf2(TEP2)はEP-Aが紐づくLeaf1(TEP1)を学習済のため、直接TEP1宛に転送する
22. Leaf1(TEP1)はパケットを受信した時点で、EP-BがLeaf2(TEP2)に紐づくことを学習しEndpoint Tableにキャッシュとしてエントリする

Packet Walk (7) EP-A宛ICMP受信

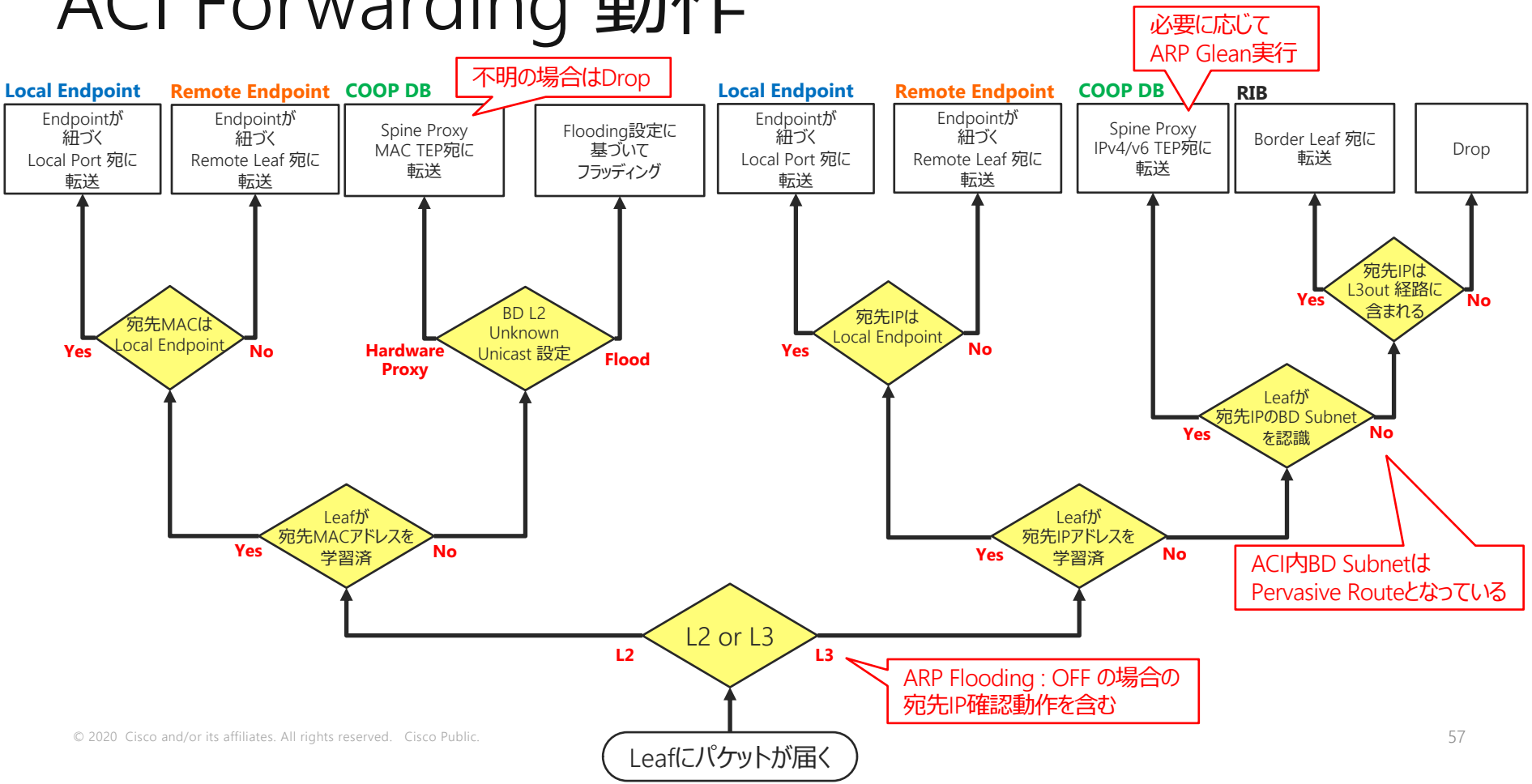


23. EP-AはEP-BからのICMP応答を受信する

24. 以降のEP-A・EP-B間通信はLeaf1(TEP1)とLeaf2(TEP2)のいずれも宛先Endpoint情報を学習済となるため、Leaf間で相互に転送を行う(Spine Proxyは利用しない)

まとめ

ACI Forwarding 動作




CISCO