Cisco
Customer Experience

Ask the Experts

Cisco ACI 向け 分散ネットワーキングのパフォーマンス チューニング ベストプラクティス (Performance Tuning Best Practices Distributed Networking for Cisco ACI)

2024年3月19日

Disclaimer

This document is Cisco Confidential information provided for your internal business use in connection with the Cisco Services purchased by you or your authorized reseller on your behalf. This document contains guidance based on Cisco's recommended practices.

You remain responsible for determining whether to employ this guidance, whether it fits your network design, business needs, and whether the guidance complies with laws, including any regulatory, security, or privacy requirements applicable to your business.

免責

この文書は、お客様またはお客様の代理人である認定リセラーが購入したシスコサービスに関連して、お客様が社内業務において使用することを目的としてシスコが提供するシスコの機密情報です。この文書にはシスコが推奨するプラクティスに基づく手引きが記載されています。

お客様は、この手引きを使用するか否かやお客様のネットワーク設計および業務上のニーズにこの手引きが適合しているか否か、さらにはこの手引きが法律（お客様の業務に適用される規制上の要件、セキュリティ上の要件およびプライバシーに関する要件を含みます）に準拠しているか否かを判断する責任を引き続き負います。

セッション対象者

- Cisco ACI Multi-Pod / Multi-Site 環境を構築予定または構築済み
- Cisco ACI を使った分散ネットワーキングについてより理解を深めたい
- ACI 分散ネットワーキングのパフォーマンスチューニング方法について知りたい

本日のトピック

01 | コントロールプレーン チューニングのベストプラクティス

02 | データプレーン チューニングのベストプラクティス

03 | Remote Leaf の考慮事項

04 | デモ

コントロールプレーン チューニングの ベストプラクティス

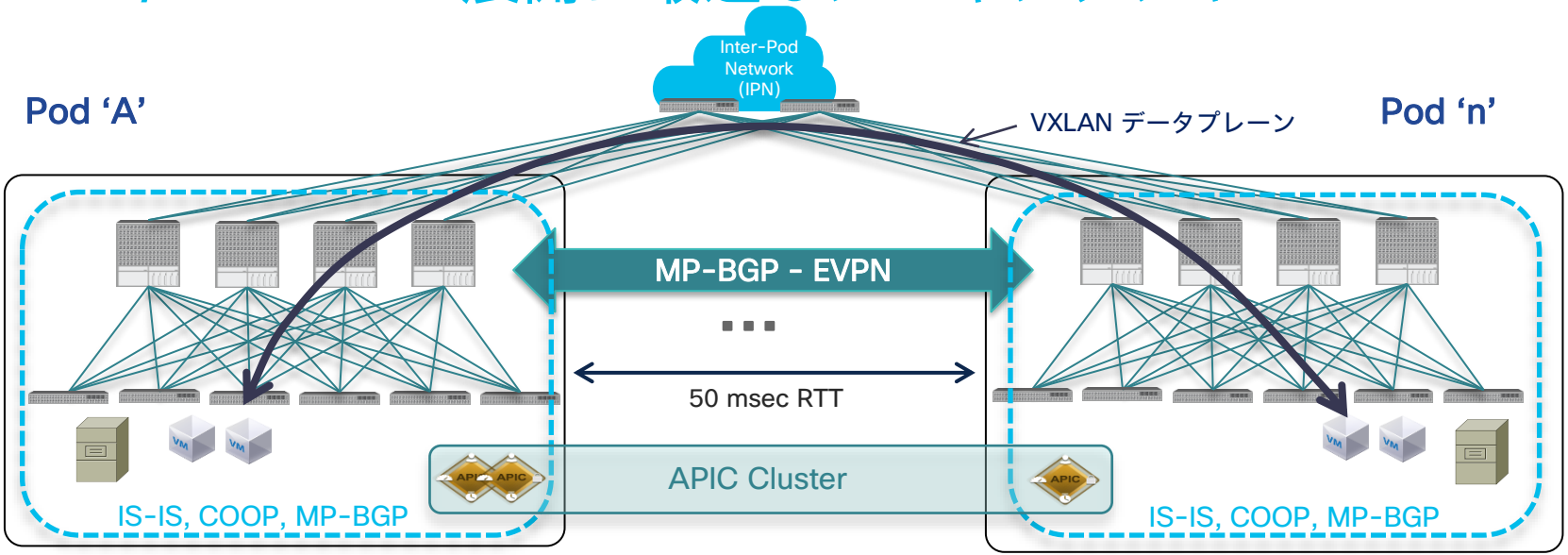
シングル Pod のチューニング項目

- Forwarding Scale Profile
  - IPv4 Endpoint 数や IPv6 Endpoint 数、Policy CAM などのリソースの配分を変更できる Profile
- Fast Link Failover
  - ファブリックリンク障害時のフェールオーバーコンバージェンスを向上させる機能
  - ハードウェアテーブルの更新とコントロールプレーンのコンバージェンスが高速化する
- Debounce Timer
  - リンクダウンイベント通知からハードウェアテーブル更新までの時間
  - デフォルト値が 100ms で最小 10ms に調整可能

Forwarding Scale Profile Policy Options	ToR Switches with EX and FX2 Names	ToR Switches with FX Names
Dual Stack	• EP MAC: 24,000 • EP IPv4: 24,000 • EP IPv6: 12,000 • LPM: 20,000 • Policy: 64,000 • Multicast: 8,000	Has the same scalability numbers as Dual Stack scale on earlier switches.
High Dual Stack	• EP MAC: 64,000 • EP IPv4: 64,000 • EP IPv6: 24,000 • LPM: 38,000 • Policy: 8,000 • Multicast: 512	• EP MAC: 64,000 • EP IPv4: 64,000 • EP IPv6: 48,000 • LPM: 38,000 • Policy: 128,000 • Multicast: 32,000
High LPM	Provides scalability similar to the dual-stack profile, except that the longest prefix match (LPM) scale is 128,000 and the policy scale is 8,000.	Has the same scalability numbers as on earlier switches.
IPv4 Scale	• EP MAC: 48,000 • EP IPv4: 48,000 • EP IPv6: 0 • LPM: 38,000 • Policy: 64,000 • Multicast: 8,000	Has the same scalability numbers as IPv4 scale on earlier switches.

ACI Multi-Pod

Active/Active DC 展開に最適なアーキテクチャ

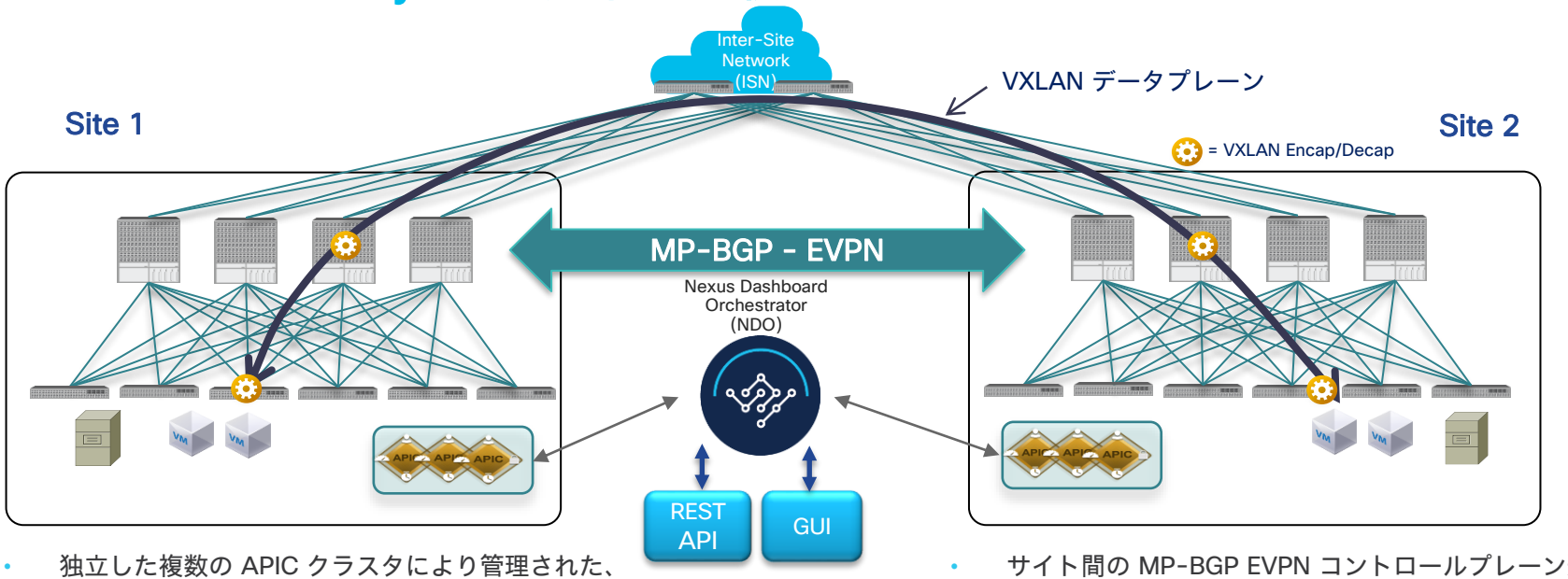


- L3 Inter-Pod Network で接続された複数の ACI Pod; 各 Pod は Leaf と Spine で構成される
- 1 つの APIC クラスタで管理
- 単一の管理およびポリシー領域

- コントロールプレーン(IS-IS, COOP, MP-BGP) は障害分離される
- Pod 間のデータプレーンは VXLAN でカプセル化
- End-to-end でポリシーを適用

ACI Multi-Site

Active/Standby DC 展開に最適なアーキテクチャ



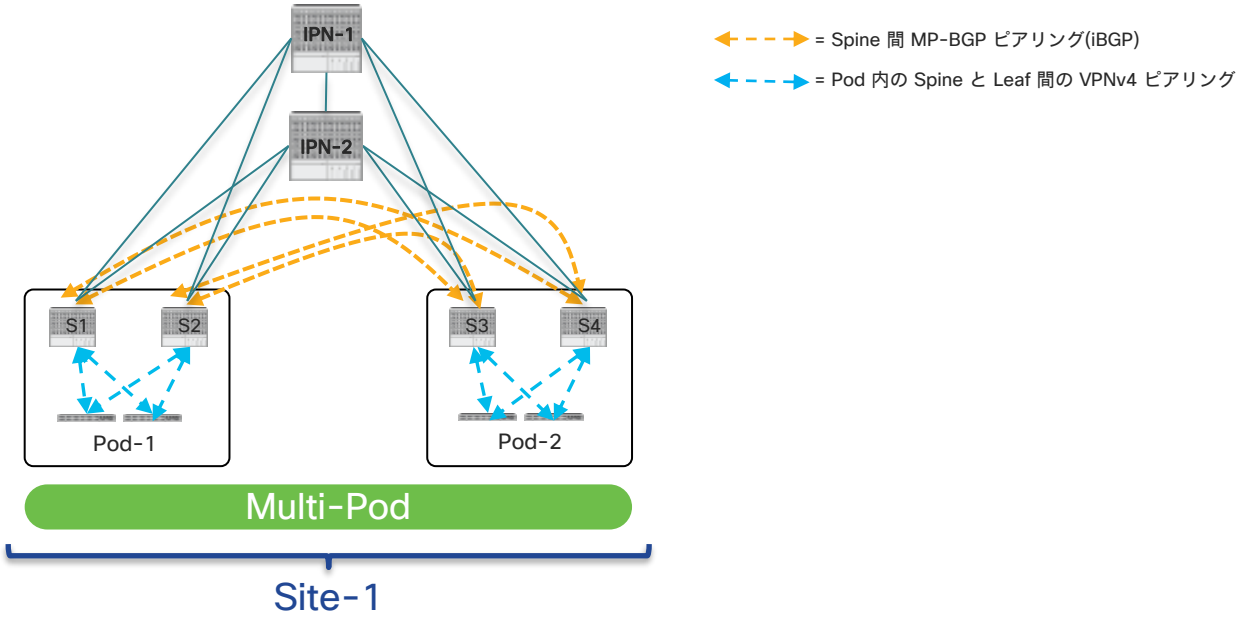
- 独立した複数の APIC クラスタにより管理された、異なる ACI ファブリック
- NDO が、ファブリックをまたがる構成を複数の APIC クラスタにプッシュし、すべての構成変更を統合管理

- サイト間の MP-BGP EVPN コントロールプレーン通信
- サイト間のデータプレーン通信は VXLAN カプセル化
- End-to-end でポリシーを適用
- サイト間の遅延制限なし

コントロールプレーンの違い

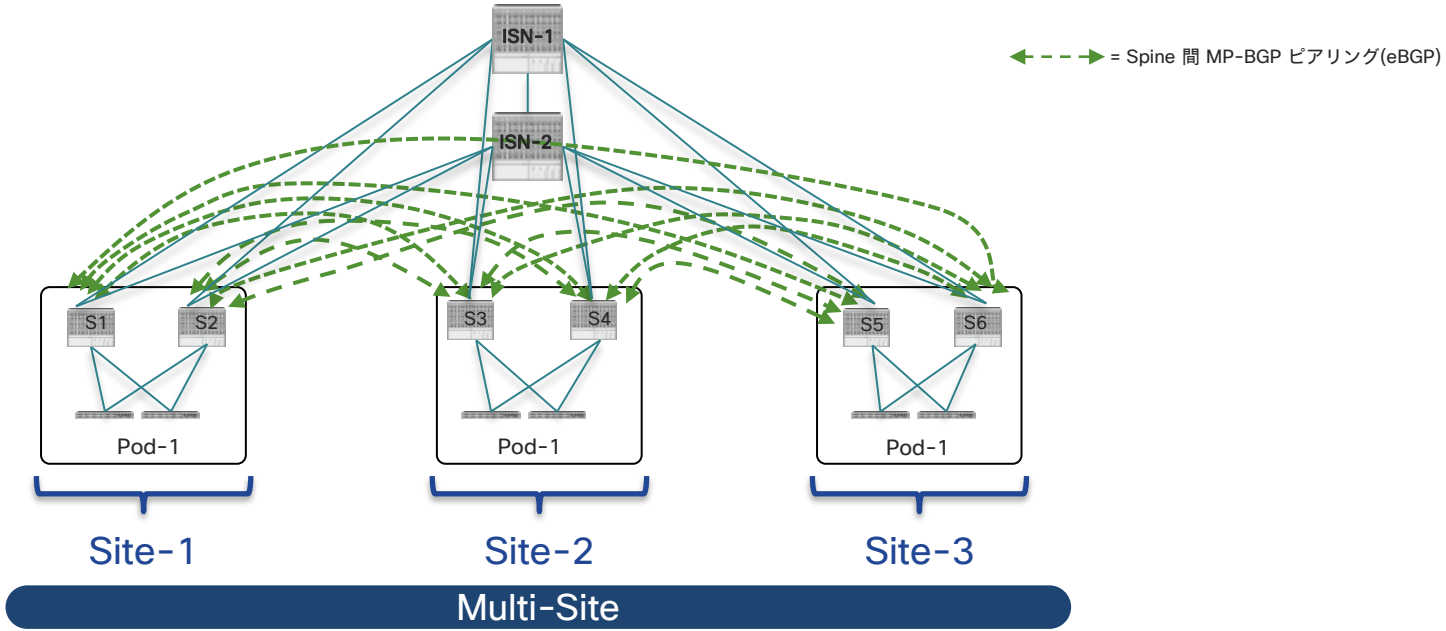
コントロールプレーンの動作	Multi-Pod	Multi-Site
Pod/Site 間での COOP 伝搬	Type-2 の L2VPN	Type-2 の L2VPN
Pod/Site 間での L3Out 共有	MP-BGP VPNv4/VPNv6を用いて、外部プレフィックス情報が交換される	ACI リリース 4.2(1) 以降実装された InterSite L3Out 機能により、MP-BGP VPNv4/VPNv6 を用いて、外部プレフィックス情報が交換される
Pod /Site 間での BGP ピア	・フルメッシュではすべての Spine 間で形成 ・ルートリフレクタ(RR)を構成した場合、Spine と RR 間で形成	・フルメッシュでは別サイトの Spine 間で形成 ・同じ AS で複数サイトを設定し、ルートリフレクタ(RR)を使用した場合、RR 間および Spine と RR 間で形成

Multi-Pod における BGP ピアリング



- Pod 間の BGP L2VPN および VPNv4 ピアリング
- 同じ Pod 内の Spine と Leaf 間の VPNv4 ピアリング

Multi-Site における BGP ピアリング



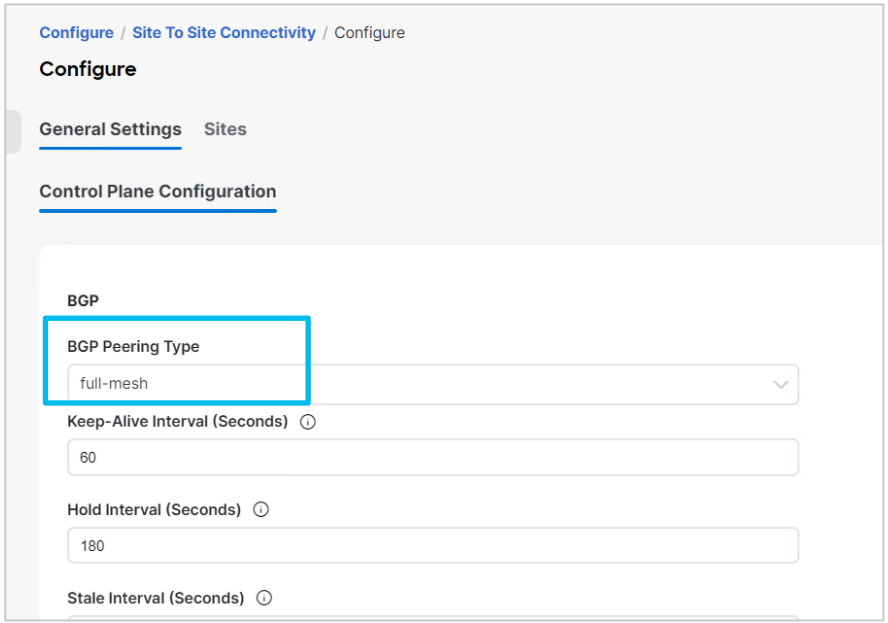
- Spine 間でサイト間 BGP L2VPN ピアリングが張られる
- InterSite L3Out を構成している場合、VPNv4 ピアリングが張られる

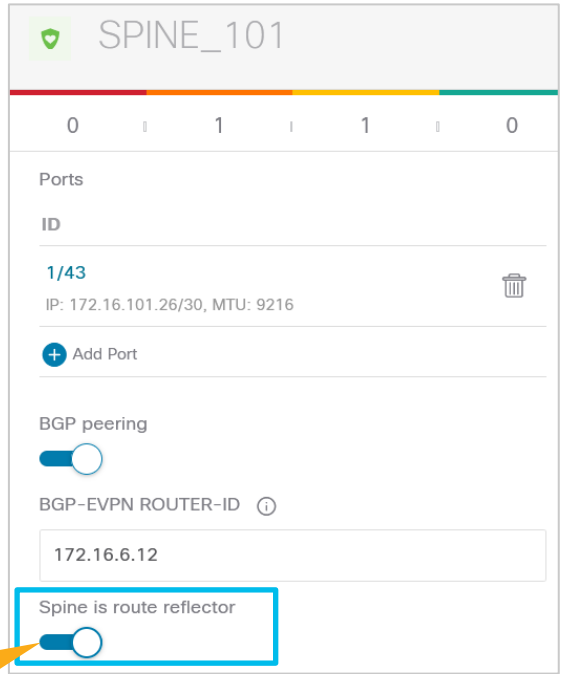
BGP ピアリング ベストプラクティス

- フルメッシュ BGP-EVPN ピアリングを使用することを推奨
- Multi-Pod で外部ルートリフレクタ(Ext-RR)を構成した場合
  - 各 Pod に RR を構成することで Spine とリモート RR 間でフルメッシュ MP-BGP(iBGP) セッションを形成
  - 最大 4 台までをサポート
  - 3 Pod 以上の場合は、最初の 3 つの Pod にのみ Ext-RR を 1 つずつ構成
  - 4 Pod 以上の場合は、フルメッシュを構成しないことを推奨

BGP 設定 / よく見られる間違い

NDO における Multi-Site の BGP 設定

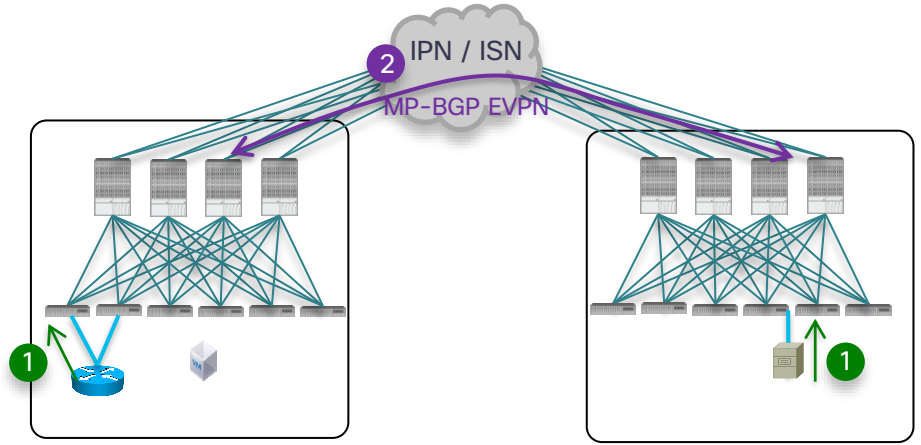




full-mesh ではこのオプションを有効にする必要はない

Multi-Pod / Multi-Site の MTU

1. Data Plane MTU
   - Endpoint (サーバ、ルータ、サービスノード等) が生成する通信トラフィックの MTU サイズ
   - VXLAN カプセル化によるオーバーヘッドのために、 + 50 byte が必要
   - MacSec / CloudSec が有効な場合、 + 40 byte が必要
2. Control Plane MTU
   - ACI ファブリック (Switch CPU) が生成する EVPN 管理通信トラフィックの MTU サイズ
   - デフォルト値は 9000 byte で、IPN/ ISN でサポートされている最大 MTU 値に調整可能
   - 内部 uplink のMTU は 9366 byte



MTU に関する考慮事項は、Remote Leaf や Cloud Network Controller などの ACI 分散ネットワークにも適用される

ACI QoS 概要

	Class of service/ QoS-group	Traffic Type	Dot1p (CoS) marking in VXLAN header	DEI Bit (Drop eligible indicator)
ユーザ定義	0	Level3 user data (default)	0	0
	1	Level2 user data	1	0
	2	Level1 user data	2	0
	4	Level4 user data	2	1
	5	Level5 user data	3	1
	6	Level6 user data	5	1
システム予約	3	APIC controller traffic	3	0
	9	SPAN traffic	4	0
	8(SUP)	Control traffic	5	0
	8(SUP)	Traceroute	6	0
	7	Copy Service	7	0

*ACI 4.0(1) からユーザ定義可能な Class が 3 つ追加された

- ユーザ定義可能な 6 つ*の Class と 4 つのシステム予約 Class が用意されている
- ACI 内部では ACI QoS Class に従って優先制御が行われる
- QoS Class に従って VXLAN Outer ヘッダの CoS と DEI がセットされる
- 着信パケットの CoS 値はデフォルトでは保持されない
- Dot1p Preserve を有効化することで元の CoS 値を維持可能

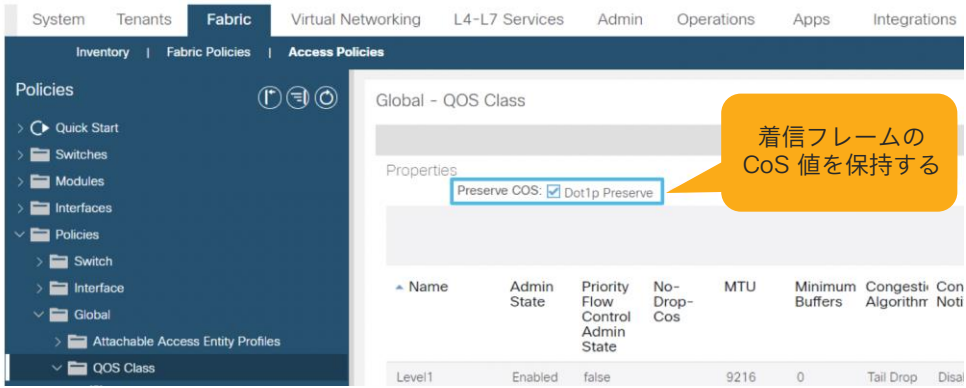
QoS マーキングを保持する方法

DSCP：

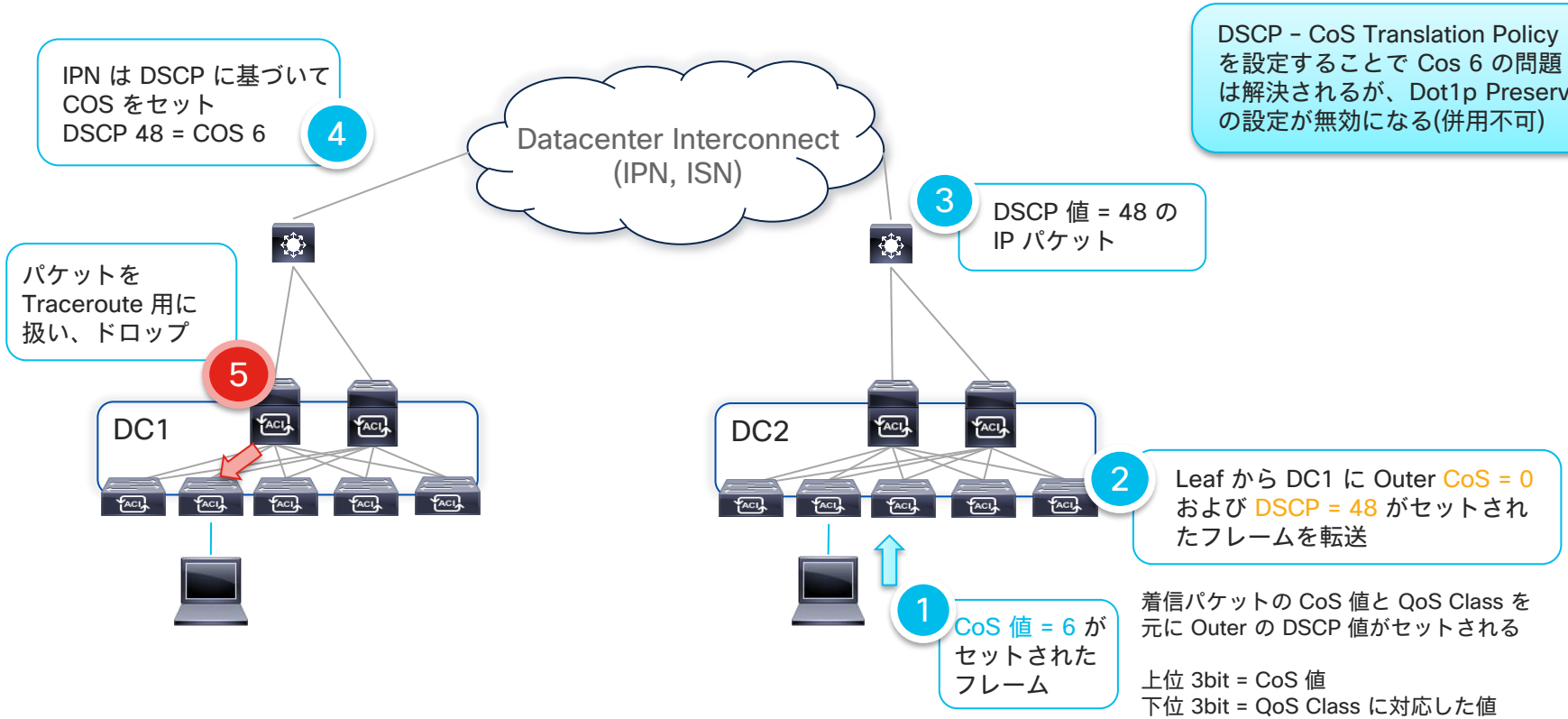
- 着信パケットの DSCP は常に保持される
- Ingress Leaf は着信パケットの DSCP を VXLAN inner DSCP に保持する
- QoS ポリシーにより DSCP のリマーキングを行わない限り、inner の DSCP 値は変更されない

CoS：

- デフォルトで保持されない
  - Ingress Leaf で VLAN header が削除され、Egress でマーキングが行われない（CoS 0が使用される）
- 着信パケットの CoS を保持したい場合は Fabric 全体で“Dot1p Preserve”を有効にする



QoS 問題: 4.0 より前のバージョン



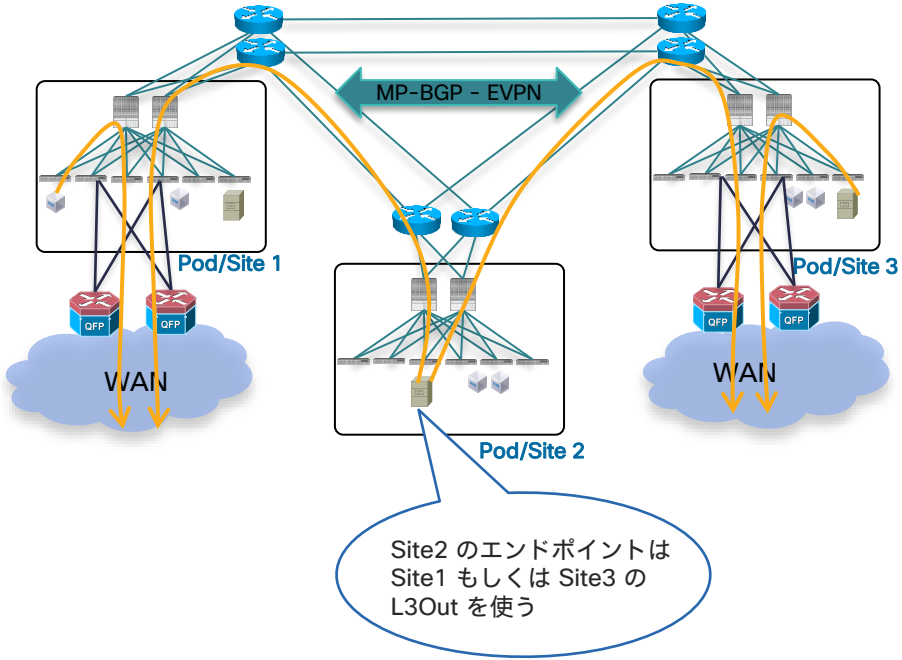
4.0 以降の QoS

- Dot1p Preserve：有効 + DSCP – CoS Translation Policy：無効
  - Egress Spine は、outer DSCP および static DSCP マッピングテーブル（編集不可）に基づいて分類を行い、ユーザ定義の QoS Class にマッピング
  - デフォルトで CoS 6 のドロップ問題を回避
- DSCP – CoS Translation Policy ：有効
  - Ingress Spine は、構成された DSCP/CoS map に基づいて Outer DSCP を変更
  - Egress Spine は、Outer DSCP に基づいて分類を行い、構成された DSCP/CoS map を使用して、ユーザ定義の QoS Class にマッピング
  - さまざまなタイプのトラフィックに対して、柔軟に優先順位を付けることが可能

データプレーン チューニングの ベストプラクティス

外部ルートチューニング
Multi-Pod / Multi-Site

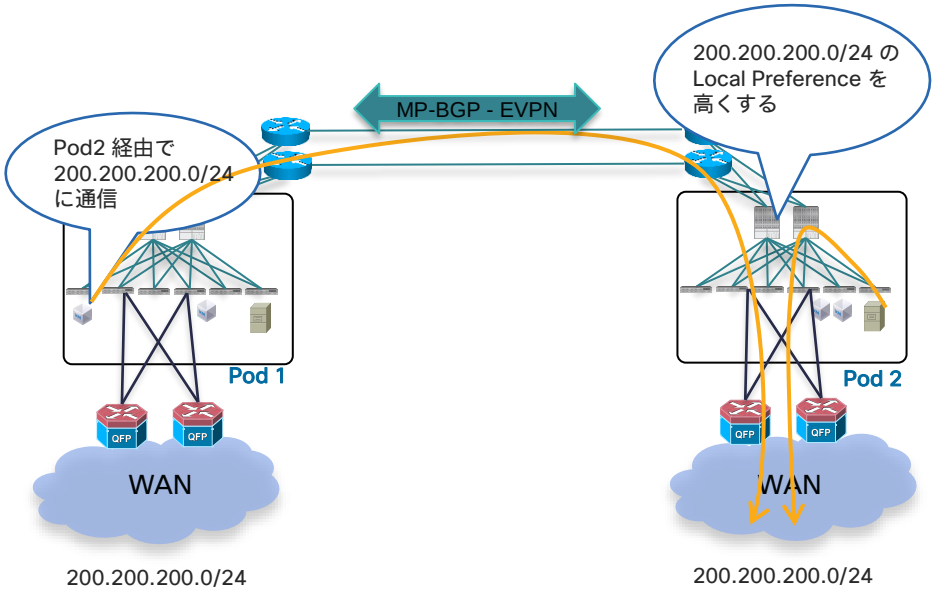
- Pod に専用の WAN 接続は必要ない（例えば、その WAN 接続が他の Pod へのトランジットサービスを提供可能)
    - このガイドラインは Multi-Site にも適用。ACI 4.2(1) 以降導入された InterSite L3Out 機能により実現可能
- 複数の WAN 接続を Pod / Site 内もしくは Pod / Site 間で展開可能
- アウトバウンドトラフィック:
デフォルトでは、常に Preferred IS-IS メトリックに基づいてローカル Pod / Site の WAN 接続を通過



アウトバウンド ルートコントロール

Multi-Pod / Multi-Site

- リモートの Pod / Site が外部ネットワークへの優先パスであることが必要な場合がある
- VRF ごと、もしくは特定の外部宛先ごとに優先度の調整が可能
- Route Map を活用して、受信した外部プレフィックスの Local Preference または MED をチューニング
- Multi-Pod の場合は、Local Preference のチューニングが十分で、Multi-Site の場合は MED のチューニングが必要な場合がある（AS 番号が異なるため）

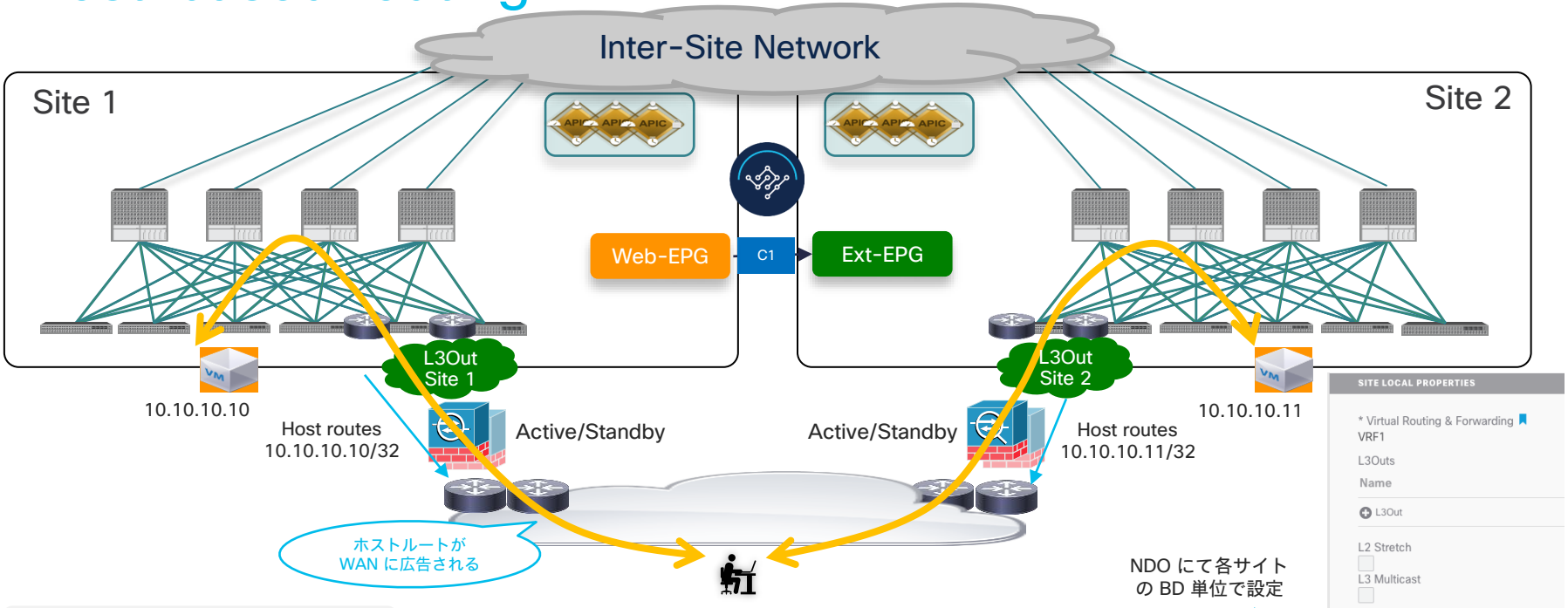


アウトバウンド ルートコントロール
考慮事項

- エンドポイントは Border Leaf に接続すべきではない
  - Border Leaf に接続されたエンドポイントは、常にローカルアウトバウンドパスを優先
- Local-Preference または MED をチューニングするルートマップは、次のプロトコルに対して、それぞれ設定が必要
  - BGP
  - OSPF, EIGRP
    ※MP-BGP に再配布するためのルートマップ
- 次の点も注意ください
  - Leaf ごとの OSPF と EIGRP ルートマップ
  - L3Out ごとの BGP ルートマップ（4.2(4) 以降からネイバーごとに設定可能）

インバウンド ルートコントロール

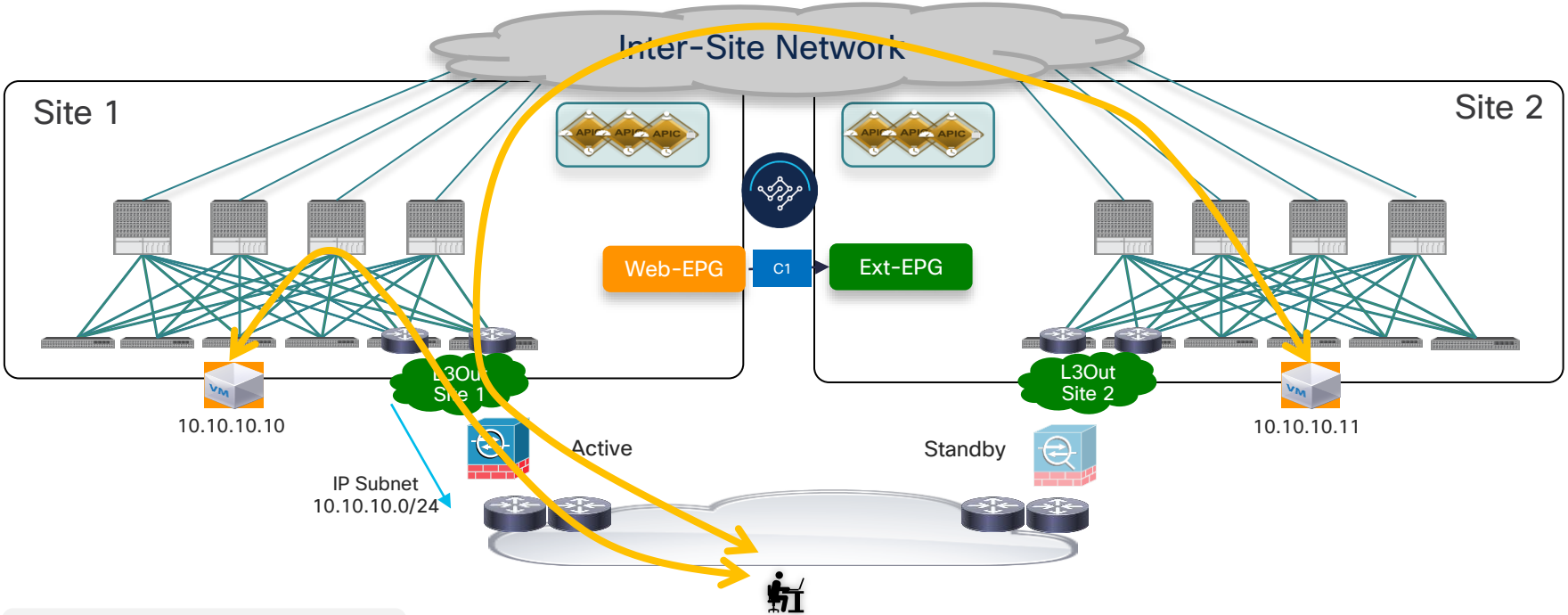
Host-based routing



Multi-Pod 構成にも適用可能

インバウンド ルートコントロール

サイト間の Active/Standby FW ペア



Multi-Pod 構成にも適用可能

Remote Leaf の考慮事項

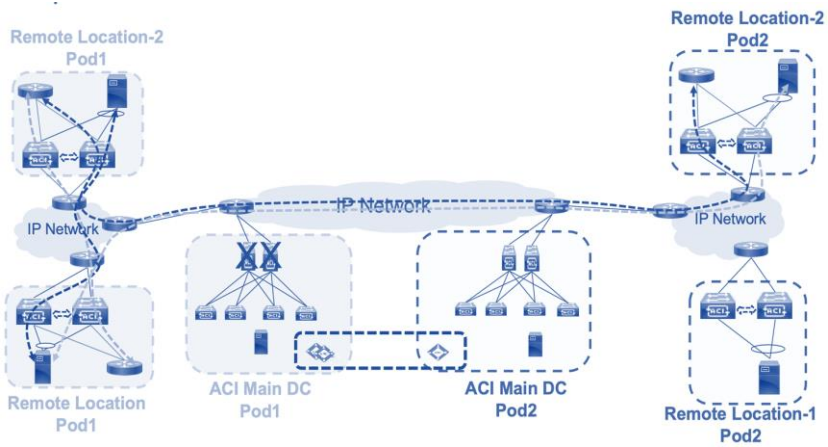
Remote Leaf の考慮事項

概要

- Remote Leaf の技術は、単一ファブリック、Multi-Pod および Multi-Site と共通部分が多い
  - Remote Leaf は COOP を実行し、Ext-RR として構成された Spine に対して VPNv4 セッションを形成
  - Multi-Pod 環境の 1 つの Pod に論理的に接続
  - IPN / ISN ではマルチキャスト対応は不要
- ACI リリース 4.1(2) 以降、メイン DC との Control-Plane および Data-Plane の通信に External TEP Pool を使用
- Multi-Pod と組み合わせて利用可能
- MTU や QoS に関する考慮事項は Multi-Pod / Multi-Site と同じ
  - ルートコントロール（Inbound および Outbound）に関するチューニングも同様

Remote Leaf と Multi-Pod の組み合わせ

- 単一 Fabric の代わりに Multi-Pod を使用すると、Remote Leaf の可用性とパフォーマンスが向上する
  - Direct Traffic Forwarding および T-Glean 機能（4.1(2) 以降実装）
  - Pod 冗長性機能（4.2(2) 以降実装）
- Pod ごとに 1 つの Ext-RR を配置
  - フルメッシュ接続を回避し、スケーラビリティを向上させる
  - 1 つの Pod が障害発生しても、他の Pod とのトラフィックが継続される
  - 代替 Pod の Spine への新しい COOP および BGP セッションが確立



デモ：
アウトバウンド ルートコントロール

キーポイント

01 フルメッシュ BGP-EVPN ピアリングを使用することを推奨。Pod 数が多い場合には、ルートリフレクタを構成可能

02 IPN / ISN デバイスは、Control Plane MTU サイズおよび Data Plane MTU サイズの両方に対応する

03 DSCP – CoS Translation Policy を設定することで、さまざまなタイプのトラフィックに対して、柔軟に優先順位を付けることが可能

04 Inbound および Outbound ルートコントールを行うことで、最適な L3Out を利用した送受信が可能

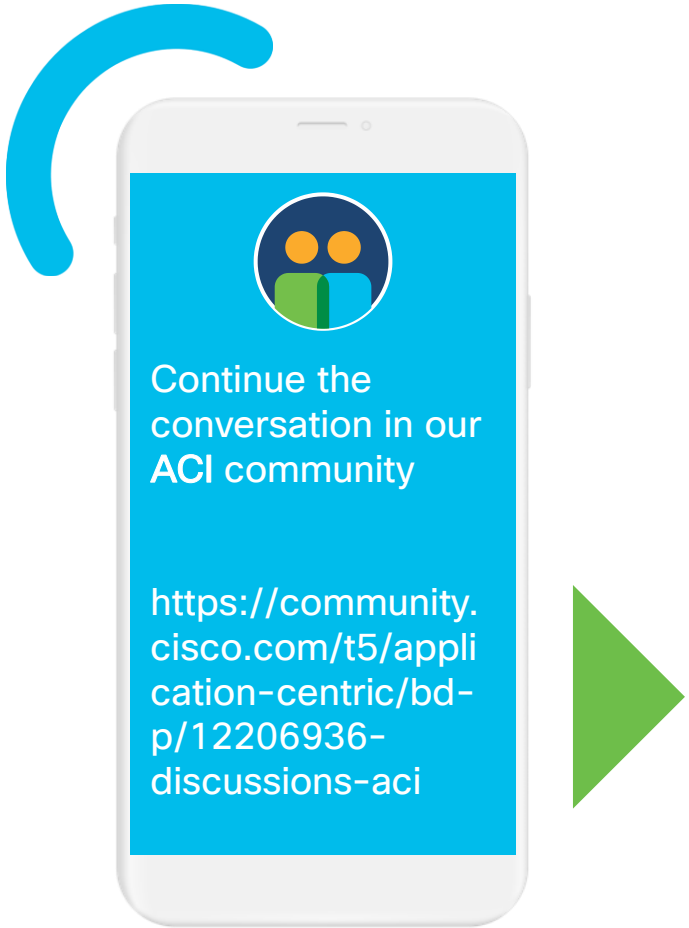
05 Remote Leaf は、Multi-Pod と組み合わせて利用可能

Resources

Ask the Experts リンク集: Cisco ACI

https://community.cisco.com/t5/-/-/ta-p/4147116

※本日の ATXs 以外のリソースリンクも確認できます。



Cisco コミュニティサイトについて

日本語サイトがあります！
言語設定を変えるだけ。

・わからない事
・知りたい事
日本語でご質問下さい！

日本語コミュニティサイト
https://community.cisco.com/t5/japan/tkbc-p/japanese-community


Cisco
Customer Experience