ACI Design #8 – Endpoint Learning

Cisco Systems G.K.
2018/1

Endpoint Learning 基本編

ACIにおけるEndpoint学習動作

従来のネットワークにおける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)。

Table	含まれる情報	概要
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1つもしくは複数 IPアドレスは /32 (IPv4) もしくは /128 (IPv6) でのみ学習する
ARP	L3out 接続については、ARPに基づいて MACアドレスに紐づくIPアドレスを学習する	

デフォルトでは、ACIはARPに依存せずに、Leafスイッチに届いたパケットの送信元
MACアドレス・IPアドレスを元にEndpointを学習する。
(DownlinkからはLocal Endpoint、UplinkからはRemote Endpointとして扱う)

ACIにおける外部アドレス学習動作

ACI範囲外の外部ネットワークとの接続においては、従来通りRIBテーブルと
ARPテーブルを利用したアドレス解決を行う (Endpointテーブルは利用しない)。
※EndpointテーブルはMACアドレスに紐付け可能なIPアドレス数に制限がある (3.1では1024まで)

また、内部Endpointとは異なり、外部ネットワークについては /32 (IPv4)や /128 (IPv6)でのアドレス管理は行わず、SubnetベースのPrefixに基づく宛先管理を行う。

ACIは宛先Prefixへ到達するためのNext-Hopのみ学習するために、従来スイッチと同じくデータプレーンからMACアドレスの学習を行い、ARPによりMACアドレスとIPアドレスの解決を行う(従来スイッチと同じ動作)。

Local Endpoint と Remote Endpoint

各Leafスイッチは、自身に直接接続したEndpoint (=Local Endpoint) を学習し、Local Endpointの通信相手となった別Leaf (TEP)に接続したEndpoint (=Remote Endpoint)をキャッシュ用として学習する。

Endpoint	学習対象	補足
Local Endpoint	MACアドレスとIPアドレス (IPアドレスの学習は構成次第) ※1 MAC + 必要数IP を紐付けて管理	COOPによるSpineスイッチへの通知対象 (Spineスイッチは全てのEndpointのLeaf配置を管理) ・Default Retention Timer：900秒 ※IPアドレスは個別にAging処理される
Remote Endpoint	単一のMACアドレスか単一のIPアドレス (どちらを学習するかはiVXLANに含まれる情報に基づく[BD=MAC, VRF=IP])	キャッシュ用途(Spine Proxyの負荷軽減と最適化) Ingress側でのPolicy適用(Fabric通信の最適化) ・Default Retention Timer：300秒

なぜ Spine Proxy があるのに Remote Endpoint を学習する?

各Leafスイッチは、Local Endpointが通信相手とするRemote Endpointの情報をキャッシュすることにより、都度Spine Proxyに問い合わせる時間と動作を削減する。

キャッシュしたRemote Endpoint宛の通信は、Spine Proxyへの問い合わせを行わずに直接宛先Leafスイッチに送信する。学習済の相手に対する通信は送信元側のLeafスイッチでPolicyを適用することが可能(Ingress Policy Enforcement)。

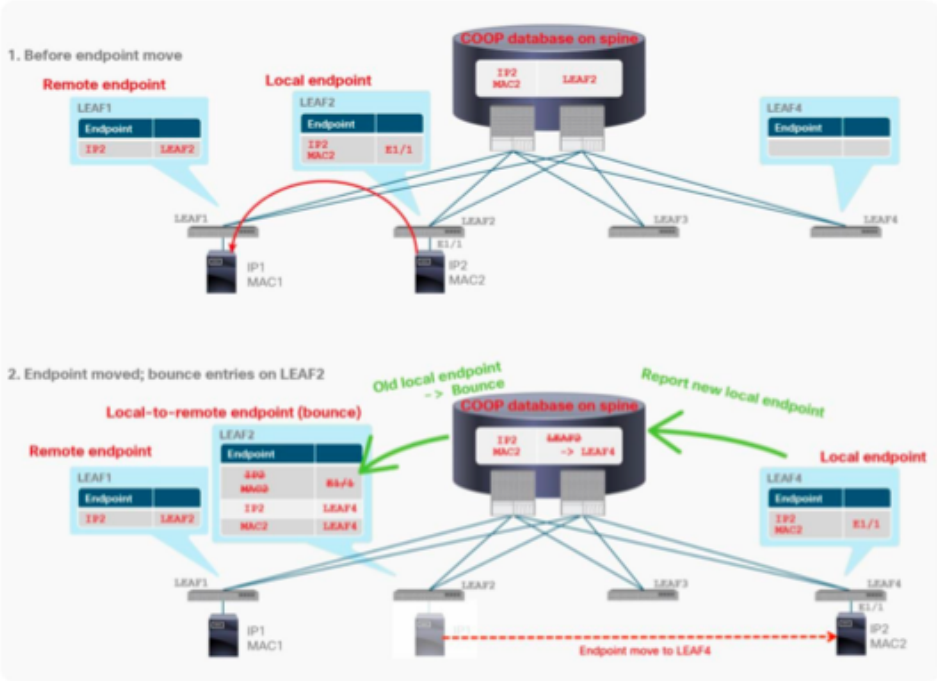
Remote Endpointについては、Local Endpointの様にMACアドレスとIPアドレスの紐付けは管理せず、宛先IPアドレスもしくはMACアドレスを個別に宛先Leafと紐付ける(L2通信相手ならMACアドレスのみ、L3通信相手ならIPアドレスのみ)。

※ただしvPC Peer同士はOrphan Portを含む全てのEndpointの情報を同期するため動作が異なる

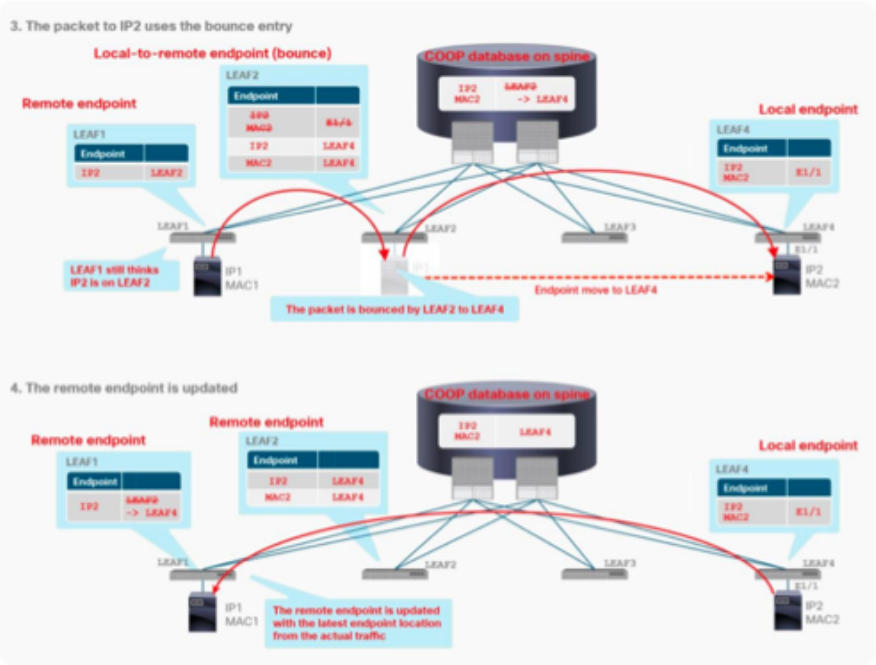
Endpointの移動とBounce Entry (1)

EndpointがLeafスイッチをまたいで
移動した場合、移動先Leafスイッチは
当該Endpointの情報をSpineスイッチに
通知し、Spineスイッチは更新情報を
移動元Leafスイッチにのみ通知する。

この時点では、移動元・移動先のLeaf
スイッチ以外のLeafスイッチは
当該Endpointの移動を知らない。
→キャッシュに基づいて当該Endpoint宛
通信を移動元に送信する可能性がある



Endpointの移動とBounce Entry (2)



移動元Leafスイッチは移動先Leafスイッチを知っているため、当該パケットを移動先Leafに転送する(Bounce)。

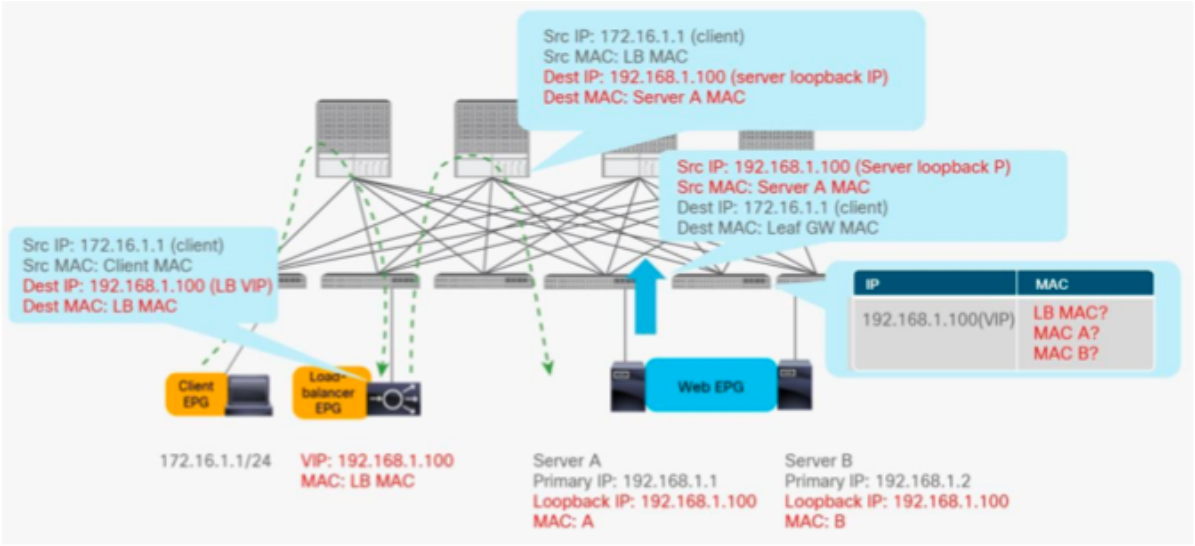
戻りの通信については、移動先Leafは送信元Leafに直接返信するため、送信元Leafはこの時点で当該Endpointの移動を認識し、自身のキャッシュを更新する。

Endpoint移動時にFabricサイズに依存しないSpine処理を実現し、必要Leafのみ必要な時点で情報を更新する為の仕組み

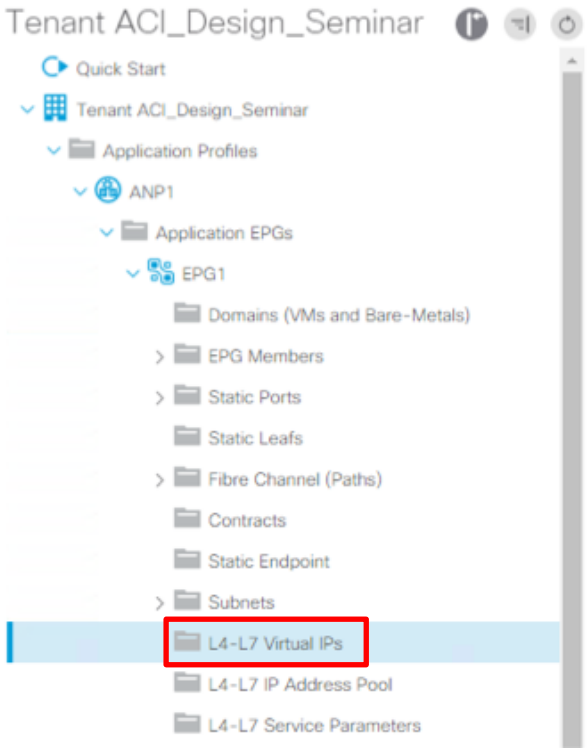
Endpoint Learning
パラメータチューニング編

EPG - L4-L7 Virtual IPs

LBのDirect Server Return (L2 DSR)動作のために、DSR用仮想IPアドレスについて
Endpoint Data-plane Learningを無効化する(フラップ抑制)。



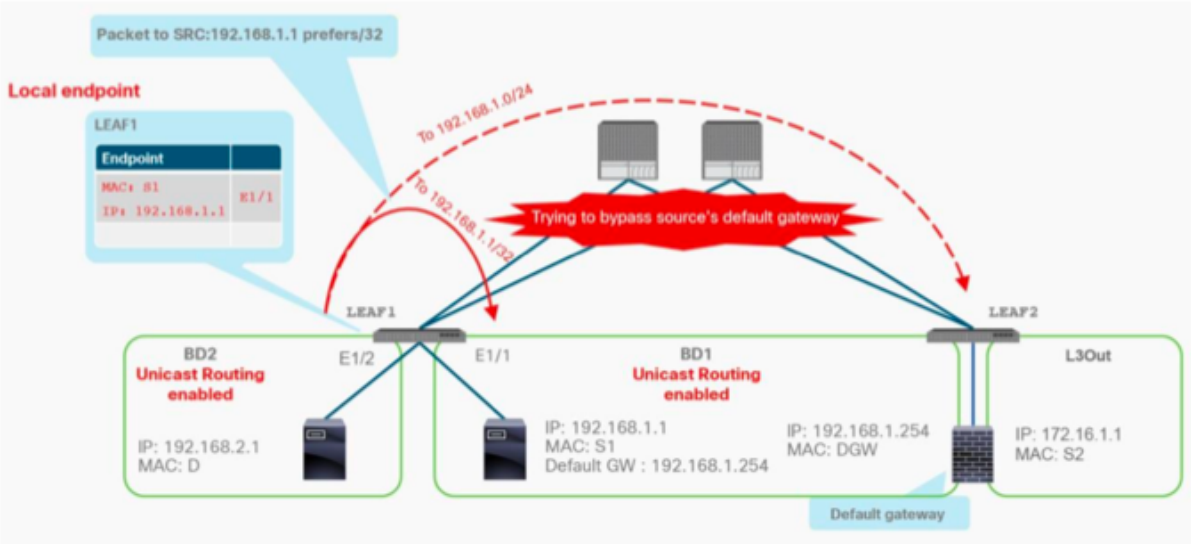
EPG - L4-L7 Virtual IPs



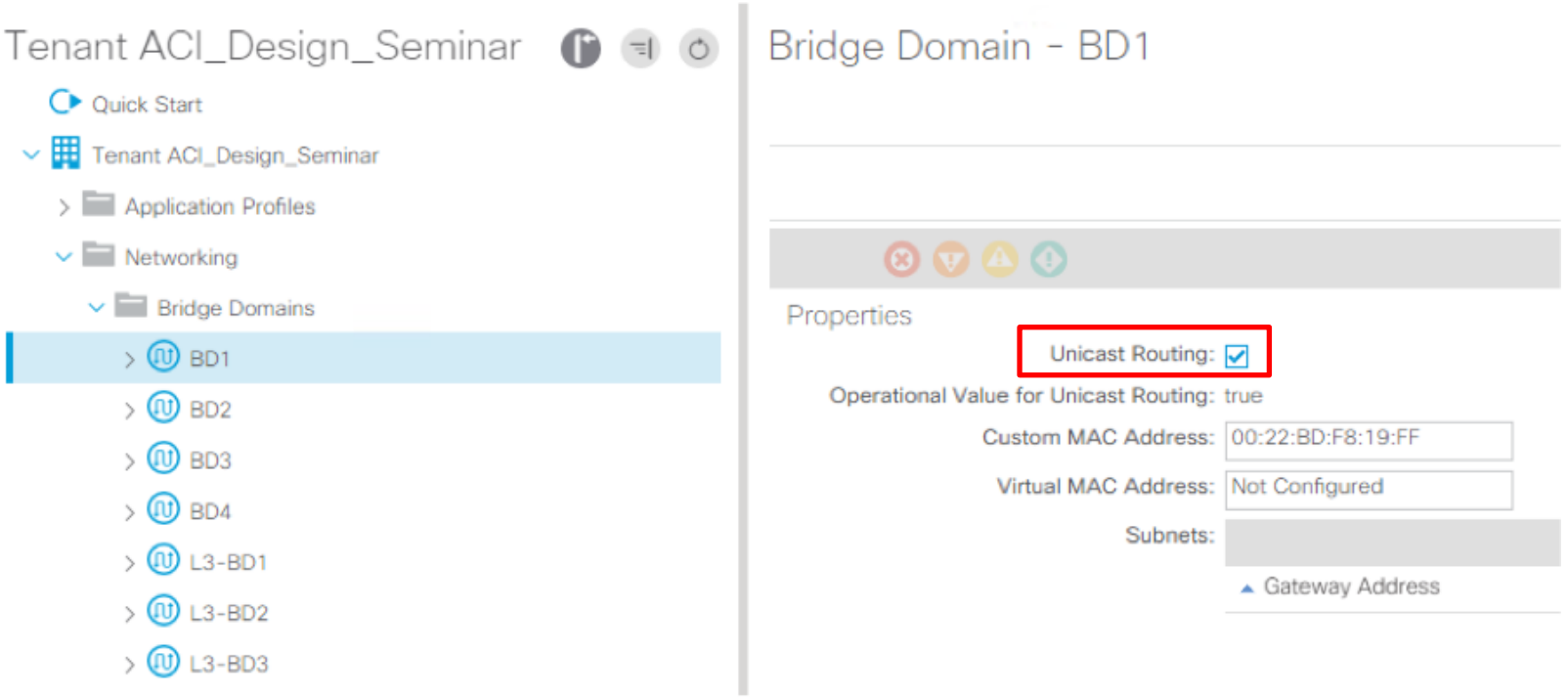
BD – Unicast Routing

BDに対して構成したSubnet設定をLeafスイッチに適用する。
(Routing動作およびEndpointのIPアドレス学習の両方に対して機能する)

BDをL2利用する場合は、
想定外のGWバイパス
動作の発生に注意する。



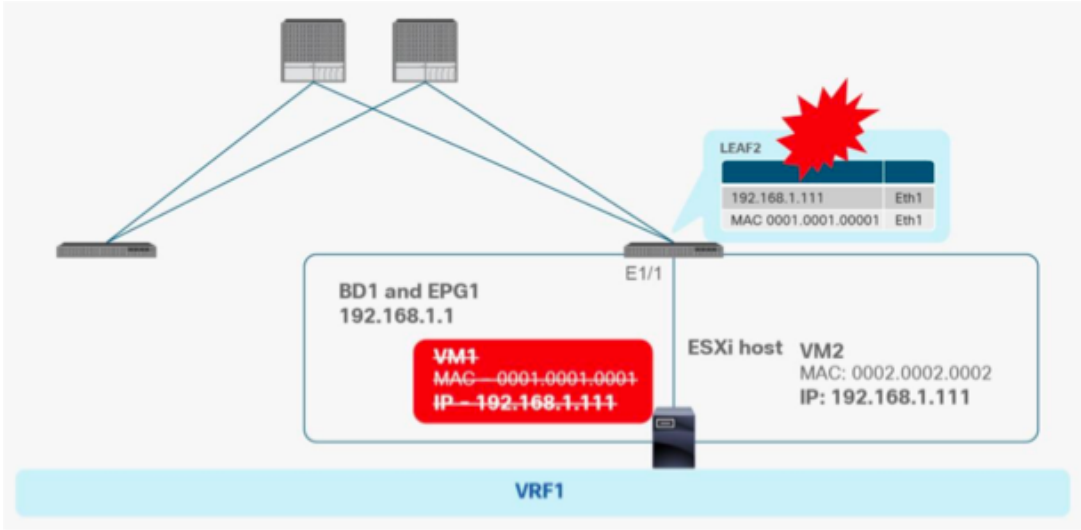
BD – Unicast Routing



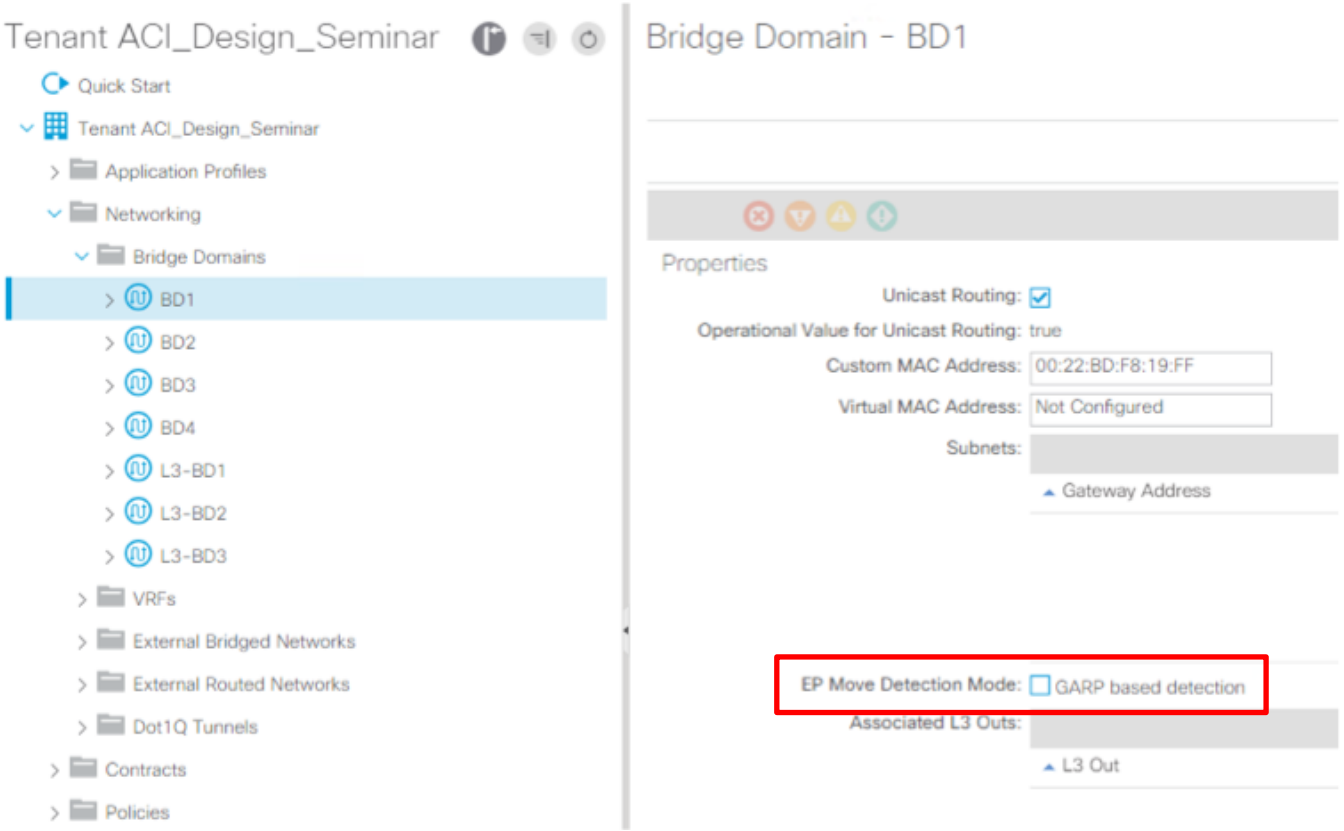
BD – GARP-based EP Move Detection

Unicast Routing有効・ARP Flooding有効に加えて有効化することで、同一ポートの配下でIPアドレスが新しいMACアドレスに紐付いたGARP通知に基づいたEndpointの学習を行う。

LeafスイッチがVM1からの通信で一旦学習したIPアドレスを別のVMが同一ホスト上で再利用した場合などに必要となる。



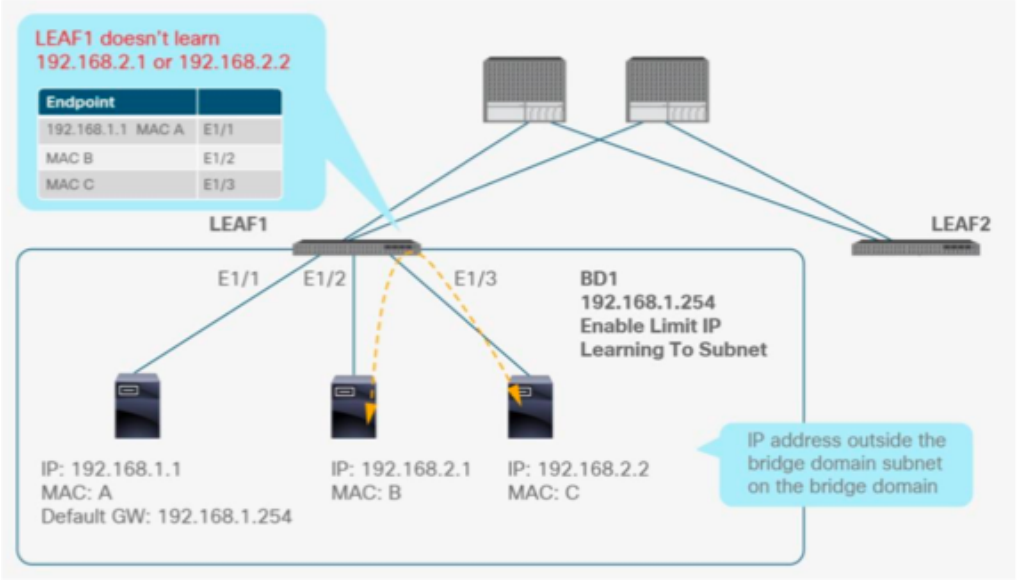
BD – GARP-based EP Move Detection



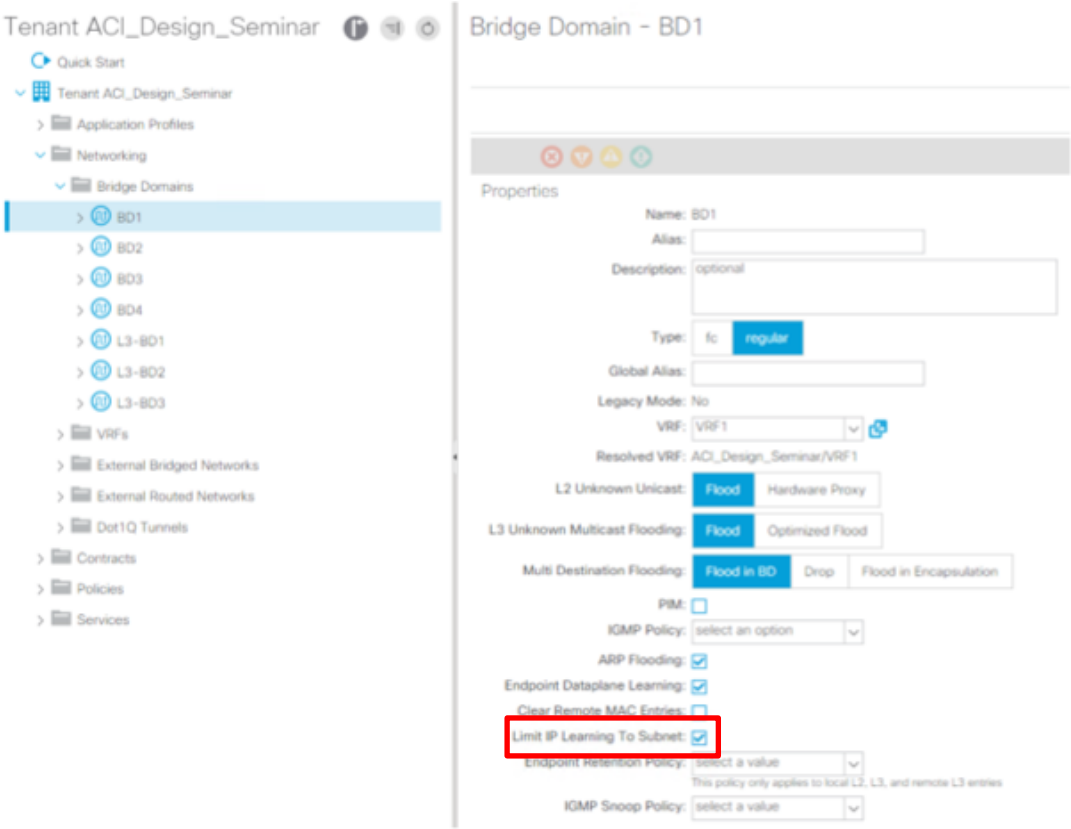
BD – Limit IP Learning To Subnet
(旧名称：Enforce Subnet Check for IP Learning)

- 2.3(1e)および3.0(1k)以降で新規に作成したBDではデフォルト有効(CSCvb16668)
- 3.0(1k)以前のバージョンで有効・無効を切り替える場合はEndpoint学習が120秒間停止するので注意(CSCve29663)

誤ったIPアドレスを設定した場合に当該IPアドレスの学習を抑制する。
(Local Endpointのみ)

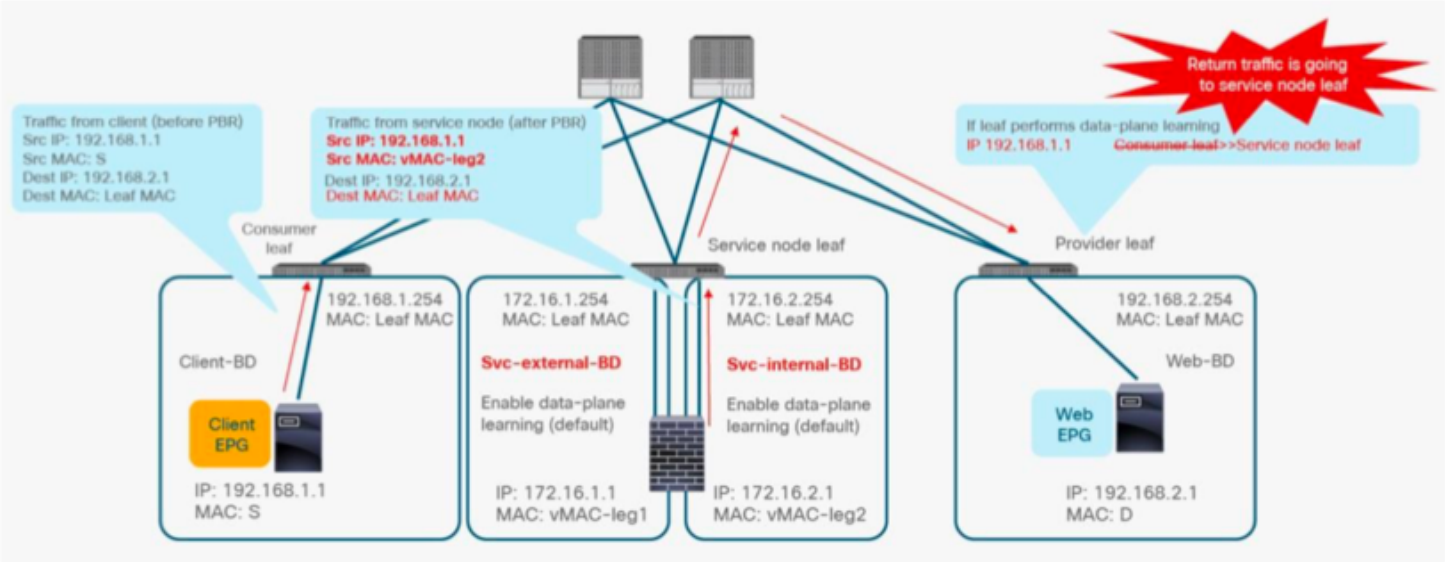


BD – Limit IP Learning To Subnet
(旧名称：Enforce Subnet Check for IP Learning)

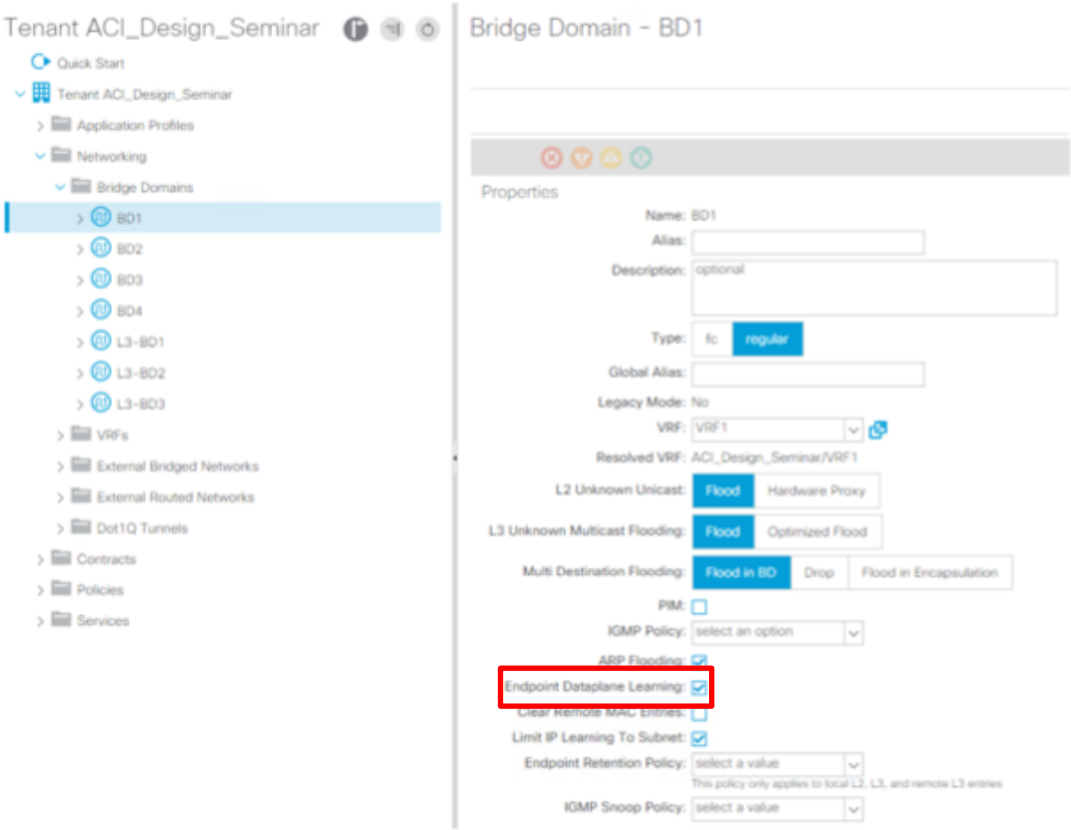


BD – Endpoint Dataplane Learning

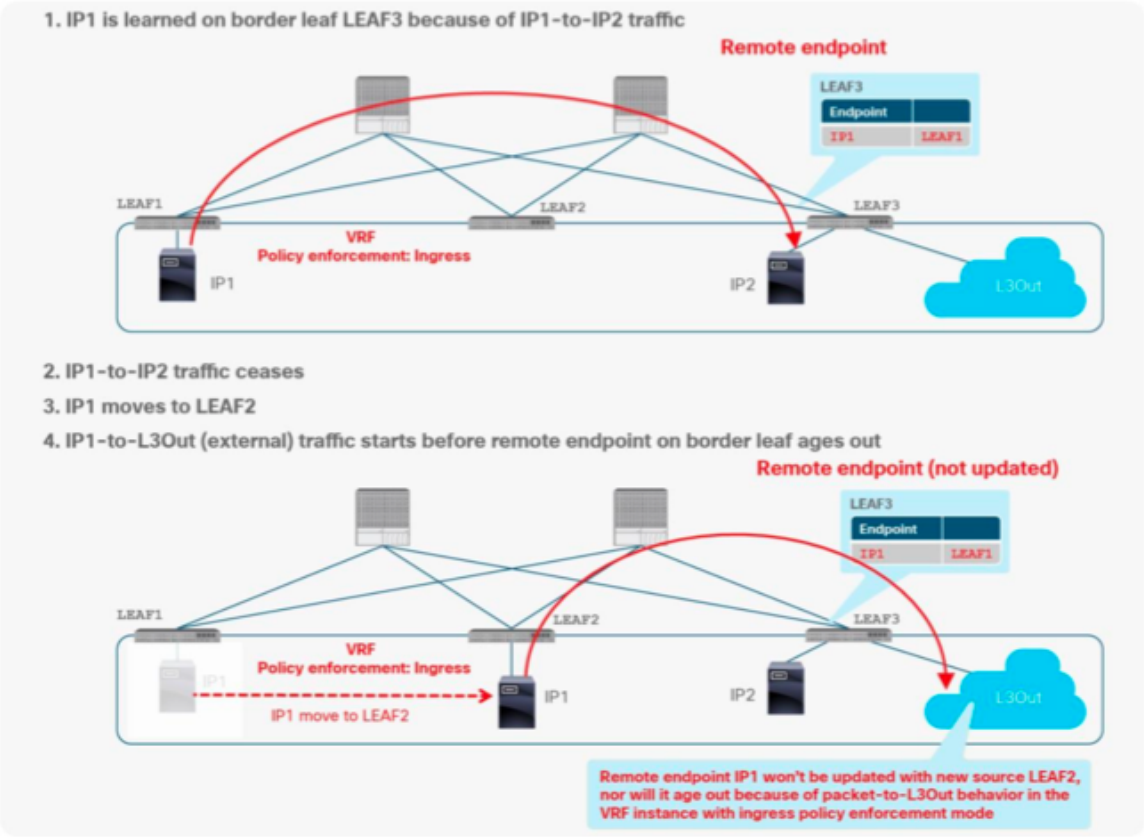
PBRで利用するService Nodeを接続するためのBDでは、Dataplane Learningは無効化する必要がある。



BD – Endpoint Dataplane Learning



Fabric - Disable Remote EP Learn (on border Leaf) ①



Endpoint接続LeafをBorder Leafとしても利用する注意点

Leaf1でBounce Entryが期限切れとなった後にLeaf3からIP1宛の通信が失敗する(Retention Timer経過時間まで)

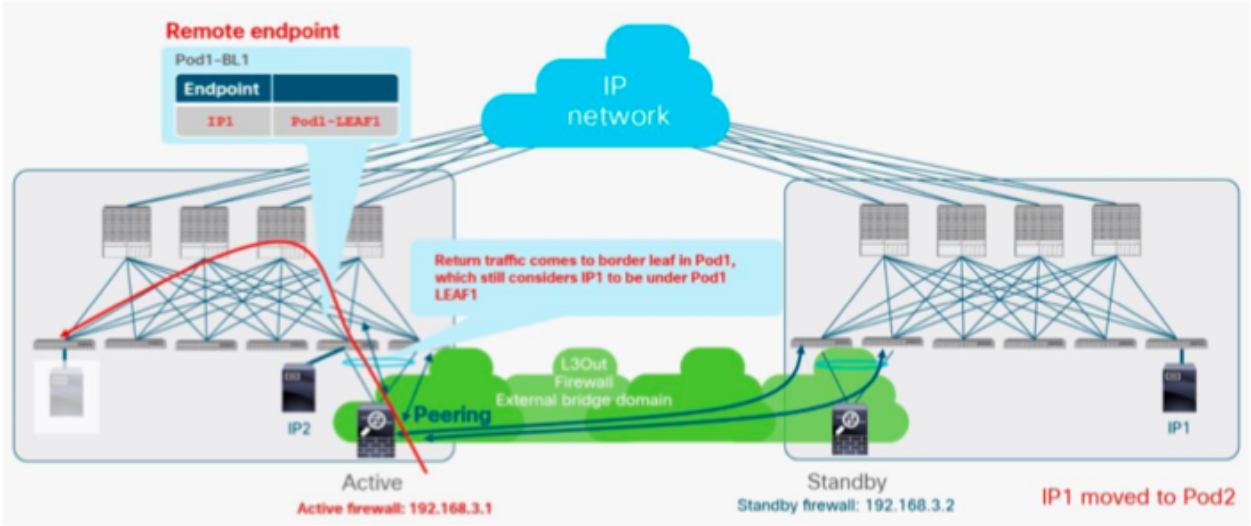
Leaf3でRemote Endpoint情報を削除することにより事象は解決する(要CLI操作)

Leaf3がBorder Leaf専用である場合や、Disable Remote EP Learnが構成されている場合も防止できる

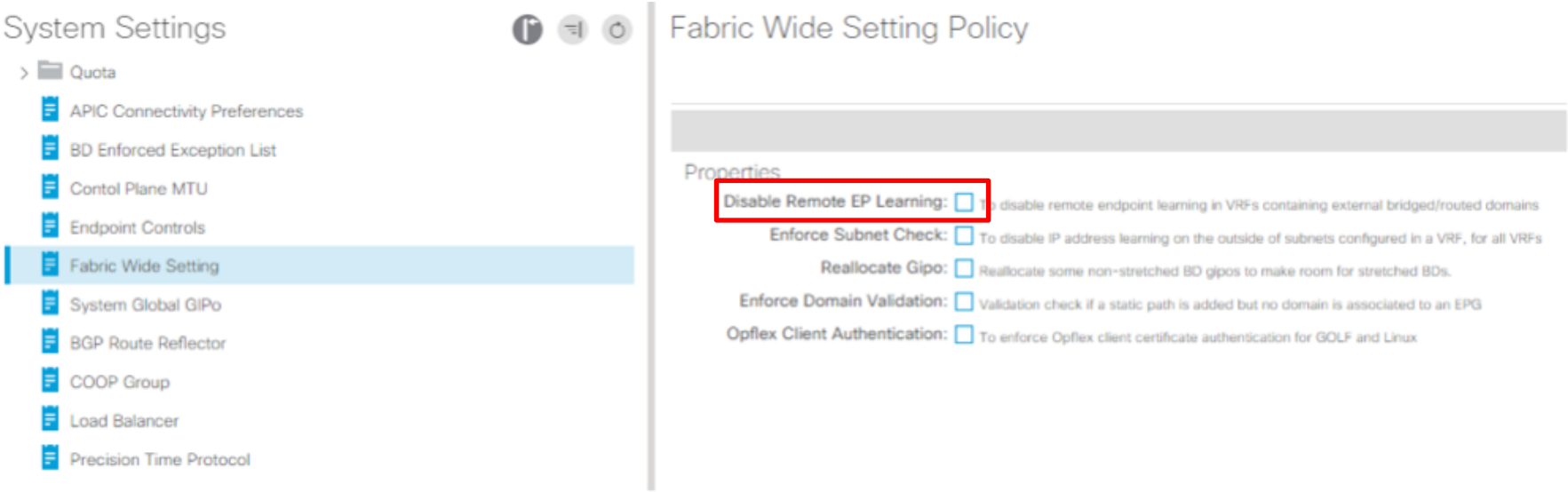
Fabric – Disable Remote EP Learn (on border leaf) ②

複数のLeafスイッチに渡って同一VLAN SVIでL3outを構成する場合に構成する。

例：Multi-PodでL3outをActive/Standby構成の環境でEndpointが移動した場合



Fabric - Disable Remote EP Learn (on border Leaf)

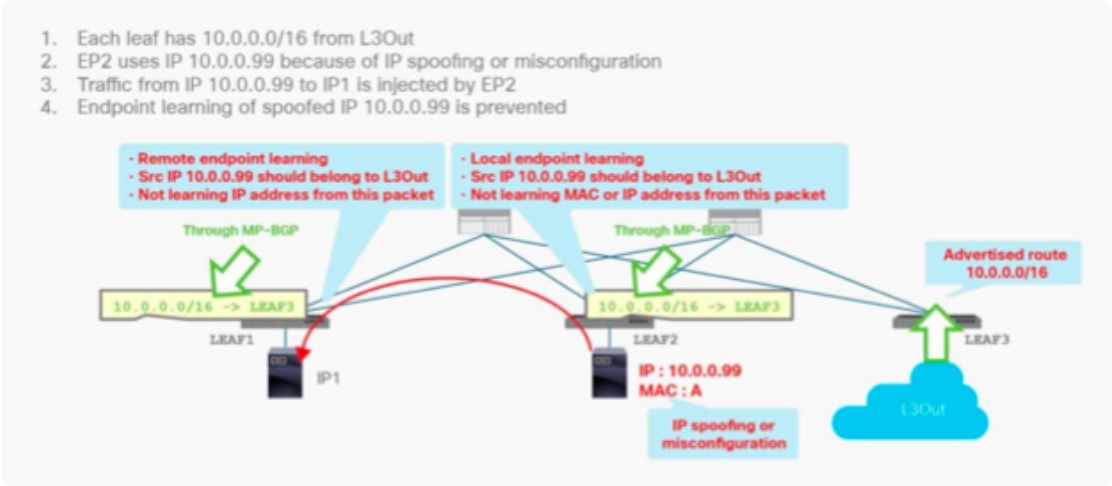


Fabric - Enforce Subnet Check ①

外部Route範囲のIPアドレスを設定したEndpointがパケットを誤送信した場合、L3outから受信している経路情報がPrefixの場合は登録は防止される(第2世代Leafスイッチの場合のみ)が、0.0.0.0/0範囲だった場合には防止できない。

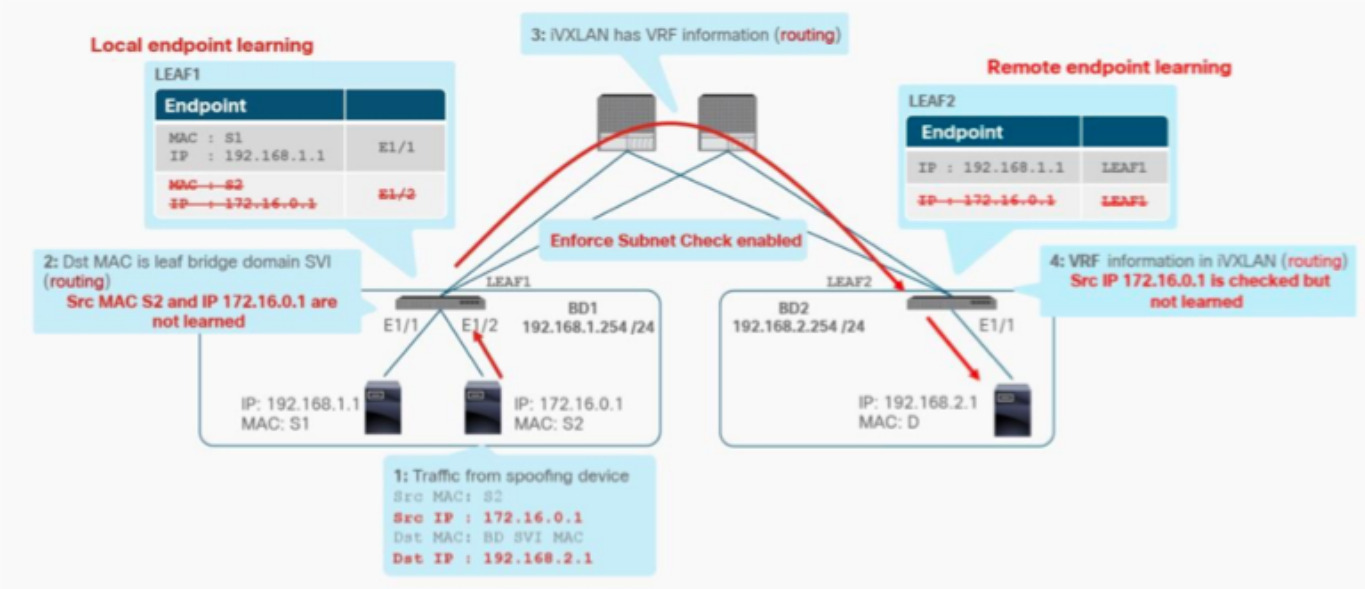
第2世代Leafスイッチの場合は[Enforce Subnet Check]を有効化して防止する。

第1世代Leafスイッチの場合は[Limit IP Learn To Subnet]と[Disable Remote EP Learn]を併用設定する。

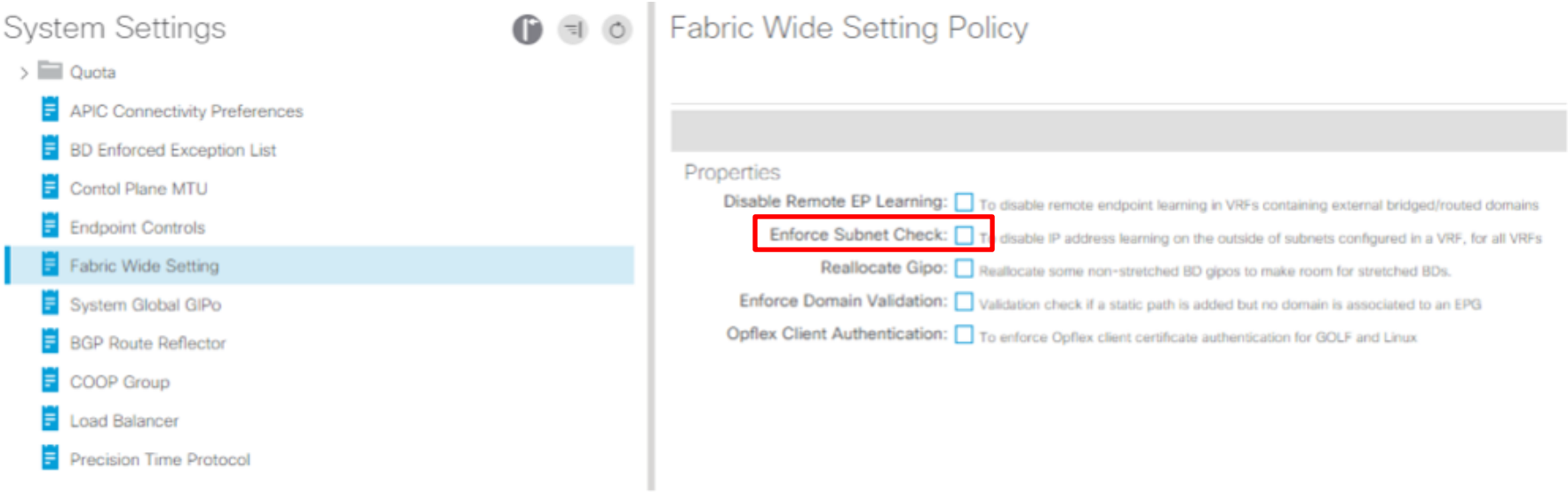


Fabric – Enforce Subnet Check ②

Limit IP Learning To Subnetよりも強力に、Local Endpointだけでなく、Remote Endpointについても誤学習を防止できるが、第2世代Leafスイッチに対してのみ有効。

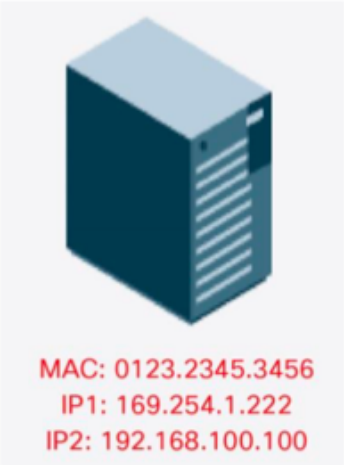


Fabric – Enforce Subnet Check



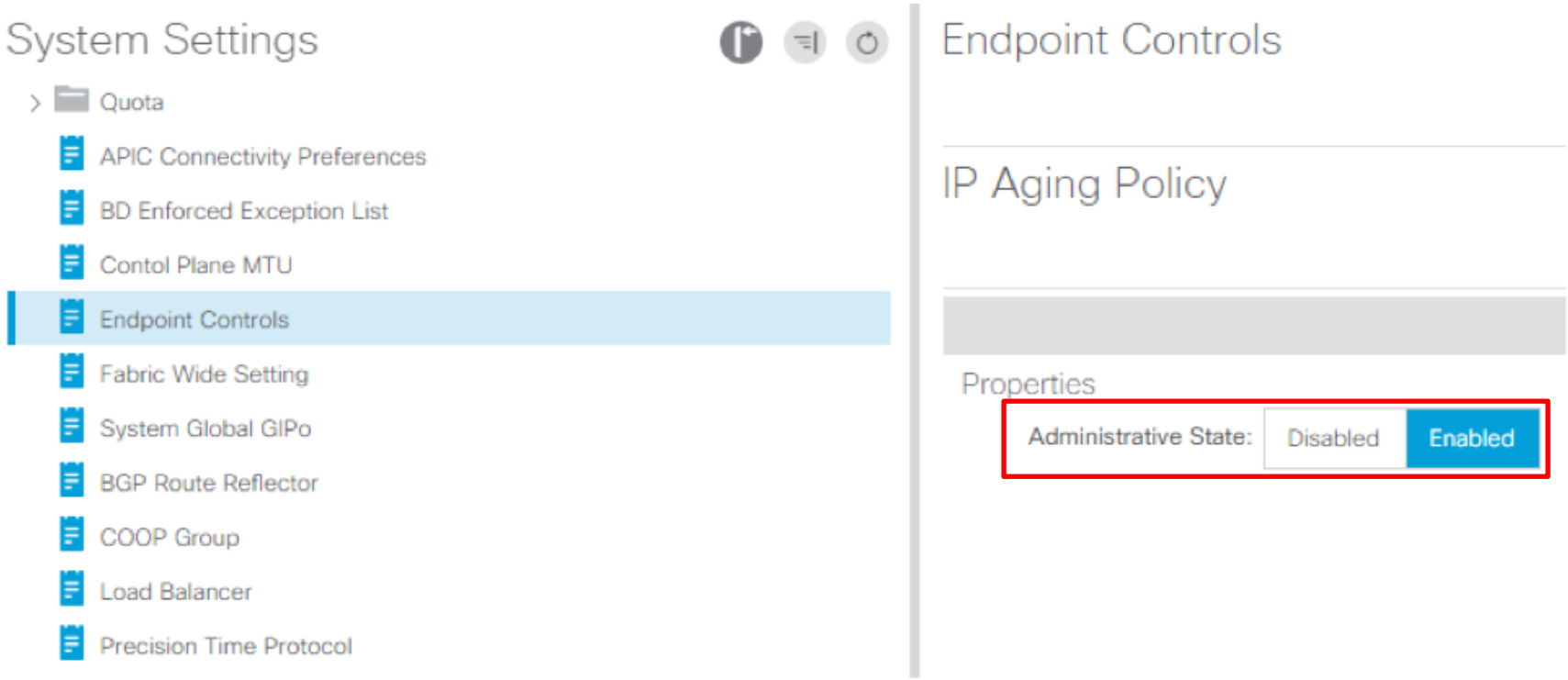
Fabric – IP Aging

不要になったIPアドレス情報をいつまでも保持することを防止する。
(DHCPでIPアドレスを取得できなかった際の169.254.x.xアドレスなど)



Aging指定時間の75%が経過した時点で、Leafスイッチは
UnicastでARPパケットを送出し、IPアドレスの存在確認を行う
(応答がない場合は当該IPアドレスエントリが削除される)

Fabric – IP Aging



推奨チューニング設定

ACIにおけるEndpointの学習動作を正しく理解した上で、適切な構成を行う。

		第1世代 Leafスイッチのみの ACI Fabric	第2世代 Leafスイッチのみの ACI Fabric	第1世代・第2世代 Leafスイッチ混在の ACI Fabric
BD	Limit IP Learning To Subnet	有効	有効	有効
Fabric	IP Aging	有効 (Default)	有効 (Default)	有効 (Default)
	Disable Remote EP Learn (on border leaf)	有効	有効	有効
	Enforce Subnet Check		有効	有効

主なEndpoint学習動作に関連する機能の実装バージョン

機能	実装バージョン	概要
L4-L7 Virtual IP	1.2(1m)	特定IPアドレスのData-plane Learningを無効化する (DSR動作に対応するため)
Unicast Routing	1.0(1e)	BDにおけるL3ルーティングとIP Learning動作
GARP-based EP Move Detection	1.1(1j)	ARP FloodingがEnable時に同一Interface配下における Endpointの移動の検出にGARPを利用する (Endpoint DBにおける IP-MAC紐付け管理)
Limit IP Learning To Subnet	1.1(1j)	BD構成Subnet範囲外のIP学習の防止 (Local Endpointのみ)
Endpoint Dataplane Learning	2.0(1m)	BDにおけるEndpoint Data-plane Learningを無効化する (PBR動作に対応するため)
Disable Remote EP Learn (on border leaf)	2.2(2e)	Ingress Policy Enforcement (Default) 動作のBorder LeafにおけるRemote Endpointの学習を無効化する (対象Border LeafスイッチはSpine Proxyのみ利用する)
Enforce Subnet Check	2.2(2q) ※2.3, 3.0(1x)では 利用不可	VRFに紐付いたBD Subnet範囲外のEndpoint学習を防止する (Local / Remote 両方) ※第2世代Leafスイッチに対してのみ有効
IP Aging	2.1(1h)	学習済のEndpointに紐付いた未使用IPアドレスの開放 ※2.1(1h)以降ではデフォルト有効、それ以前からの場合デフォルト無効

詳しくは・・・

ACI Fabric Endpoint Learning ホワイトペーパーを参照ください!!

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


CISCO