



ACI Endpoint Learning Deep Dive

Cisco Systems Inc.

2020年 4月

Agenda

1. Endpoint Learning **要点** : 最適化パラメータ、Gen 2 Leaf 動作、ベストプラクティス
2. Endpoint **基礎** : Endpoint Table, Local / Remote Endpoint, Spine Proxy
3. Endpoint Table と COOP DB : COOP DB, Endpoint Table, Verified Scalability
4. Forwarding **基礎** : L2転送, L3転送, L3out転送
5. Endpoint **学習動作** : Local Endpoint学習, Remote Endpoint学習, L3out通信時 + **Issue Case #1-4**
6. Endpoint **移動への対応** : Bounce Entry, Retention Timer + **Issue Case #5-7**
7. Endpoint **サイレントホスト対応** : Unknown MAC, Unknown IP, ARP Flooding
8. Endpoint Learning **パラメータ詳細**

參考資料

- 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-2641 ACI Troubleshooting – Endpoints

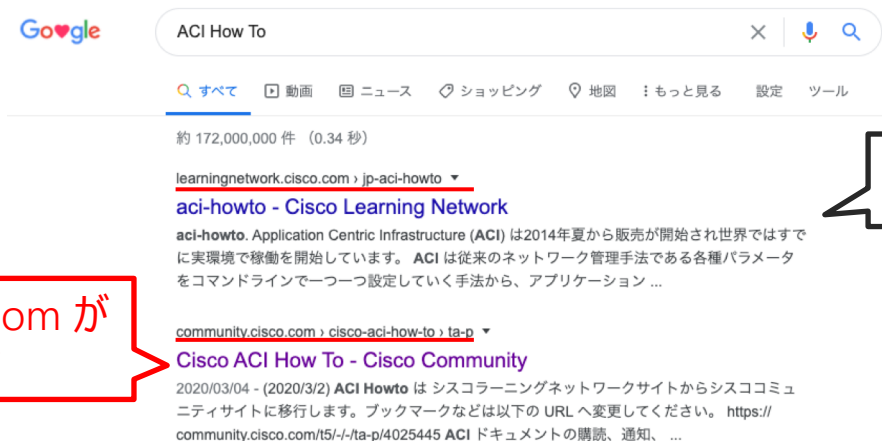
<https://www.ciscolive.com/c/dam/r/ciscolive/emea/docs/2020/pdf/BRKACI-2641.pdf>

ACI How To は移行しました!!

ACI How To は、learningnetwork から community に移行しました。

<https://community.cisco.com/t5/%E3%83%87%E3%83%BC%E3%82%BF%E3%82%BB%E3%83%B3%E3%82%BF%E3%83%BC-%E3%83%89%E3%82%AD%E3%83%A5%E3%83%A1%E3%83%B3%E3%83%88/cisco-aci-how-to/ta-p/4039933>

旧ページの資料は当面維持されますが、システム更新に伴い全ページのURLが変更されています。また、新しい情報については community 側でのみ更新されますのでご注意ください!!



learningnetwork.cisco.com は
古い方です

community.cisco.com が
新しい方です

Endpoint Learning 要点

ACI Endpoint Learning 最適化パラメータ

| | パラメータ [関連スライド番号] | 設定箇所 | 実装 バージョン | 用途 | 考慮点 |
|----------------|---|---|--------------------|---|---|
| EPG ↑ BD | L4-L7 Virtual IPs [74-74] | Tenant > Application Profiles > EPG / uSeg EPG | 1.2(1m) | DSR動作のための仮想IPアドレスに対するData-Plane Learningを無効化する | 用途としてDSR対応以外では正式にはサポートされない(開発側として未検証) |
| | Unicast Routing [27, 67, 77] | Tenant > Networking > Bridge Domains | 1.0(1e) | 当該BDでL3 Routing動作をするためにEndpointのIPアドレスを学習する | Disableにすると当該BD範囲でIPアドレスは学習しない |
| | EP Move Detection Mode (GARP based detection) [67, 78] | | 1.1(1j) | GARP情報からIPの移動を検出することで、同一Interface配下でのMACアドレス間での特定IPアドレスの移行に対応する | Unicast RoutingおよびARP Floodingを必ず有効化する必要がある (Gen 2 Leafでは対処不要) |
| | Limit IP Learning To Subnet [7, 36, 44-46, 79] | | 1.1(1j) | BDに構成したSubnet範囲外のIPアドレスはLocal Endpointとして学習しない | Remote Endpointとしての学習を防止することはできない |
| | IP Data-plane Learning (BD) [6, 51, 80] | | 2.0(1m) | 当該BDでData-planeからのIPアドレスの学習を無効化する (PBR用のL4-7デバイス対応) | PBRを利用するService Graphのための機能 |
| VRF | IP Data-plane Learning (VRF) [6, 49-50, 82, 91] | Tenant > Networking > VRFs | 4.0(1h) | VRF範囲でData-planeからのIPアドレスの学習を無効化する | |
| Fabric ↑ | Disable Remote EP Learning (on Border Leaf) [6, 7, 36, 45, 63-64, 84] | System > System Settings > Fabric-Wide Settings | 2.2(2e) 3.0(1k) | Ingress Policy Enforcementが構成されているVRFに紐づくL3outを持つBorder LeafでRemote EndpointのIPアドレス学習を無効化する | Gen 2 Leafの場合はL3 Multicastからは学習する (Gen 1 LeafはL3 Multicastをサポートしない) |
| | Enforce Subnet Check [7, 34, 36, 44, 47, 60, 85] | | 2.2(2q) 3.0(2h) | VRFに紐づくBDに構成されているSubnet範囲外のIPアドレスをLocal / Remote Endpointとして学習しない | Gen 2 Leaf に対してのみ有効 ※2.3および3.0(1)では利用できない |
| | IP Aging [7, 42, 49-50, 88-89] | System > System Settings > Endpoint Controls | 2.1(1h) | Endpoint情報としてMACアドレスに紐付けて学習されているIPアドレスについて個別にAgingによりタイムアウト管理する | ACI 2.1(1h)以降ではデフォルト有効 |
| | Rouge EP Control [39, 87] | | 3.2(1l) | 一定以上のEndpointの頻繁な移動に対する負荷軽減 | 問題解決には直接はつながらない |
| | EP Loop Protection [40, 86] | | 1.0(1e) | 一定以上のEndpointの頻繁な移動をLoopとみなす | 対象以外にも影響を与える場合がある |

ACI Endpoint Data-plane Learning 動作

| 構成 | | | Data-plane 学習動作 | | | | |
|------------------------------|---------------------------------|----------------------------------|-----------------|-----------------------|---------------------|------------------------|---------------------------------|
| IP Data-plane Learning (VRF) | IP Data-plane Learning (BD) | Disable Remote EP Learn (Global) | Local MAC | Local IP | Remote MAC | Remote IP (Unicast) | Remote IP (Multicast) |
| Disabled | Enable/Disable どちらでも動作は変化しない | Enable/Disable どちらでも動作は変化しない | 学習する | 学習しない ARP/NDでは学習する | 学習する ※Gen1は学習しない | 学習しない | 学習する ※Gen1はL3 Multicast非サポート |
| Enabled (Default) | Disabled | Enable/Disable どちらでも動作は変化しない | 学習する | 学習しない ARP/NDでは学習する | 学習しない | 学習しない | 学習しない |
| Enabled (Default) | Enabled (Default) | Disabled | 学習する | 学習する | 学習する | Border Leafのみ 学習しない | 学習する |
| Enabled (Default) | Enabled (Default) | Enabled (Default) | 学習する | 学習する | 学習する | 学習する | 学習する |

Gen 2 モデルとは、型番の最後が -EX, -FX, -FX2, -GX などハイフンの後にXを含むものを指す。
-E はGen 1.5であり、広義にはGen 1 に含まれる。

ACI Endpoint Learning : ベストプラクティス

| パラメータ | Gen 1 only | Gen 2 only | Gen 1 & Gen 2 mix |
|--|---|---|---|
| Limit IP Learning To Subnet | Enabled ※2.3(1e)および3.0(1k)以降でデフォルト有効 | 構成しなくてもよい ※ただし2.3(1e)および3.0(1k)以降でデフォルト有効 | Enabled ※2.3(1e)および3.0(1k)以降でデフォルト有効 |
| Disable Remote EP Learning (on Border Leaf) | ACI 3.2(2l)より前 かつ Policy Control Enforcement : Ingress の場合 Enabled ※必要に応じて手動での学習済Remote Endpointのクリアを実施 ACI 3.2(2l)以降の場合 構成不要 (下記参照) | ACI 3.2(2l)より前 かつ Policy Control Enforcement : Ingress の場合 Enabled ※必要に応じて手動での学習済Remote Endpointのクリアを実施 ACI 3.2(2l)以降の場合 構成不要 (下記参照) | ACI 3.2(2l)より前 かつ Policy Control Enforcement : Ingress の場合 Enabled ※必要に応じて手動での学習済Remote Endpointのクリアを実施 ACI 3.2(2l)以降の場合 構成不要 (下記参照) |
| Enforce Subnet Check | 無効 | Enabled | Enabled ※効果があるのはGen 2のみ |
| IP Aging | Enabled ※2.1(1h)以降でデフォルト有効 | Enabled ※2.1(1h)以降でデフォルト有効 | Enabled ※2.1(1h)以降でデフォルト有効 |

ACI 3.2(2l)以降では[EP Announce on Bounce Delete]機能が自動的に利用可能となるため、意図しないRemote Endpoint情報がエントリされたままとなる問題を防止できる(CSCvj17665)。

Endpoint 基礎

Endpoint とは ACI 内部の接続端末情報

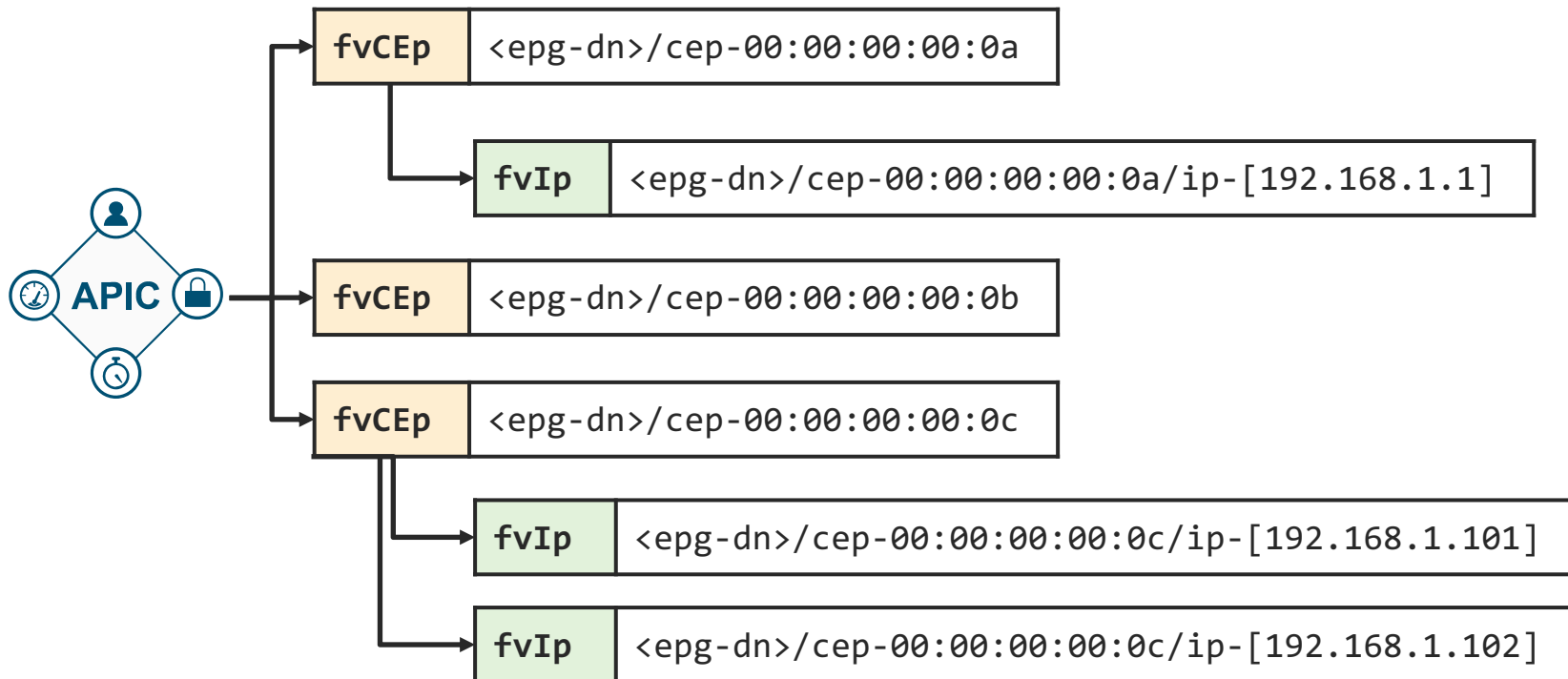
従来のネットワーク機器におけるMACアドレスとIPアドレスの学習動作とは異なり、ACIではホストルートとしてMACアドレステーブルとARPテーブルではなく**Endpointテーブル**を利用する(外部接続を除く)。ホストルート以外はRIBが合わせて参照される。

| 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

APIC における Endpoint = Object

MACアドレス Object と、それに紐づく IPアドレス Object



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。

Endpoint学習動作の違い

Local Endpointは学習したDownlinkポートとEncap VLANに紐付いてData-planeとControl-plane (ARP/ND)を通じて学習される。

Remote EndpointはVXLANの対向VNIDとTunnelインターフェイスに紐付いてData-planeを通じて学習される。

LeafスイッチはData-planeとControl-planeを通じてEndpointを学習する。

→ **Endpoint Table**

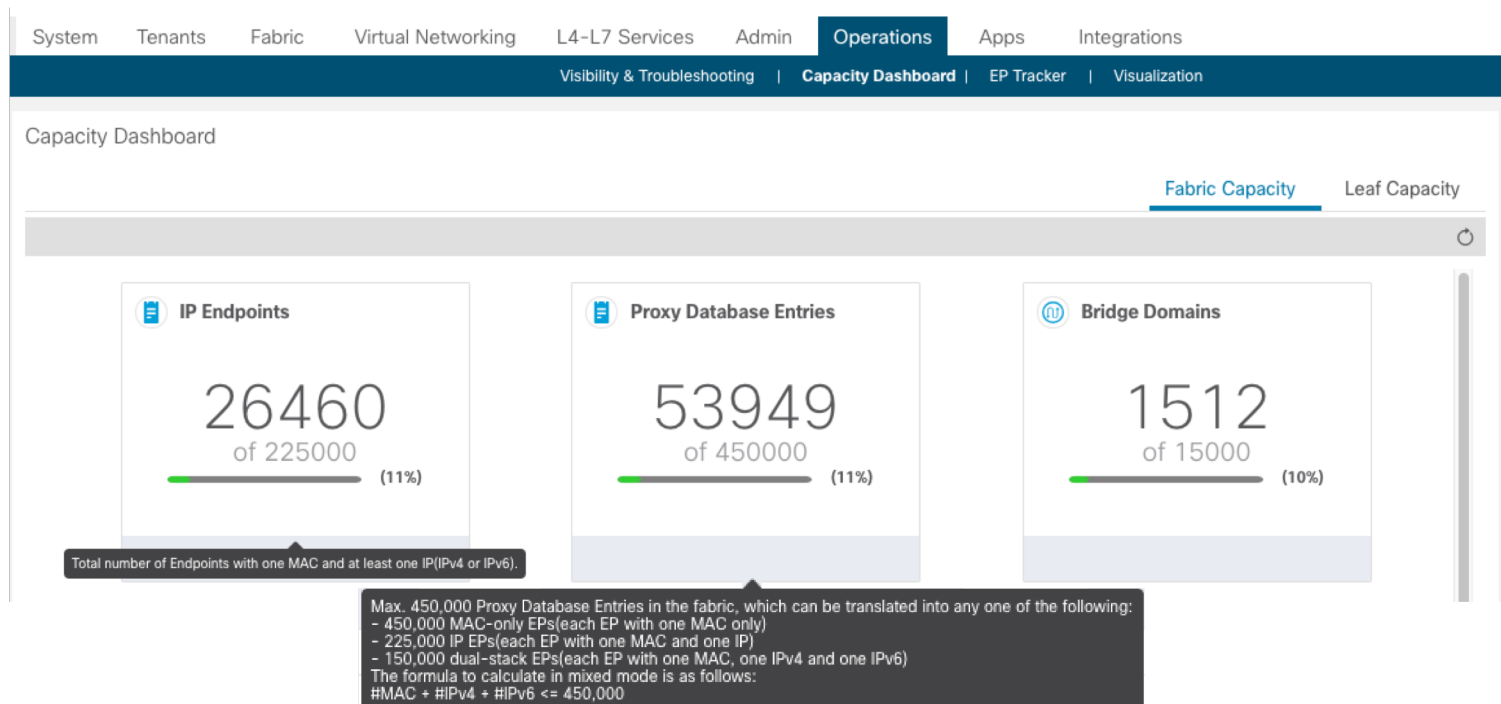
SpineスイッチはCOOPを通じてのみEndpointを学習する。

→ **COOP DB**

Endpoint Table と COOP DB

APIC : Capacity Dashboard

APIC [Operations] > [Capacity Dashboard] の数字はどう計算されている？

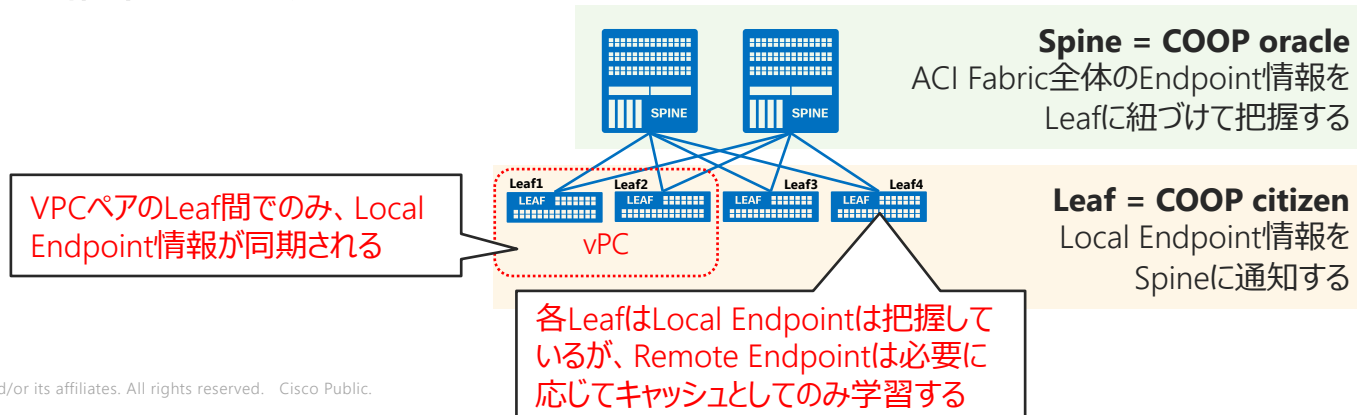


COOP (Council of Oracles Protocol) DB

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。

- COOP citizen である Leaf スイッチは、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)
- COOP oracle である Spine スイッチは、COOP 情報を相互に同期する

ただし、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。



COOP DB エントリ

```
Pod1-Spine1# show coop internal info repo ep key 16220082 0050.5680.e818
```

```
EP bd vnid : 16220082
EP mac : 00:50:56:80:E8:18
flags : 0x80
repo flags : 0x122
Vrf vnid : 2326530
Epg vnid : 0
EVPN Seq no : 0
Remote publish timestamp: 01 01 1970 09:00:00 0
Snapshot timestamp: 03 23 2020 10:03:43 337833079
Tunnel nh : 10.0.120.64
MAC Tunnel : 10.0.120.64
IPv4 Tunnel : 10.0.120.64
IPv6 Tunnel : 10.0.120.64
ETEP Tunnel : 0.0.0.0
num of active ipv4 addresses : 1
num of anycast ipv4 addresses : 0
num of ipv4 addresses : 1
Primary Path:
Current published TEP : 10.0.120.64
Current citizen (publisher_id): 10.0.120.64
Previous citizen : 10.0.120.64
Prev to Previous citizen : 10.0.120.64
Real IPv4 EP : 192.168.1.221
Current publisher_id: 10.0.120.64
MAC Tunnel : 10.0.120.64
IPv4 Tunnel : 10.0.120.64
```

BD VNID

※出力を一部省略しています

```
Pod1-Spine1# show coop internal info ip-db key 2326530 192.168.1.221
```

```
IP address : 192.168.1.221
Vrf : 2326530
Flags : 0
EP bd vnid : 16220082
EP mac : 00:50:56:80:E8:18
Publisher Id : 10.0.120.64
Record timestamp : 03 23 2020 10:03:43 337833079
Publish timestamp : 03 23 2020 10:03:43 340608835
Seq No: 0
Remote publish timestamp: 01 01 1970 09:00:00 0
URIB Tunnel Info
Num tunnels : 1
    Tunnel address : 10.0.120.64
    Tunnel ref count : 1
```

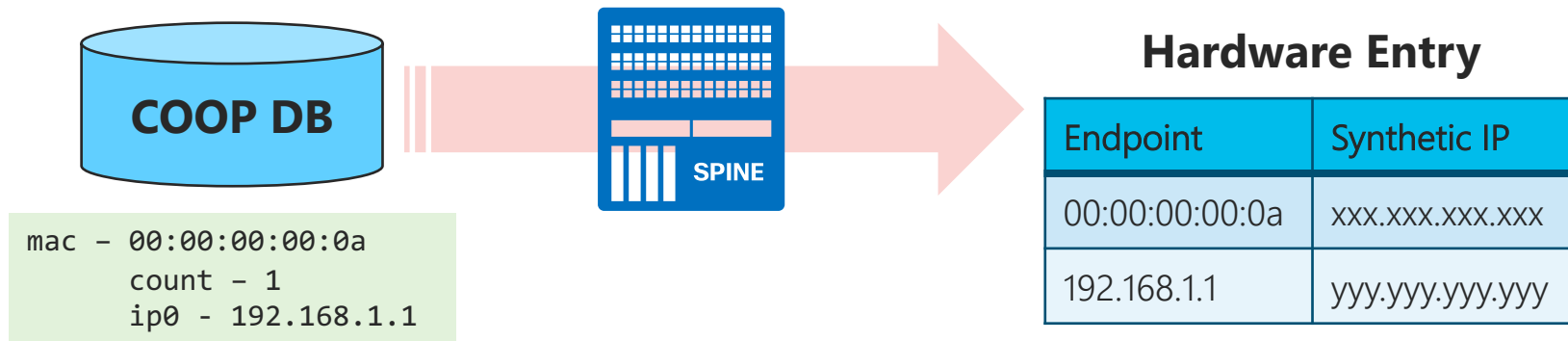
VRF VNID

```
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|
```

Spine における Endpoint = COOP DB → HW

COOP DB のエントリは Spine ハードウェアエントリとなり ハードウェア処理される



Remote Endpoint 学習と Spine Proxy

各Leafスイッチは自身が送信先を把握していないEndpoint宛の通信をSpineに転送する Spine Proxy の機能を備えているが、Remote Endpoint の情報をキャッシュすることで、毎回Spineを経由することなく直接VXLAN Tunnel経由でのLeaf間転送が可能となる(Spineの負荷軽減と転送処理時間の最小化)。

また、送信元Leaf側でSource EPGとDestination EPGの組み合わせを把握できていれば、送信元Leaf側でPolicyの適用が可能となるため、無駄な通信をFabric内に流さずに済む(Ingress Policy Enforcement)。

ACI構成ノードにおける MAC/IP アドレスの扱い

Spineが持つCOOP DBでは、1つのEndpoint情報として1つのMACアドレスに加えて、必要に応じて1つもしくは複数のIPアドレスが紐付けて管理されるが、Endpoint TableとしてはそれぞれのMACアドレスとIPアドレスは分けて構成される。そのため、ScalabilityはMACアドレス、IPv4アドレス、IPv6アドレスのそれぞれの最大値と合計した値の両方で決定される。

- Leaf の Scalability は、HWモデルと構成している Scale Profile 次第で決定される
- Spine の Scalability は、Modular・Fixedの選択で決定される

Leaf / Spine の Endpoint 最大エン트리数

Maximum number of endpoints (EPs)

Leaf側のScalability
(使用モデルとScale Profileの選択で異なる)

| ALE/LSE Type | ACI-Supported ToR switches |
|--------------|---|
| ALE v1 | <ul style="list-style-type: none"> N9K-C9396PX + N9K-M12PQ N9K-C93128TX + N9K-M12PQ N9K-C9396TX + N9K-M12PQ |
| ALE v2 | <ul style="list-style-type: none"> N9K-C9396TX + N9K-M6PQ N9K-C93128TX + N9K-M6PQ N9K-C9396PX + N9K-M6PQ N9K-C9372TX 64K N9K-C9332PQ N9K-C9372PX |
| LSE | <ul style="list-style-type: none"> N9K-C93108TC-EX N9K-C93180YC-EX N9K-C93180LC-EX N9K-C9336C-FX2 N9K-C93216TC-FX2 N9K-C93240YC-FX2 N9K-C93360YC-FX2 |
| LSE2 | <ul style="list-style-type: none"> N9K-C93108TC-FX N9K-C93180YC-FX N9K-C9348GC-FXP N9K-C93600CD-GX |

Default (Dual Stack) profile:

- ALE v1 and v2:
 - MAC: 12,000
 - IPv4: 12,000 or
 - IPv6: 6,000 or
 - IPv4: 4,000
IPv6: 4,000

Default profile or High LPM profile:

- LSE or LSE2:
 - MAC: 24,000
 - IPv4: 24,000
 - IPv6: 12,000

IPv4 scale profile:

- LSE and LSE2:
 - MAC: 48,000
 - IPv4: 48,000
 - IPv6: Not supported
- ALE v1 and v2: Not supported

High Dual Stack scale profile:

- LSE:
 - MAC: 64,000
 - IPv4: 64,000
 - IPv6: 24,000

- LSE2:
 - MAC: 64,000
 - IPv4: 64,000
 - IPv6: 48,000

ALE v1 and v2: Not supported

High Policy profile:

- LSE2 (N9K-C93180YC-FX and N9K-C93600CD-GX with 32GB of RAM only):
 - MAC: 24,000
 - IPv4: 24,000
 - IPv6: 12,000

Modular spine switches:

Max. 450,000 Proxy Database Entries in the fabric, which can be translated into any one of the following:

- 450,000 MAC-only EPs (each EP with one MAC only)
- 225,000 IPv4 EPs (each EP with one MAC and one IPv4)
- 150,000 dual-stack EPs (each EP with one MAC, one IPv4, and one IPv6)

The formula to calculate in mixed mode is as follows:
#MAC + #IPv4 + #IPv6 <= 450,000

NOTE: Four fabric modules (N9K-C9508-FM-E) are required on all spines in the fabric to support above scale.

Spine側のScalability
(Modular / Fixed で異なる)

Fixed spine switches (N9K-C9364C and N9K-C9316D-GX):

Max. 180,000 Proxy Database Entries in the fabric, which can be translated into any one of the following:

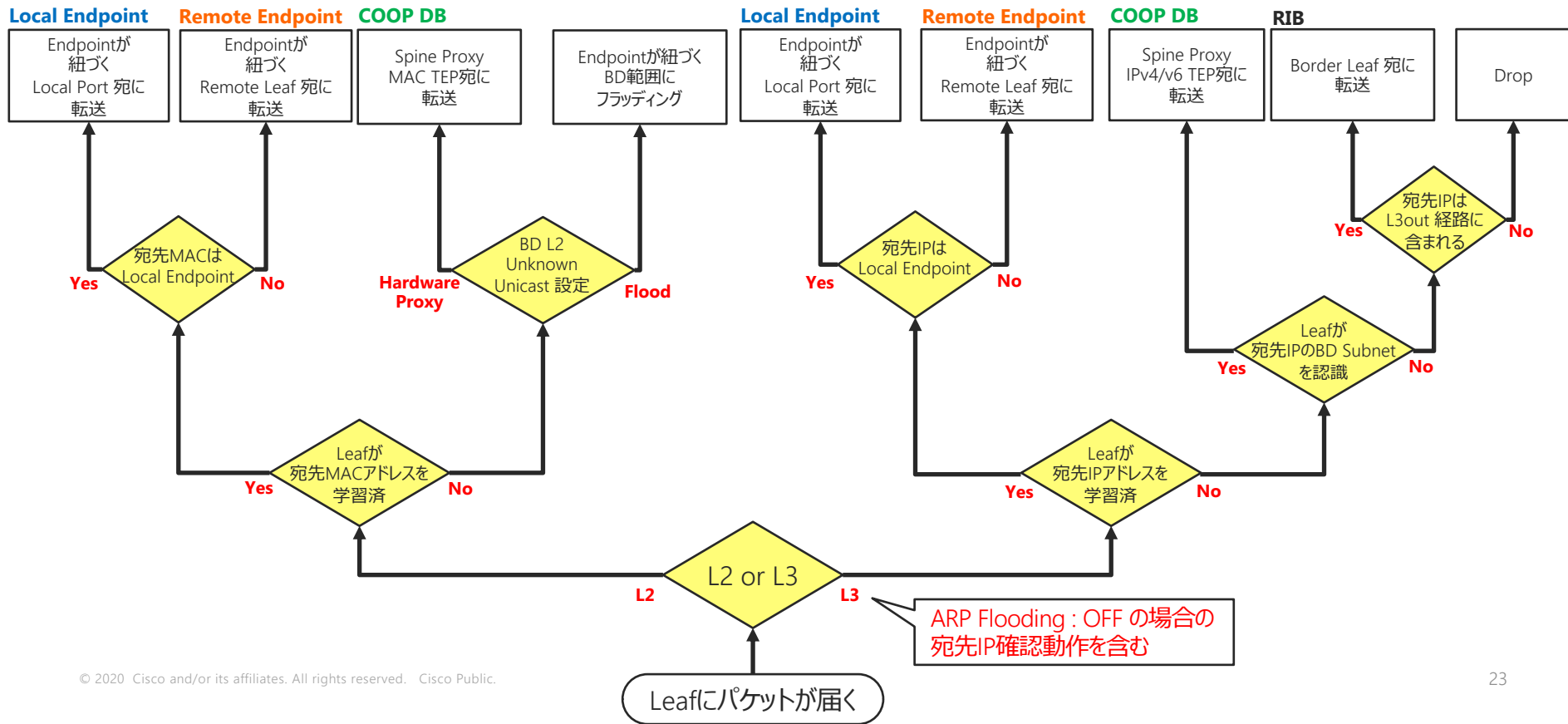
- 180,000 MAC-only EPs (each EP with one MAC only)
- 90,000 IPv4 EPs (each EP with one MAC and one IPv4)
- 60,000 dual-stack EPs (each EP with one MAC, one IPv4, and one IPv6)

The formula to calculate in mixed mode is as follows:
#MAC + #IPv4 + #IPv6 <= 180,000

Forwarding 基礎

※詳細は “ACI Design : Forwarding Behavior Deep Dive” 参照

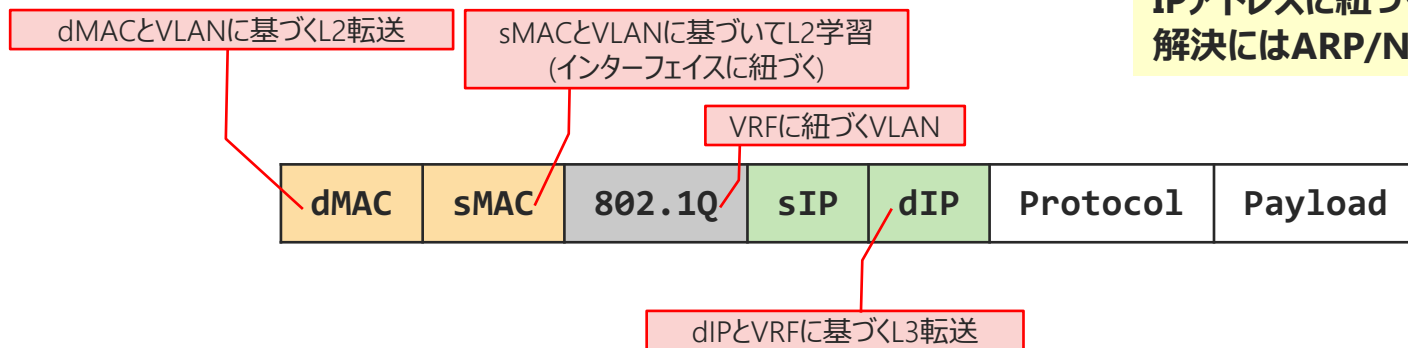
ACI Forwarding 概要



Endpoint 學習動作

参考：従来のスイッチにおける学習動作

IPアドレスに紐づくMACアドレスの
解決にはARP/NDを利用する



L2転送

| | |
|------------------------|----------|
| dMAC + VLAN → エントリにヒット | 宛先ポートに転送 |
| dMAC + VLAN → GW MAC | ルーティング処理 |
| dMAC + VLAN → ミス | Flooding |

L3転送 (LPM)

| | |
|----------------------|--------------|
| VRF + dIP → エントリにヒット | Next Hop に転送 |
| VRF + dIP → ミス | Drop |

Local Endpoint の学習方法

Local Endpoint は 当該Leafのフロントパネルポート(非Fabricポート)から学習する

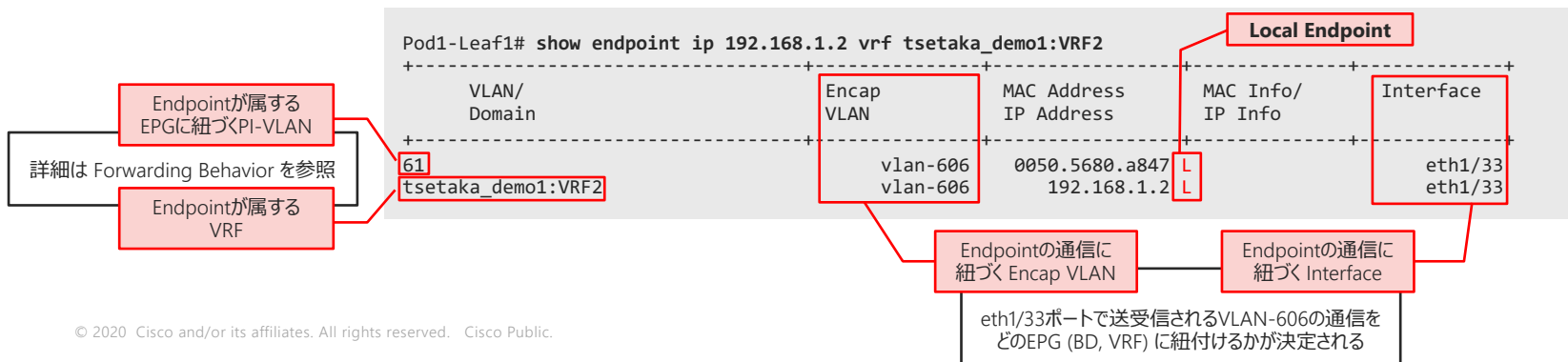
Local Endpoint 学習方法

MAC Address フロントパネルポートから受信したパケット・フレームのSource MACアドレスを学習する

フロントパネルポートから受信したACI BD Gateway宛パケットのSource IPアドレスを学習する (Data-plane Learning)

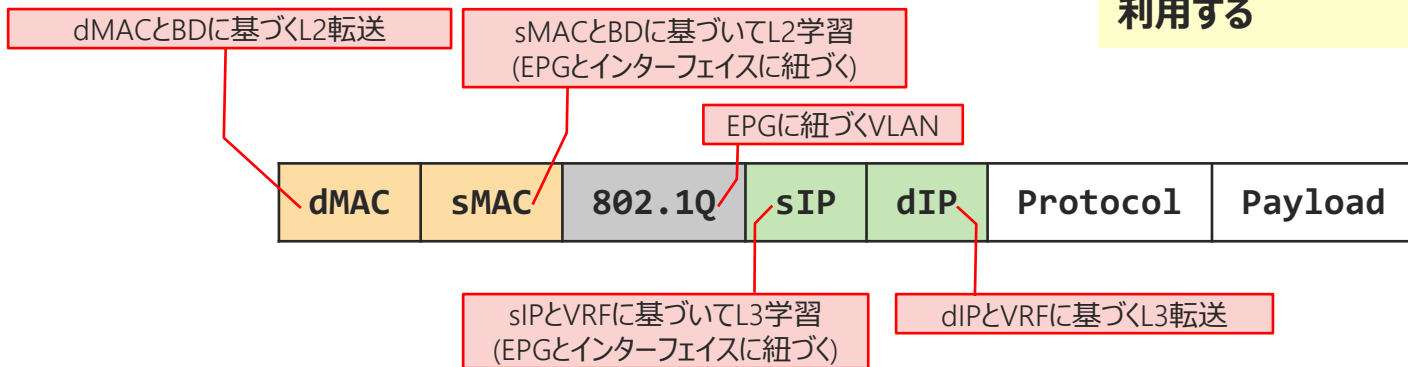
IP Address

フロントパネルポートから受信したARP/NDリクエストのSender IPアドレスを学習する (Control-plane Learning)



Local Endpoint 学習動作

IPアドレスに紐づくMACアドレスの解決にはFloodもしくはSpine Proxyを利用する



L2転送

| | |
|----------------------|-------------------------------------|
| dMAC + BD → エントリにヒット | 宛先ポート もしくは Next Hop に転送 |
| dMAC + BD → GW MAC | ルーティング処理 |
| dMAC + BD → ミス | Flooding もしくは Spine Proxy + Drop |

L3転送 (LPM)

| | |
|----------------------|--|
| VRF + dIP → エントリにヒット | 宛先ポート もしくはNext Hop に転送 |
| VRF + dIP → ミス | Drop + Spine Proxy (必要に応じてSpine側でARP Glean) |

Unicast Routing 設定と Endpoint 学習

Local Endpointの学習動作は紐づくBDの設定によって学習対象が異なる。
COOPによる通知対象。デフォルトの Retention Timer は 900秒 (15分)。

| BD設定 | MACアドレス | IPアドレス |
|---------------------------|---------------|---|
| Unicast Routing 無効 | Data-plane 学習 | 学習しない |
| Unicast Routing 有効 | Data-plane 学習 | ARP/ND および Data-plane学習 ※DP学習はRouting通信の場合のみ |

※例外として、Infra VLANについてはIPアドレスの学習は行われない

Remote Endpoint の学習方法

Remote Endpoint は 当該LeafのFabricポート(Spine接続ポート)から学習する

Remote Endpoint

学習方法

MAC Address

Fabricポートから受信したEthernetフレーム(L2)のSource MACアドレスを学習する (iVXLANヘッダに含まれるVNIDはBDに紐づく) ※IPアドレスとは紐付けられない

IP Address

Fabricポートから受信したIPパケット(L3)のSource IPアドレスを学習する (iVXLANヘッダに含まれるVNIDはVRFに紐づく) ※MACアドレスとは紐付けられない

Pod1-Leaf1# show endpoint ip 192.168.1.1 vrf tsetaka_demo1:VRF2

| VLAN/ Domain | Encap VLAN | MAC Address IP Address | MAC Info/ IP Info | Interface |
|--------------------|---------------|---------------------------|----------------------|-----------|
| tsetaka_demo1:VRF2 | | 192.168.1.1 | | tunne15 |

Pod1-Leaf1# show endpoint mac 0050.5680.a836 vrf tsetaka_demo1:VRF2

| VLAN/ Domain | Encap VLAN | MAC Address IP Address | MAC Info/ IP Info | Interface |
|-----------------------|----------------|---------------------------|----------------------|-----------|
| 58/tsetaka_demo1:VRF2 | vxlan-16646017 | 0050.5680.a836 | | tunne15 |

Remote Endpointが紐づくVRF

Remote IPアドレスは VRF VNIDに紐づく

表示されるPI-VLANは自身が管理するBDに紐づくVLAN

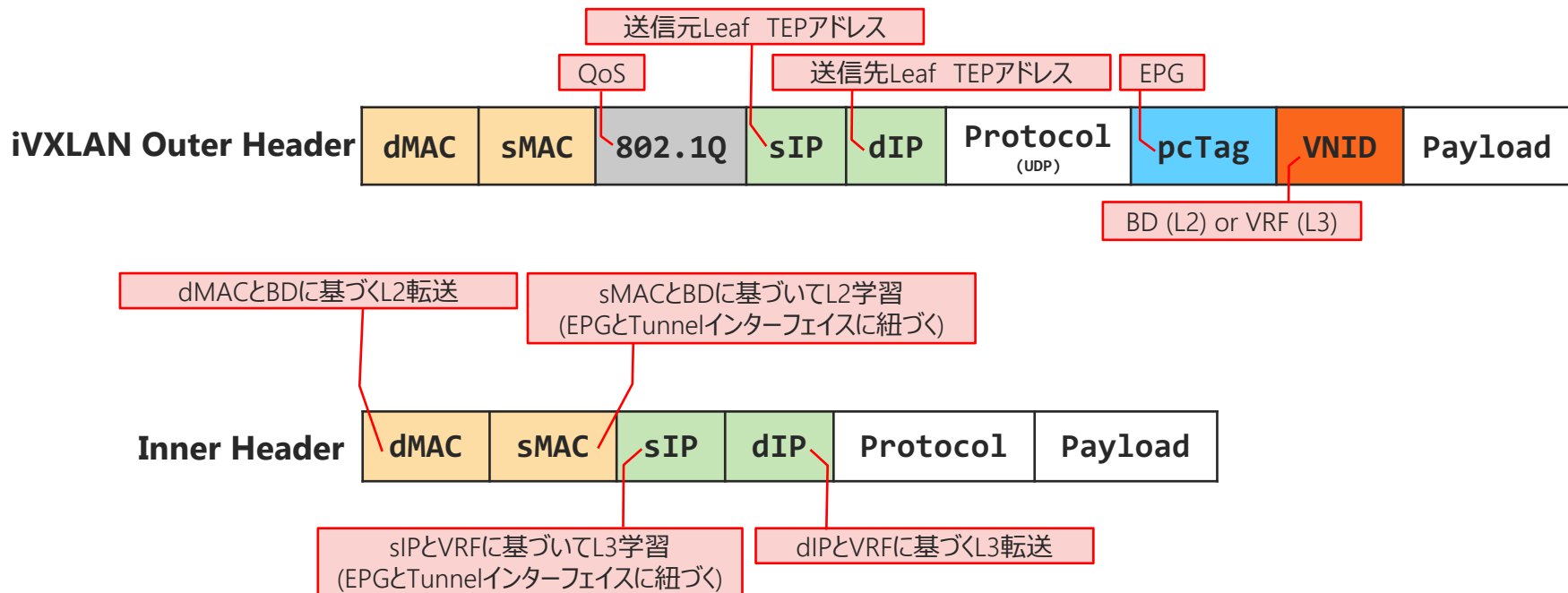
Remote Endpointが紐づくPI-VLANとVRF

Remote Endpointが存在するノードとの間のVXLANトンネル Interface

詳細は Forwarding Behavior をあわせて参照

BDが紐づくVXLAN VNID

Remote Endpoint 学習動作



VNID と Remote Endpoint 学習

Remote Endpointの学習動作はVXLANヘッダに含まれるVNIDにより学習対象が異なる。COOPの通知対象ではない。デフォルトの Retention Timer は 300秒 (5分)。

| VNID | MACアドレス | IPアドレス |
|-------------|---------|--------|
| BDに紐づくVNID | 学習する | 学習しない |
| VRFに紐づくVNID | 学習しない | 学習する |

同じEndpointのMACアドレスとIPアドレスを対象としている場合であっても、Remote Endpointとしては別々に扱われる(Local EndpointのようにMACアドレスとIPアドレスの紐付けは管理されない)。

Endpoint 学習タイプの分類

Endpoint Tableには様々な学習タイプのEndpointがエントリされている

```
Pod1-Leaf1# show endpoint vrf tsetaka_demo1:VRF2
```

```
Legend:
```

```
s - arp          H - vtep          V - vpc-attached  p - peer-aged
R - peer-attached-r1 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/ Domain | Encap VLAN | MAC Address IP Address | MAC Info/ IP Info | Interface |
|-----------------------|----------------|---------------------------|----------------------|-----------|
| 61 | vlan-606 | 0050.5680.a847 | L | eth1/33 |
| tsetaka_demo1:VRF2 | vlan-606 | 192.168.1.2 | L | eth1/33 |
| 58/tsetaka_demo1:VRF2 | vxlan-16646017 | 0050.5680.a836 | | tunne15 |
| tsetaka_demo1:VRF2 | | 192.168.1.1 | | tunne15 |
| 50 | vlan-2004 | 0050.5680.a045 | LV | po1 |
| tsetaka_demo1:VRF2 | vlan-2004 | 192.168.2.2 | LV | po1 |
| 48/tsetaka_demo1:VRF2 | vlan-2443 | 0050.5685.3a8a | O | tunne14 |
| 63/tsetaka_demo1:VRF2 | vxlan-15269818 | 0050.5685.ce43 | L | eth1/22 |

Local Endpoint
(VLAN)

当該Leafに接続している
Endpoint

Local Endpoint
(VXLAN)

VXLANで接続している
Endpoint

- AVE/AVS接続VM
- OpenStack (VXLAN)
- Kubernetes ACI CNI
等

当該Leafに接続している
Endpointが通信するために
キャッシュとして学習している
他のノードに接続している
Endpoint

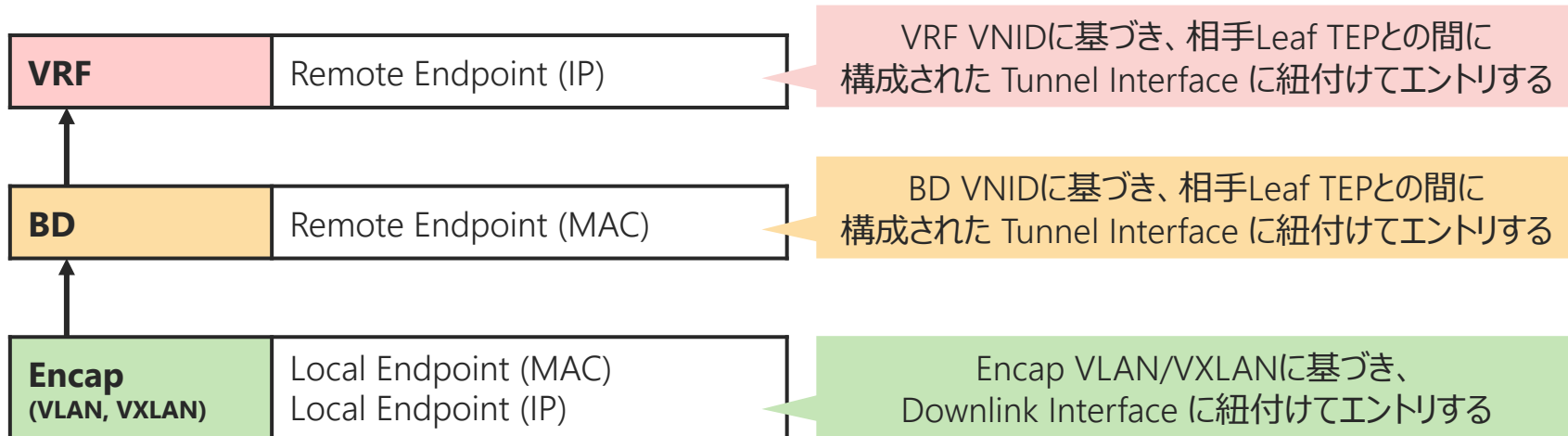
Remote
Endpoint

VPC接続
Endpoint

VPC接続
Endpoint
(Orphan Port)

VPCを構成しているため
相互に学習ステータスを
同期している対象のEndpoint

Endpoint Table における紐付け先の違い



※ネットワークとしての識別にはVNIDが用いられるが、合わせてセキュリティの識別にはpcTagに基づくEPGへの紐付けによって識別される

ACI外部は Endpoint としては管理しない

Endpointテーブルでは1つのMACアドレスに対して紐付けることができるIPアドレスの数に制限がある(ACI3.2以降では4096,それ以前では1024)ため、ACIが外部として扱う範囲(L3out)に対してはLocal Endpointとしてではなく、従来同様にARPによって隣接ノードのIPアドレスに紐づくMACアドレスを学習する。

| 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Leafに対してプログラムされる |
| 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MACアドレスを学習する | |

ACI外部の経路情報については、Static Routeを利用した静的な定義と、OSPF, BGP, EIGRPを利用したDynamic Routingによる動的な経路情報を通じて設定される

L3out 通信時の Endpoint 学習動作

L3outとの間での通信では、通常のEndpoint同士の通信とは動作が異なる。

| L3out通信 | MACアドレス | IPアドレス |
|--------------------|--|---|
| L3out → ACI | Local (Border Leaf) : 学習する Remote : 学習しない | Local (Border Leaf) : 学習しない Remote : 学習しない |
| ACI → L3out | Remote : Gen 1 Leaf からの送信である場合、送信元MAC/IPアドレスともに Remote Endpointとしては学習しない(DLビットが付加される) | |

Gen2 Leafスイッチの場合、[Enforce Subnet Check]機能を有効化することで、L3out経路範囲に含まれる送信元IPアドレスはEndpointとして学習しない(0.0.0.0/0を除く)。

Issue Case #1

Teaming構成のミスによるIPアドレスのFrapping

スイッチ側のPort-channelをあわせて構成する必要があるTeaming方式を、サーバ側だけで構成してしまった場合、Teamingインターフェイスに構成したIPアドレスが複数の物理アダプタのMACアドレス間でフラッピングする事象が発生する。

ACIでこのような事態が発生すると、LeafのCPU使用率の上昇や、COOP処理を通じたSpineのCPU使用率の上昇などが発生してしまうことがある。



Rogue EP Control : Enabled

指定した間隔の間に指定した回数以上のEndpointの移動が行われた場合、指定された間、対象のEndpoint情報を静的にエントリしてしまうことで新規Endpointとしての学習を防止するとともに、当該Endpointの通信に対してDLビットを付加してRemote Endpointとしての学習を防止する。

この対応は問題の解決にはつながらないが、ACI Fabric側の過負荷を抑制するとともに、管理者側が構成上の問題を認識するキッカケを提供する(Warning Fault扱い)。

The screenshot shows the ACI System Settings page. The left sidebar lists various settings, with 'Endpoint Controls' selected. The main content area is titled 'Endpoint Controls' and has tabs for 'Ep Loop Protection', 'Rogue EP Control', and 'Ip Aging'. Under 'Rogue EP Control', there are tabs for 'Policy' and 'History'. The 'Properties' section shows the following configuration:

| | | |
|---|----------|---------|
| Administrative State: | Disabled | Enabled |
| Rogue EP Detection Interval: | 60 | ↑ ↓ |
| Rogue EP Detection Multiplication Factor: | 4 | ↑ ↓ |
| Hold Interval (sec): | 1800 | ↑ ↓ |

デフォルト値は60秒間に4回以上のEndpoint移動を検出した場合に、1800秒(30分間) Endpoint Learning動作を停止する
※ただし機能自体はデフォルト無効(Disabled)

EP Loop Protection : Enabled

指定した間隔の間に指定した回数以上のEndpointの移動が行われた場合、以下の指定されたアクションを実行する。

- BDにおける新規学習の停止 (Action: BD Learn Disable)
- ポート停止 (Action: Port Disable)

どちらのアクションを選択した場合も、問題の対象となっているEndpoint以外に対しても与える影響が非常に大きくなるため、利用は可能な限り避ける(非推奨)。

The screenshot shows the Cisco NCS interface for configuring Endpoint Controls. The left sidebar lists various system settings, with 'Endpoint Controls' selected. The main panel shows the 'Ep Loop Protection' configuration. The 'Administrative State' is set to 'Enabled'. The 'Loop Detection Interval' is 60, and the 'Loop Detection Multiplication Factor' is 4. The 'Action' section has 'BD Learn Disable' and 'Port Disable' checked.

Issue Case #2

リンクローカルアドレスなどの情報が残り続ける

Endpoint に紐づくIPアドレスが個別にタイムアウトしない場合、一時的に構成されたリンクローカルアドレス(169.254.x.x/16)などが残り続けてしまう。

IP Agingを有効化することで、EndpointのMACアドレスに複数のIPアドレスが紐付けられている場合において個別のIPアドレス毎にAging Timerが構成される。

The screenshot displays the Cisco ACI System Settings interface. The 'System' menu item is highlighted in the top navigation bar. The 'System Settings' menu is open on the left, with 'Endpoint Controls' selected. In the main content area, the 'Endpoint Controls' page is shown, with the 'Ip Aging' tab selected. The 'Administrative State' is set to 'Enabled', which is highlighted with a red box. A callout box points to the 'Enabled' button with the text 'ACI 2.1(1h)以降でデフォルト有効'.

IP Aging Policy

- Endpointとしてエントリされた送信元IPアドレスからのフレームを受信する度にセットされるHit-Bitに基づいてリセットされるAging Timer
- Aging Timerがタイムアウトすると、当該エントリはEndpoint Tableから削除される
- タイムアウト値は Endpoint Retention Policy 設定に基づく (VRF, BDで構成可)
- Local Endpointの場合、Aging Timer設定値の75%経過時にLeafスイッチからUnicast ARP/NDが3回送信され、存在確認が実行される
※LeafからのARP/NDに応答する限り、Endpoint Tableから削除されることはない
- Endpointに複数のIPアドレスがある場合、個別IP毎にAging Timerが設定される
- 2つのLeafがVPCドメインを構成している場合、VPCホストとOrphanホストの両方についてタイムアウト時には両ノードからエントリは削除される

Endpoint Retention Policy

無指定の場合は、以下の設定値がデフォルトとして使われる。

Tenants > common > Policies > Protocol > End Point Retention > default

| Timer | Timeout デフォルト値 | BD | VRF |
|---------------------------------------|-------------------|--|-----|
| Local Endpoint Aging Interval | 900秒 (15分) | MAC/IP ※IP Aging Policyにより個別に Agingする | — |
| Bounce Entry Aging Interval | 630秒 (10分30秒) | MAC | IP |
| Remote Endpoint Aging Interval | 300秒 (5分) | MAC | IP |
| Move Frequency | 256回/秒 | — | — |
| Hold Interval | 300秒 (5分) | — | — |

※各テナントで Endpoint Retention Policy を構成することが可能

※Endpoint Retention Policy は、VRF や BD の Policy パラメータとして紐付けることが可能

Issue Case #3

※“ACI Endpoint Learning : ベストプラクティス”を
合わせて参照下さい

Endpointで誤ったIPアドレスを構成してしまった

Endpointで誤って意図しないIPアドレスを構成してしまった場合、そのIPアドレスがEndpointとしてエントリされてしまうとタイムアウトするまでの間、当該IPアドレスを持つ本来通信できるべき宛先との間で通信ができなくなってしまう。

この問題への対処方法は以下の通り(詳細後述)。

- **Limit IP Learning To Subnet : Enabled**

- ※ACI 1.1(1)以降で利用可能、ACI 2.3(1e)および3.0(1k)以降で新規に作成されたBDではデフォルトで有効
 - ※以下の [Enforce Subnet Check]が利用できる かつ 全LeafスイッチがGen 2であれば、構成不要

- **Enforce Subnet Check : Enabled** ←必要要件を満たすならば、こちらが推奨

- ※ACI 2.2(2q)および3.0(2)以降 かつ Gen 2 Leaf に対してのみ有効
(ただし、ACI 2.3および3.0(1)には実装されていないので注意)

Limit IP Learning To Subnet : Enabled

BD/EPGに構成されているSubnet範囲外のIPアドレスについて、Local Endpointとして学習しない。ただし、Border LeafがRemote Endpointとして学習することを防止することはできないため、完全な解決にはならない(VRFのVNIDのみで判断するため)。

Border Leaf において Remote Endpoint として学習してしまうことに対しては、合わせて [Disable Remote EP Learning] を有効化することで対処する。

The screenshot displays the Cisco ACI GUI interface. On the left, a navigation pane shows the hierarchy: System > Tenants > tsetaka_demo1 > Networking > Bridge Domains > BD1-1. The main content area shows the configuration for 'Bridge Domain - BD1-1'. Under the 'Policy' tab, the 'General' sub-tab is active. In the 'Properties' section, the 'Limit IP Learning To Subnet' checkbox is checked and highlighted with a red box. Other visible settings include 'IP Data-plane Learning' set to 'no', 'Multi Destination Flooding' set to 'Flood in BD', and 'ARP Flooding' checked. The top navigation bar includes tabs for System, Tenants, Fabric, Virtual Networking, L4-L7 Services, Admin, Operations, Apps, and Integrations.

Limit IP Learning To Subnet 注意点

[Limit IP Learning To Subnet]オプションを、すでに利用中のBDに対してEnabled / Disabled の設定変更を実施した場合、ACI 3.0(1k)より前では以下の状態が発生する。

- 当該BD範囲において全ての学習済Endpointが一旦消去される
- 120秒間 MACアドレス および IPアドレス の学習が停止される

3.0(1k)以降では、上記に関して、以下の改善がされている(CSCve29663)。

- 有効化・無効化時において、学習済のEndpoint情報のフラッシュは行われ
ない(ただし、BD Subnet範囲外のEndpoint IPアドレス情報については消去される)
- 120秒間の Endpoint 学習の停止は発生しない

Enforce Subnet Check : Enabled

ACI 2.2(2q)~
ACI 3.0(2h)~

VRF範囲に構成されたBD/EPGのSubnet範囲外のIPアドレスはEndpointとして学習しない(Local Endpoint / Remote Endpoint いずれとしても)。

Gen 2 以降の Leaf スイッチに対してのみ有効。

The screenshot displays the Cisco ACI System Settings page. The 'System' tab is selected in the top navigation bar. The 'System Settings' menu on the left has 'Fabric-Wide Settings' highlighted. The main content area shows the 'Fabric-Wide Settings Policy' configuration. Under the 'Properties' section, the 'Enforce Subnet Check' checkbox is checked and highlighted with a red box. The description for this option is: 'To disable IP address learning on the outside of subnets configured in a VRF, for all VRFs'. Other options like 'Disable Remote EP Learning', 'Enforce EPG VLAN Validation', 'Enforce Domain Validation', 'Enable Remote Leaf Direct Traffic Forwarding', 'Opflex Client Authentication', and 'Reallocate Gipo' are also visible.

Issue Case #4

ACIに接続するデバイスを制御できないケース

ACIのData-planeにおけるEndpoint学習の仕組みは、ACIの様々な機能の基盤となっているため、基本的には利用することが強く推奨されますが、従来のスイッチと同じ様なARP/NDのみによるIP学習動作を求められるようなケースにおいては、Data-plane Learning を無効化することが可能。

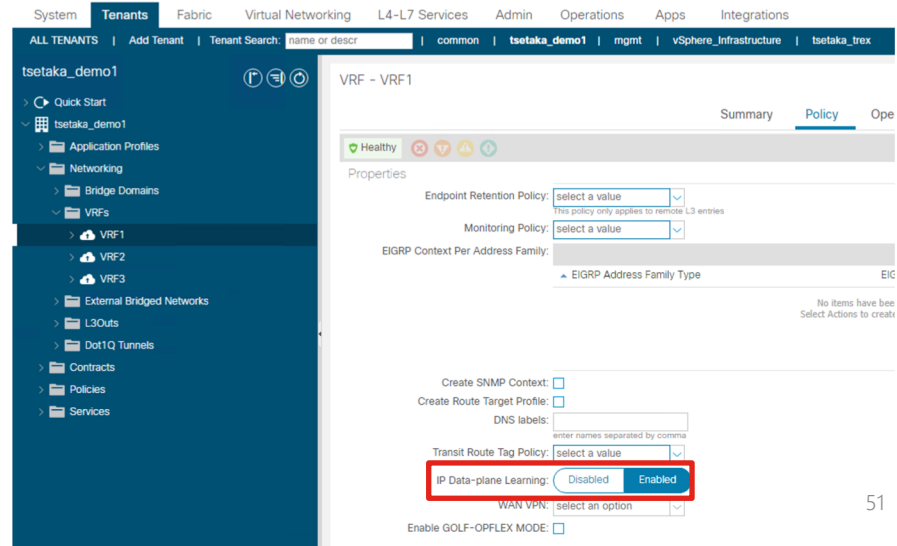
例：ACI側での適切な構成を行うことなく(ネットワーク管理者が接続種別に合わせて必要な設定を定義することが出来ない)、EndpointとしてRouting動作をするデバイスやロードバランス動作、同一IPアドレスをARPベースでフェイルオーバー利用するサーバなどを接続して利用される可能性があるケース

※IP Data-plane Learning 無効化時には、IP Agingを有効化することを推奨 (Default 有効)

IP Data-plane Learning : Disabled (VRF)

IP Data-plane Learning を VRF で Disabled とした場合、以下の動作となる。

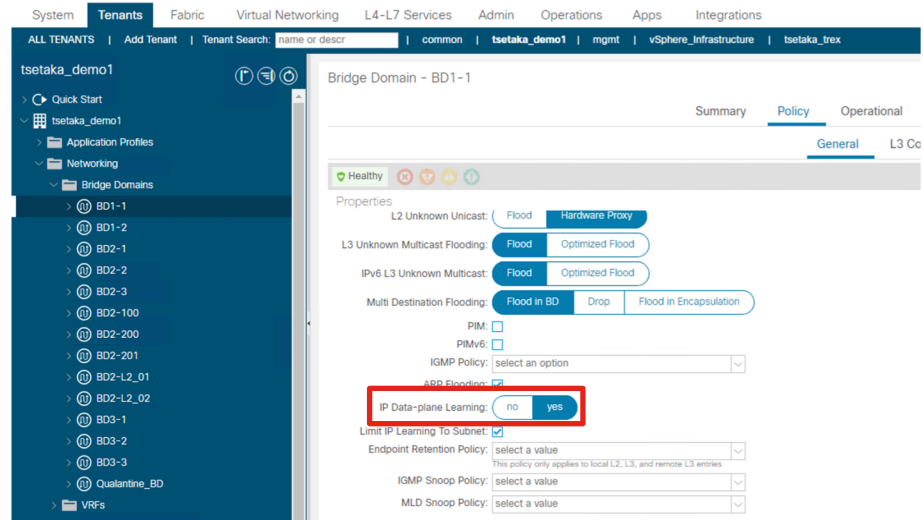
- Local Endpoint として MACアドレスは引き続きData-planeから学習する(Gen1/2)
- Remote Endpoint として MACアドレスは Data-planeから学習する (Gen2のみ)
※Gen 1がRemote Endpointを学習しないため、L2 Unknown Unicast 設定は Hardware Proxy を必ず利用する必要がある
- Local Endpoint の IPアドレスは Control-planeを通じたARP/NDでのみ学習する
※IP Aging は必ず Enable とする必要がある
- Remote Endpoint の IPアドレスは Unicast からは学習せず、Routed Multicast からのみ学習する



IP Data-plane Learning : Disabled (BD)

IP Data-plane Learning を BD で Disabled とする使い方は、Service Graph で PBRを利用することのみが現在はサポートされている。
(要Gen 2 LeafスイッチへのService Node接続)

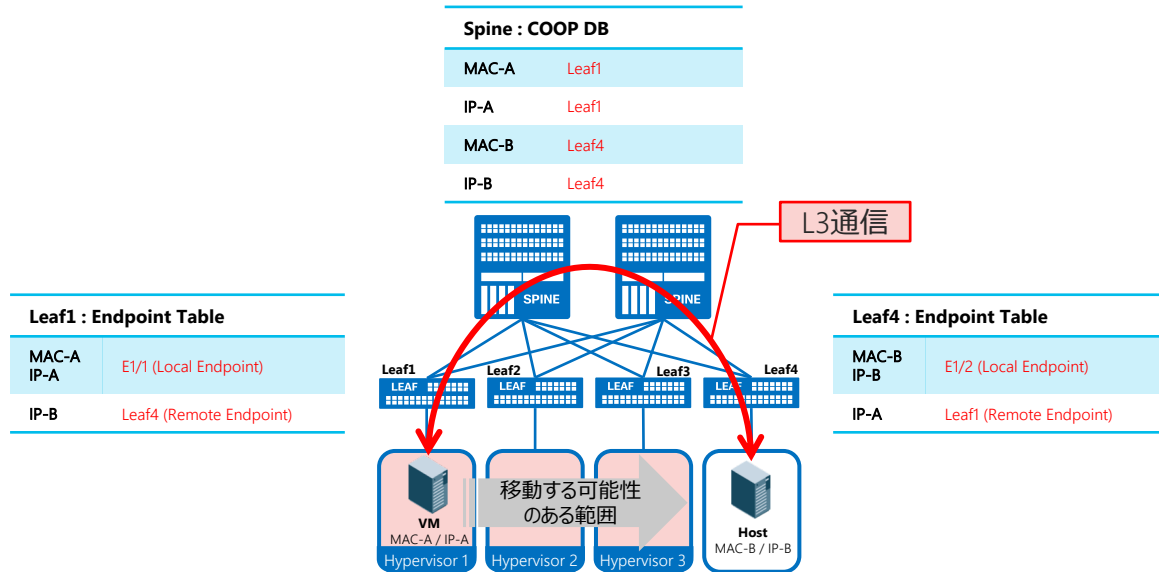
ACI 3.1以降では自動的にPBR利用される Service Node接続EPG単位で自動的に IP Data-plane LearningがDisabledとなるため、上記を手動で対応する必要はない。



Endpoint 移動への対応

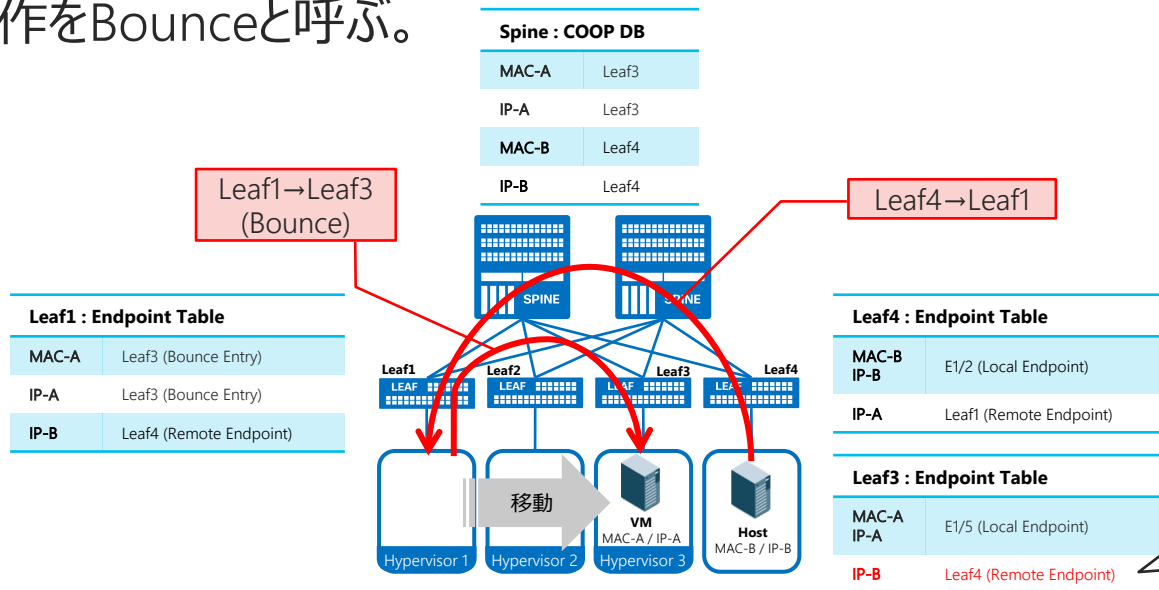
Bounce Entry (1/4)

EndpointがVMの場合などでは、Endpointは異なるLeaf配下へ動的に移動する可能性がある。しかし、ACIではRemote Endpoint情報をキャッシュとして保持する為、転送先Leafにはすでに当該Endpointが存在しない状態が発生しうる。



Bounce Entry (3/4)

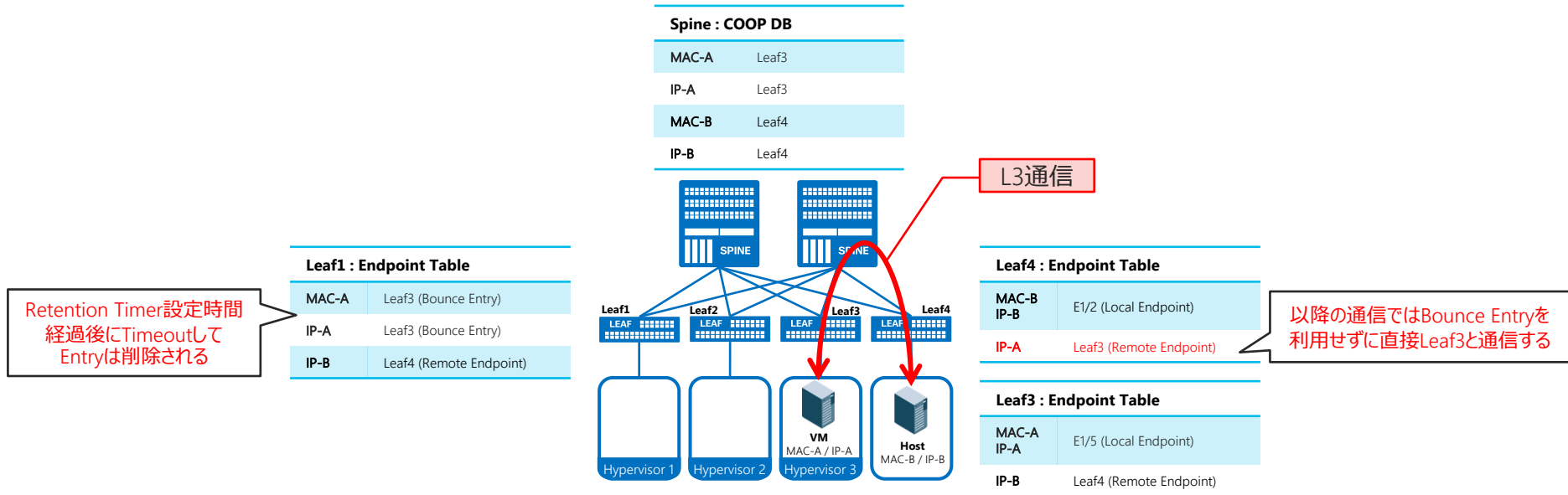
Remote Endpoint情報を保持したままとなっているLeaf4は、この時点ではVMの移動を認識していないため、VM宛の通信をLeaf1に対して転送する。Leaf1は当該VMがLeaf3に移動していることを認識しているため、Leaf4からの転送をLeaf3に再転送する。この動作をBounceと呼ぶ。



Bounce 通信では送信元LeafのTEPアドレスは書き換えられないため、Leaf4からの送信として学習する

Bounce Entry (4/4)

Leaf3配下に移動したVMからの戻り通信は、Leaf3がBounceによりLeaf4を宛先TEPとして学習済のため、直接Leaf4に戻される。この時点で、Leaf4はVMの移動を認識し、自身のRemote Endpointエントリを更新する。



Bounce Entry と Endpoint Retention Timer

Bounce Entryは通常のRemote Endpointよりも長い(倍+30秒)Retention Timerが設定されるため、Bounce EntryがTimeoutする前にRemote EndpointのエントリがTimeoutする。これにより、Bounce EntryがTimeoutしても問題は発生しない。

| Timer | Timeout デフォルト値 | BD | VRF |
|---------------|-------------------|--------|-----|
| Local | 900秒(15分) | MAC/IP | — |
| Bounce | 630秒 (10分30秒) | MAC | IP |
| Remote | 300秒(5分) | MAC | IP |
| Move | 256回/秒 | — | — |
| Hold | 300秒 | — | — |

Bounce Entryの仕組みにより、ACI Fabricの規模が拡大してもEndpoint移動時にSpineが更新を通知する対象は元々当該Endpointを紐付けていた移動元Leafのみのままとするため、負荷が増大しない(通知対象が増加しない)。

Endpoint Retention Policy は、BD単位やVRF単位でのカスタム構成が可能

Bounce Entry 削除時の通知：3.2(2l)～

ACI 3.2(2l)以降において、Bounce Entry削除時に全Leafに対して該当Remote Endpoint情報の削除を通知を行う機能が実装された(CSCvj17665)。これは極稀な例外的なケース(CSCva56754)としてBorder LeafにおいてRemote Endpoint情報が残ってしまう場合への対処のためである。この機能を利用する上で必要な構成はない(自動的に利用可能となる)。

Issue Case #5

ACI 内から ACI 外への Subnet の移行

ACI内からL3outの外にSubnetを移行した場合、Bounce Entryは構成されない。

L3out通信

IPアドレス

L3out → ACI

Remote : 学習しない

Remote Endpointエントリを持っているLeafからは、L3outの外に移行した宛先にはタイムアウトするまで疎通できない。

→ Leaf毎に Remote Endpoint 情報を手動でクリアすることが可能

特定Remote EndpointのIPアドレスエントリのみを削除する場合

```
LEAF1# clear system internal epm endpoint key vrf <vrf-name> ip <ip-address>
```

全Remote Endpointのエントリを削除する場合(MACアドレスを含む)

```
LEAF1# clear system internal epm endpoint vrf <vrf-name> remote
```

※Gen2 Leafの場合は[Enforce Subnet Check]を有効化することで、BD Subnetの削除によるEndpoint情報の削除が可能。

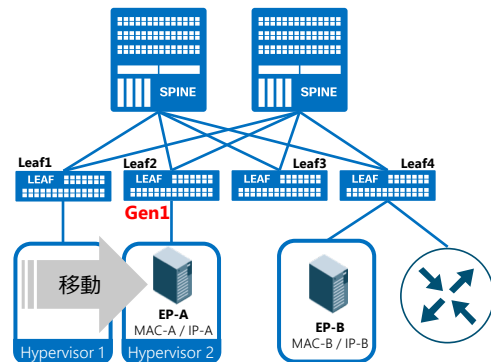
Issue Case #6

※“ACI Endpoint Learning : ベストプラクティス”を
合わせて参照下さい

送信元LeafでGen1利用の場合の問題点(1/2)

以下の条件を満たしてしまった場合、Border LeafにおいてRemote Endpointのエントリが更新されない状況が発生する。

- Ingress Policy Enforcement (Default)を利用
- Border LeafにEndpoint-Aと通信を行ったEndpoint-Bが存在する (Border LeafはEndpoint-AをRemote Endpointとして学習している)
- 送信元Endpoint-Aが Leaf2 (Gen1)配下へと移動した後は、Endpoint-Bとは通信せずに当該Border Leafに構成されているL3outの先との間でのみ通信を行う
(L3outとの通信によってRemote EndpointのRetention Timerは更新されるが、Leaf2への移動は学習しない = Leaf1に紐付いたままとなる)
- 移動前のLeafに構成されていたBounce Entryタイムアウト後Endpoint-AはL3outとの疎通ができなくなる



Leaf4 : Endpoint Table

| | |
|---------------|----------------------------|
| MAC-B IP-B | E1/2 (Local Endpoint) |
| IP-A | Leaf1 (Remote Endpoint) |

L3out通信

MACアドレス

IPアドレス

ACI → L3out

Remote : Gen 1 Leaf からの送信である場合、送信元MAC/IPアドレスともにRemote Endpointとしては学習しない(DLビットが付加される)

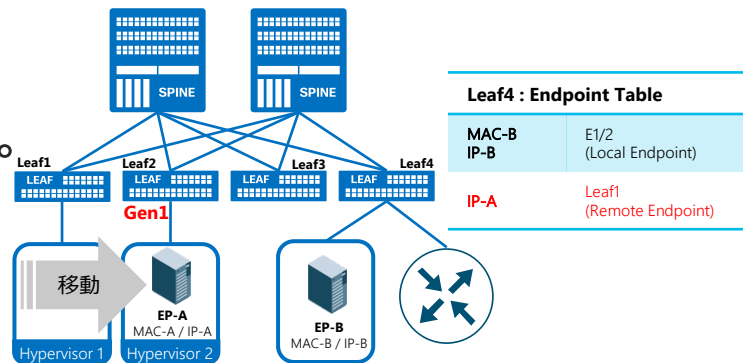
送信元LeafでGen1利用の場合の問題点(2/2)

前項の問題への対処方法は以下の通り。

1.もしくは2.を満たす場合、3.および4.の対応は不要。

4.を満たす場合、3.の対応は不要。

→3.2(2)以降へのアップグレードによる対応を推奨



1. Border Leaf (Leaf4) に Endpoint が同時に存在しない場合はこの問題は発生しない
2. 移動したEndpointの接続先Leaf (Leaf2) がGen2である場合はこの問題は発生しない
3. ACI 2.2(2)以降もしくはACI 3.0(1)以降の場合、[Disable Remote EP Learning]を有効にすることで、Border Leaf側でのRemote Endpointの学習をしないように構成することで問題の発生を防止できる
4. ACI 3.2(2)以降の場合、[EP Announce on Bounce Delete]が機能するため、Bounce Entryの削除時に合わせて、残ってしまっていたRemote Endpointエントリも削除される。これにより、問題の発生を防止できる (設定不要)

Disable Remote EP Learning : Enabled

2.2(2e)~
3.0(1k)~

VRFにL3outが紐付けられている場合に、Border LeafでのRemote Endpointの学習を無効化することで、問題の発生を予防する。

※ACI 3.2(2l)以降の場合は構成の必要なし

The screenshot shows the Cisco ACI System Settings interface. The top navigation bar includes 'System', 'Tenants', 'Fabric', 'Virtual Networking', 'L4-L7 Services', 'Admin', 'Operations', 'Apps', and 'Integrations'. The 'System Settings' tab is selected. The left sidebar shows 'System Settings' with 'Fabric-Wide Settings' highlighted. The main content area displays the 'Fabric-Wide Settings Policy' configuration page. The 'Properties' section is visible, and the 'Disable Remote EP Learning' checkbox is checked and highlighted with a red box. Other settings include 'Enforce Subnet Check', 'Enforce EPG VLAN Validation', 'Enforce Domain Validation', 'Enable Remote Leaf Direct Traffic Forwarding', 'Opflex Client Authentication', and 'Reallocate Gipo'.

※VPCペアの片方のノードのみにL3outが構成されている場合、Disable Remote EP Learning を構成している場合でもVPCノード間での同期によってRemote Endpointを学習してしまうため、Border Leafの構成には注意が必要 (CSCvi50954)

Issue Case #7

同一ポート配下での別MACアドレスへのIP移行

MAC-Aに紐づくIP-Aを Endpoint として学習済の状態において、同一ポート配下で IP-AがMAC-Aを持つHost-Aから、MAC-Bを持つHost-Bに付け替えられた場合、Gen 1 LeafモデルはデフォルトではEndpoint Tableを更新しない(CSCus77627)。



EP Move Detection Mode : Enabled

EP Move Detection Mode を有効化すると、Leaf スイッチは GARP パケットに基づいて Endpoint の移動 (別 MAC アドレスへの IP アドレスの付け替え) を認識し エントリを書き換える。

ただし、この機能を利用するためには、以下の条件を満たす必要がある。

- BD において Unicast Routing が有効となっている (ACI as Default G/W)
- ARP Flooding が有効となっている

The screenshot displays the Cisco ACI GUI for configuring a Bridge Domain. The left sidebar shows a tree view of tenants and bridge domains. The main panel shows the configuration for 'Bridge Domain - BD1-1'. The 'Policy' tab is selected, and the 'L3 Configuration' sub-tab is active. The 'Properties' section shows 'Unicast Routing' is checked. Below it, there are fields for 'Operational Value for Unicast Routing' (true), 'Custom MAC Address' (00:22:BD:F8:19:FF), and 'Virtual MAC Address' (Not Configured). A table lists subnets with columns for Gateway Address, Scope, Primary IP Address, and Virtual IP. At the bottom, a red box highlights the 'EP Move Detection Mode' checkbox, which is checked and labeled 'GARP based detection'. The page number '68' is visible in the bottom right corner.

| Gateway Address | Scope | Primary IP Address | Virtual IP |
|------------------|----------------|--------------------|------------|
| 192.168.1.254/24 | Private to VRF | False | False |

Endpoint サイレントホスト対応

L2 : Unknown MAC Address

一切自発的にフレームを送信しないL2 Endpointが存在した場合、L2スイッチング通信を成立するためには宛先MACアドレスがEndpoint TableかCOOP DBにエントリされている必要がある。

BDでUnicast Routingが有効化されていない場合、ACIはMACアドレスのみを学習するため、Unicast ARPによる存在確認を行うことが出来ない。
そのため、L2 Unknown Unicast設定を[Flood]とすることで、BUMトラフィックをBD範囲でFloodingする様に構成する必要がある。

| L2 Unknown Unicast | Endpoint Table | COOP DB |
|------------------------------------|---|------------------------------------|
| Hardware Proxy (Default) | 送信元LeafのEndpoint Tableにエントリがない場合、Spine Proxyに転送する | COOP DBに当該MACアドレスのエントリがない場合、Dropする |
| Flood (ARP Floodingも有効化される) | Endpointが紐づくBD範囲でFloodingされる(宛先Endpointに到達する) | |

L3 : Unknown IP Address

一切自発的にフレームを送信しないL3 Endpointが存在した場合、SpineのCOOP DBにエントリが存在しないとSpineはDropすると同時にLeafスイッチに対してARP Glean要求を出す。当該BDを持つLeafスイッチはPervasive Gatewayアドレスから当該IPアドレス宛のUnicast ARPを送信し存在確認を行う(ARP Glean動作)。

応答を受け取ったLeafは通常のEndpoint登録と同様にCOOPでSpineに通知するため、以降の通信に対してはSpine Proxyが成立する。

IPアドレスを登録されたEndpointはAging Timerによるタイムアウト前にUnicast ARPによる存在確認が行われるため、1度検出された後もサイレントホストだったとしても問題ない。

ARP Glean動作の詳細はForwarding Behavior資料にて解説

ARP Broadcast Flooding

Broadcast宛先のARPリクエストに対する動作は、BDにおける[ARP Flooding]設定次第で変化する。

ARP Flooding

Disabled

ARP要求の宛先IPアドレス宛にUnicastで転送する(Fabric内の通信が最適化される)。
COOP DBにエントリがない場合は、ARP Gleanによる存在確認が実行される。

Enabled

BD範囲にARP要求はFloodingされる。GARPによる通知などを行わずに移動するEndpointが存在する場合にはARP FloodingをEnableにする必要がある。
※明示的に構成しなくても、BDのUnicast Routing無効(ACIがGatewayではない)の場合はARP Floodingは有効となる

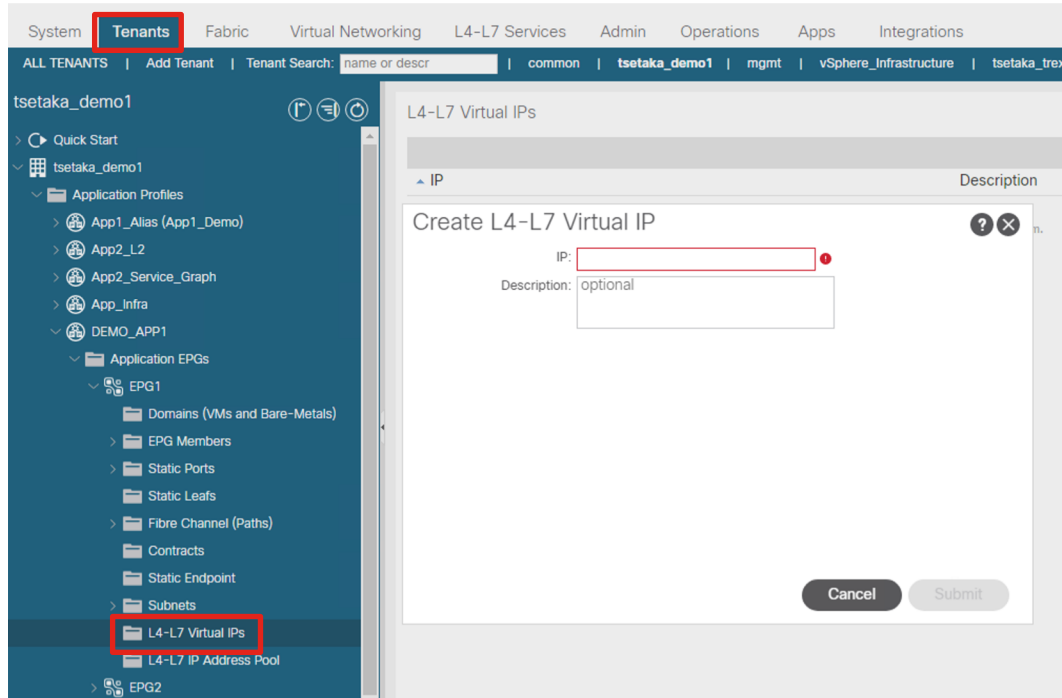
※ACI 4.2～ GUIでBDを単体で作成した場合において、[ARP Flooding]はデフォルトで有効となるように変更された。VRFとまとめてウィザード作成する場合や、API経由の場合のデフォルトは後方互換性維持のため従来どおり無効が規定値となるので注意

Endpoint Learning パラメータ詳細

EPG

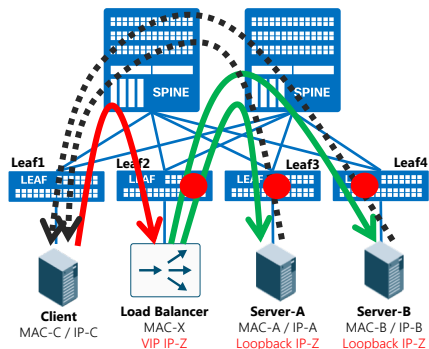
L4-L7 Virtual IPs

- Tenants > Application Profiles > AP > Application EPGs > EPG



L4-L7 Virtual IPs ユースケース

L4-L7 Virtual IP として指定したIPアドレスは、Data-plane Learningの学習から除外されます。EPGにおけるこの設定は、L2 DSR (Direct Server Return) を利用する場合のみがテスト済のユースケースです。



- ClientはLBのVIP IP-Z宛に通信(→)
 - LBは宛先MACをMAC-AもしくはMAC-B宛にして転送(→)
(送信元IPはClientのまま、宛先IPはVIPのまま)
 - Server-A/Bは送信元IPアドレスであるClientに直接返信(→)
- VIP-Z アドレスが LB, Server-A, Server-B でフラップしてしまう

VIP-Zアドレスを L4-L7 Virtual IP として設定することで、Data-plane学習を無効化することで、フラップすることなく正しく動作するようになる。

BD

Unicast Routing

- Tenants > Networking > Bridge Domains > BD > Policy > L3 Configurations

The screenshot displays the Cisco DNA Center interface for configuring a Bridge Domain (BD1-1). The navigation path is: Tenants > Networking > Bridge Domains > BD1-1 > Policy > L3 Configurations. The 'Unicast Routing' checkbox is checked, and the 'Operational Value for Unicast Routing' is set to 'true'. The 'Custom MAC Address' is 00:22:BD:F8:19:FF, and the 'Virtual MAC Address' is 'Not Configured'. A subnets table is shown below, with one entry for 192.168.1.254/24. The 'EP Move Detection Mode' is set to 'GARP based detection' (unchecked). The 'Associated L3 Out' is 'L3 Out'.

| Gateway Address | Scope | Primary IP Address | Virtual IP | Subnet Control |
|------------------|----------------|--------------------|------------|----------------|
| 192.168.1.254/24 | Private to VRF | False | False | |

EP Move Detection Mode

- Tenants > Networking > Bridge Domains > BD > Policy > L3 Configurations

The screenshot displays the Cisco DNA Center interface for configuring a Bridge Domain (BD1-1). The navigation path is: Tenants > Networking > Bridge Domains > BD1-1 > Policy > L3 Configurations. The 'EP Move Detection Mode' is currently set to 'GARP based detection'.

System **Tenants** Fabric Virtual Networking L4-L7 Services Admin Operations Apps Integrations

ALL TENANTS | Add Tenant | Tenant Search: name or descr | common | tsetaka_demo1 | mgmt | vSphere_Infrastructure | tsetaka_trex

tsetaka_demo1

- Quick Start
- tsetaka_demo1
 - Application Profiles
 - Networking
 - Bridge Domains
 - BD1-1**
 - BD1-2
 - BD2-1
 - BD2-2
 - BD2-3
 - BD2-100
 - BD2-200
 - BD2-201
 - BD2-L2_01
 - BD2-L2_02
 - BD3-1
 - BD3-2
 - BD3-3

Bridge Domain - BD1-1

Summary **Policy** Operational Stats Health Faults History

General **L3 Configurations** Advanced/Troubleshooting

Healthy

Properties

Unicast Routing:

Operational Value for Unicast Routing: true

Custom MAC Address: 00:22:BD:F8:19:FF

Virtual MAC Address: Not Configured

Subnets:

| Gateway Address | Scope | Primary IP Address | Virtual IP | Subnet Control |
|------------------|----------------|--------------------|------------|----------------|
| 192.168.1.254/24 | Private to VRF | False | False | |

EP Move Detection Mode: GARP based detection

Associated L3 Outs:

- L3 Out

Limit IP Learning To Subnet

- Tenants > Networking > Bridge Domains > BD > Policy > General

The screenshot displays the Cisco SD-WAN configuration interface. The top navigation bar includes 'System', 'Tenants', 'Fabric', 'Virtual Networking', 'L4-L7 Services', 'Admin', 'Operations', 'Apps', and 'Integrations'. The 'Tenants' tab is active, showing a list of tenants: 'ALL TENANTS', 'Add Tenant', and 'Tenant Search: name or descr'. The 'tsetaka_demo1' tenant is selected, and the 'Networking' > 'Bridge Domains' > 'BD1-1' path is highlighted in the left sidebar. The main content area shows the configuration for 'Bridge Domain - BD1-1'. The 'Policy' tab is selected, and the 'General' sub-tab is active. The 'Limit IP Learning To Subnet' checkbox is checked. Other configuration options include 'L2 Unknown Unicast' (Hardware Proxy), 'L3 Unknown Multicast Flooding' (Optimized Flood), 'IPv6 L3 Unknown Multicast' (Optimized Flood), 'Multi Destination Flooding' (Flood in BD), 'PIM', 'PIMv6', 'IGMP Policy', 'ARP Flooding' (checked), 'IP Data-plane Learning' (no), 'Endpoint Retention Policy', 'IGMP Snoop Policy', and 'MLD Snoop Policy'. The 'Healthy' status is indicated at the top of the configuration area.

IP Data-plane Learning (BD)

- Tenants > Networking > Bridge Domains > BD > Policy > General

The screenshot displays the Cisco SD-WAN GUI configuration page for Bridge Domain - BD1-1. The navigation path is: Tenants > Bridge Domains > BD1-1 > Policy > General. The 'IP Data-plane Learning' setting is set to 'no'. The 'General' tab is selected under the Policy section.

Properties:

- L2 Unknown Unicast: Flood, Hardware Proxy
- L3 Unknown Multicast Flooding: Flood, Optimized Flood
- IPv6 L3 Unknown Multicast: Flood, Optimized Flood
- Multi Destination Flooding: Flood in BD, Drop, Flood in Encapsulation
- PIM:
- PIMv6:
- IGMP Policy: select an option
- ARP Flooding:
- IP Data-plane Learning: no, yes
- Limit IP Learning To Subnet:
- Endpoint Retention Policy: select a value
- IGMP Snoop Policy: select a value
- MLD Snoop Policy: select a value

※正式にサポートされる用途は Service Graph + PBR 目的のみ

VRF

IP Data-plane Learning (VRF)

- Tenants > Networking > VRFs > VRF > Policy

The screenshot displays the Cisco SD-WAN GUI configuration page for a VRF Policy. The navigation path is Tenants > Networking > VRFs > VRF > Policy. The 'Policy' tab is selected. The 'IP Data-plane Learning' setting is highlighted with a red box and is currently set to 'Enabled'. Other settings include Endpoint Retention Policy, Monitoring Policy, EIGRP Context Per Address Family, Create SNMP Context, Create Route Target Profile, DNS labels, Transit Route Tag Policy, WAN VPN, and Enable GOLF-OPFLEX MODE.

System **Tenants** Fabric Virtual Networking L4-L7 Services Admin Operations Apps Integrations

ALL TENANTS | Add Tenant | Tenant Search: name or descr | common | tsetaka_demo1 | mgmt | vSphere_Infrastructure | tsetaka_trex

tsetaka_demo1

- Quick Start
- tsetaka_demo1
 - Application Profiles
 - Networking
 - Bridge Domains
 - VRFs
 - VRF1**
 - VRF2
 - VRF3
 - External Bridged Networks
 - L3Outs
 - Dot1Q Tunnels
 - Contracts
 - Policies
 - Services

VRF - VRF1

Summary **Policy** Open

Healthy

Properties

Endpoint Retention Policy: select a value
This policy only applies to remote L3 entries

Monitoring Policy: select a value

EIGRP Context Per Address Family:

- EIGRP Address Family Type

No items have been created. Select Actions to create.

Create SNMP Context:

Create Route Target Profile:

DNS labels:
enter names separated by comma

Transit Route Tag Policy: select a value

IP Data-plane Learning: Disabled Enabled

WAN VPN: select an option

Enable GOLF-OPFLEX MODE:

Fabric

Disable Remote EP Learning

- System > System Settings > Fabric-Wide Settings

The screenshot displays the Cisco Fabric Controller web interface. The navigation menu at the top includes 'System', 'Tenants', 'Fabric', 'Virtual Networking', 'L4-L7 Services', 'Admin', 'Operations', 'Apps', and 'Integrations'. The 'System' menu is expanded, showing 'System Settings' as the selected option. Under 'System Settings', the 'Fabric-Wide Settings' option is highlighted. The main content area is titled 'Fabric-Wide Settings Policy' and contains several configuration options, with 'Disable Remote EP Learning' being the primary focus.

System Settings

- Quota
- APIC Connectivity Preferences
- System Alias and Banners
- System Response Time
- Global AES Passphrase Encryption Settings
- BD Enforced Exception List
- Fabric Security
- Control Plane MTU
- Endpoint Controls
- Fabric-Wide Settings**
- Remote Leaf POD Redundancy Policy

Fabric-Wide Settings Policy

Policy History

Properties

- Disable Remote EP Learning:** To disable remote endpoint learning in VRFs containing external bridged/routed domains
- Enforce Subnet Check: To disable IP address learning on the outside of subnets configured in a VRF, for all VRFs
- Enforce EPG VLAN Validation: Validation check that prevents overlapping VLAN pools from being associated to an EPG
- Enforce Domain Validation: Validation check if a static path is added but no domain is associated to an EPG
- Enable Remote Leaf Direct Traffic Forwarding: Enable Remote Leaf direct communication with routable IP connectivity between Remote Leafs and Fabric, once enabled you cannot disable it again
- Opflex Client Authentication: To enforce Opflex client certificate authentication for GOLF and Linux
- Reallocate Gipo: Reallocate some non-stretched BD gipos to make room for stretched BDs

Enforce Subnet Check

- System > System Settings > Fabric-Wide Settings

The screenshot displays the Cisco Fabric Controller user interface. The top navigation bar includes 'System', 'Tenants', 'Fabric', 'Virtual Networking', 'L4-L7 Services', 'Admin', 'Operations', 'Apps', and 'Integrations'. The 'System' menu is expanded, showing 'System Settings' as the selected option. Under 'System Settings', the 'Fabric-Wide Settings' sub-menu is highlighted. The main content area is titled 'Fabric-Wide Settings Policy' and contains a list of configuration options under the 'Properties' section. The 'Enforce Subnet Check' option is highlighted with a red box. It is currently unchecked, with a tooltip that reads: 'To disable IP address learning on the outside of subnets configured in a VRF, for all VRFs'. Other visible options include 'Disable Remote EP Learning', 'Enforce EPG VLAN Validation', 'Enforce Domain Validation', 'Enable Remote Leaf Direct Traffic Forwarding', 'Opflex Client Authentication', and 'Reallocate Gipo'.

System Settings

- Quota
- APIC Connectivity Preferences
- System Alias and Banners
- System Response Time
- Global AES Passphrase Encryption Settings
- BD Enforced Exception List
- Fabric Security
- Control Plane MTU
- Endpoint Controls
- Fabric-Wide Settings**
- Remote Leaf POD Redundancy Policy

Fabric-Wide Settings Policy

Policy History

Properties

- Disable Remote EP Learning: To disable remote endpoint learning in VRFs containing external bridged/routed domains
- Enforce Subnet Check: To disable IP address learning on the outside of subnets configured in a VRF, for all VRFs**
- Enforce EPG VLAN Validation: Validation check that prevents overlapping VLAN pools from being associated to an EPG
- Enforce Domain Validation: Validation check if a static path is added but no domain is associated to an EPG
- Enable Remote Leaf Direct Traffic Forwarding: Enable Remote Leaf direct communication with routable IP connectivity between Remote Leafs and Fabric, once enabled you cannot disable it again
- Opflex Client Authentication: To enforce Opflex client certificate authentication for GOLF and Linux
- Reallocate Gipo: Reallocate some non-stretched BD gipos to make room for stretched BDs

EP Loop Protection

- System > System Settings > Endpoint Controls > EP Loop Protection

The screenshot displays the Cisco system settings interface. The top navigation bar includes 'System', 'Tenants', 'Fabric', 'Virtual Networking', 'L4-L7 Services', 'Admin', 'Operations', 'Apps', and 'Integrations'. The 'System Settings' section is expanded, showing a list of settings including 'Quota', 'APIC Connectivity Preferences', 'System Alias and Banners', 'System Response Time', 'Global AES Passphrase Encryption Settings', 'BD Enforced Exception List', 'Fabric Security', 'Control Plane MTU', 'Endpoint Controls', 'Fabric-Wide Settings', and 'Remote Leaf POD Redundancy Policy'. The 'Endpoint Controls' section is highlighted. The main content area shows the 'Endpoint Controls' configuration page, with 'Ep Loop Protection' selected. The 'Administrative State' is set to 'Disabled'. The 'Loop Detection Interval' is 60, and the 'Loop Detection Multiplication Factor' is 4. The 'Action' section has 'BD Learn Disable' unchecked and 'Port Disable' checked.

System Settings

Endpoint Controls

Ep Loop Protection

Rogue EP Control

Ip Aging

Policy

History

Properties

Administrative State: Disabled Enabled

Loop Detection Interval: 60

Loop Detection Multiplication Factor: 4

Action: BD Learn Disable Port Disable

Rogue EP Control

- System > System Settings > Endpoint Controls > Rogue EP Control

The screenshot displays the Cisco system settings interface. The top navigation bar includes 'System', 'Tenants', 'Fabric', 'Virtual Networking', 'L4-L7 Services', 'Admin', 'Operations', 'Apps', and 'Integrations'. The 'System Settings' menu is expanded, showing options like 'Quota', 'APIC Connectivity Preferences', 'System Alias and Banners', 'System Response Time', 'Global AES Passphrase Encryption Settings', 'BD Enforced Exception List', 'Fabric Security', 'Control Plane MTU', 'Endpoint Controls', 'Fabric-Wide Settings', and 'Remote Leaf POD Redundancy Policy'. The 'Endpoint Controls' section is active, showing 'Rogue EP Control' under 'Ep Loop Protection'. The 'Administrative State' is set to 'Disabled'. The 'Rogue EP Detection Interval' is 60, the 'Rogue EP Detection Multiplication Factor' is 4, and the 'Hold Interval (sec)' is 1800.

System Settings

Endpoint Controls

Ep Loop Protection

Rogue EP Control

Ip Aging

Policy

History

Properties

Administrative State: Disabled Enabled

Rogue EP Detection Interval: 60

Rogue EP Detection Multiplication Factor: 4

Hold Interval (sec): 1800

IP Aging Policy

- System > System Settings > Endpoint Controls > IP Aging

The screenshot displays the Cisco SD-WAN management interface. The top navigation bar includes 'System', 'Tenants', 'Fabric', 'Virtual Networking', 'L4-L7 Services', 'Admin', 'Operations', 'Apps', and 'Integrations'. The 'System' menu is expanded, showing 'System Settings' as the selected option. Under 'System Settings', the 'Endpoint Controls' menu item is highlighted. The main content area shows the 'Endpoint Controls' configuration page. The 'Ip Aging' tab is selected, and the 'Administrative State' is set to 'Enabled'. The 'Policy' and 'History' tabs are also visible.

System Settings

- Quota
- APIC Connectivity Preferences
- System Alias and Banners
- System Response Time
- Global AES Passphrase Encryption Settings
- BD Enforced Exception List
- Fabric Security
- Control Plane MTU
- Endpoint Controls**
- Fabric-Wide Settings

Endpoint Controls

Ep Loop Protection Rogue EP Control **Ip Aging**

Policy History

Administrative State: Disabled **Enabled**

まとめ

Local Endpoint Learning

- デフォルトではData-planeでIPアドレス"も"学習する点が、ACIならではの動作

| フレーム種別 | 転送動作種別 | 学習対象 | |
|------------------|-------------------------|--|---------------------------------|
| Non-IP | Bridged (L2) | MACアドレス (sMAC) | |
| ARP | | MACアドレス (Sender MAC) IPアドレス (Sender IP) | ※宛先がBD GW (BD-SVI)宛の場合 |
| IPv4 | Unicast Routed (L3) | MACアドレス (sMAC) IPアドレス (sIP) | } Data-plane IP Learning |
| IPv6 | Unicast Routed (L3) | MACアドレス (sMAC) IPアドレス (sIP) | |
| IPv6 / ND | Neighbor Discovery (ND) | MACアドレス (Source MAC) IPアドレス (Source IP) | |

※L3out と Infra VLAN で受信したフレームからはIPアドレスの学習は行わない

