MP-BGP EVPN コントロール プレーン搭載 VXLAN ネットワーク

Contents

はじめに

Virtual Extensible LAN（VXLAN）は、ネットワーク仮想化のオーバーレイ テクノロジーです。IP User Datagram Protocol（MAC in IP / UDP）トンネリング カプセル化の MAC アドレスを使用して、共有レイヤ 3 アンダーレイ インフラストラクチャ ネットワーク上でレイヤ 2 拡張を提供します。オーバーレイ ネットワークでレイヤ 2 拡張を行う目的は、物理サーバ ラックと地理的位置の境界の制限を克服し、データ センター内または異なるデータ センター間のワークロード配置の柔軟性を実現することです。

初期の IETF VXLAN 標準（RFC 7348）では、コントロール プレーンを使用しないマルチキャストベースのフラッドアンドラーニング VXLAN を定義しました。リモート VXLAN トンネル エンドポイント（VTEP）ピア ディスカバリおよびリモート エンドホストラーニングのデータ駆動型フラッドアンドラーニング動作に依存します。オーバーレイ ブロードキャスト、不明なユニキャスト、およびマルチキャスト トラフィックは、マルチキャスト VXLAN パケットにカプセル化され、アンダーレイ マルチキャスト転送を介してリモート VTEP スイッチに転送されます。このような展開でのフラッディングは、ソリューションのスケーラビリティに課題をもたらす可能性があります。一部の組織ではデータセンターや WAN ネットワークでマルチキャストを有効にすることを望んでいないため、アンダーレイ ネットワークでマルチキャスト機能を有効にする要件も課題となります。

RFC 7348 で定義されているフラッドアンドラーニング VXLAN の制限を克服するために、組織は VXLAN のコントロール プレーンとしてマルチプロトコル ボーダー ゲートウェイ プロトコル イーサネット バーチャル プライベート ネットワーク（MP-BGP EVPN）を使用できます。MP-BGP EVPN は、VXLAN オーバーレイの標準ベースのコントロール プレーンとして IETF によって定義されています。MP-BGP EVPN コントロール プレーンは、プライベートおよびパブリック クラウドに適した、よりスケーラブルな VXLAN オーバーレイ ネットワーク設計を可能にする、プロトコルベースの VTEP ピア検出とエンドホスト到達可能性情報配信を提供します。MP-BGP EVPN コントロール プレーンは、オーバーレイ ネットワークでのトラフィック フラッディングを削減または排除する機能セットを導入し、西北および南北両方のトラフィックの最適な転送を可能にします。

このドキュメントでは、MP-BGP EVPN の機能と設定について説明し、MP-BGP EVPN を使用した一般的な VXLAN オーバーレイ ネットワークの設計について説明します。

このドキュメントでは、VXLAN、マルチキャストベースのフラッドアンドラーニングモードの VXLAN、または関連するネットワーク設計オプションの基本については説明しません。VXLAN およびマルチキャストベースのフラッドアンドラーニングを使用した VXLAN の詳細については、次のドキュメントを参照してください。

- VXLAN の概要：CiscoNexus® 9000 シリーズ スイッチ：http://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-729383.html.
- Cisco Nexus 9300 プラットフォーム スイッチを使用した VXLAN の設計：http：//www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-732453.html

このドキュメントでは、BGP、MP-BGP、および BGP とマルチプロトコル ラベル スイッチング（BGP / MPLS）IP VPN に関する予備知識があることを前提としています。詳細については、次の IETF RFC ドキュメントを参照してください。

- RFC 4271-Border Gateway Protocol 4（BGP-4）：https：//tools.ietf.org/html/rfc4271

- RFC 4760-Multiprotocol Extensions for BGP-4：https://tools.ietf.org/html/rfc4760
- RFC 4364-BGP / MPLS IP VPNs：https://tools.ietf.org/html/rfc4364#page-15

MP-BGP EVPN コントロール プレーン：概要

MP-BGP EVPN は、業界標準に基づく VXLAN のコントロール プロトコルです。EVPN 以前は、VXLAN オーバーレイ ネットワークはフラッドアンドラーニング
モードで動作していました。このモードでは、エンドホスト情報学習と VTEP
ディスカバリは両方ともデータプレーン駆動型であり、VTEP
間でエンドホスト到達可能性情報を配信するためのコントロール プロトコルはありません。MP-BGP EVPN はこのモデルを変更します。これにより、リモート VTEP の背後にあるエンド ホスト向けのコントロール プレーン学習が導入されます。これは、コントロール プレーンとデータ プレーンの分離と、VXLAN オーバーレイ ネットワークでのレイヤ 3 およびレイヤ 2 両方向け転送用の統合コントロール プレーンを提供します。

MP-BGP EVPN コントロール プレーンには、次の主な利点があります。

- MP-BGP EVPN プロトコルは業界標準に基づいており、マルチベンダーの相互運用性を実現します。
- エンドホストのレイヤ 2 およびレイヤ 3 の到達可能性情報をコントロール プレーンで学習できるため、組織はより堅牢でスケーラブルな VXLAN オーバーレイ ネットワークを構築できます。
- 10 年前の MP-BGP VPN テクノロジーを使用して、スケーラブルなマルチテナント VXLAN オーバーレイ ネットワークをサポートします。
- EVPN アドレス ファミリは、レイヤ 2 とレイヤ 3 の両方の到達可能性情報を伝送するため、VXLAN オーバーレイ ネットワークに統合されたブリッジングとルーティングを提供します。
- これは、ローカル VTEP でのプロトコルベース ホスト MAC/IP ルート配布および Address Resolution Protocol（ARP）抑制によるネットワーク フラッディングを最小限に抑えます。
- これは、東西および北南のトラフィックに最適な転送を提供し、分散型エニーキャスト機能でワークロード モビリティをサポートします。
- VTEP ピアの検出と認証を提供し、VXLAN オーバーレイ ネットワークでの不正な VTEP のリスクを軽減します。
- レイヤ 2 でアクティブ/アクティブ マルチホーミングを構築するためのメカニズムを提供します。

MP-BGP EVPN コントロール プレーンのソフトウェアおよびハードウェア サポート

デバイスが MP-BGP EVPN VXLAN ネットワークで果たす役割に応じて、MP-BGP EVPN コントロール プレーンを備えた VXLAN ネットワークのコントロール プレーン機能のみ、またはコントロール プレーンとデータ プレーンの両方の機能をサポートする必要がある場合があります。

MP-BGP EVPN を実行する IP トランスポート デバイス

IP トランスポート デバイスは、アンダーレイ ネットワークで IP ルーティングを提供します。MP-BGP EVPN プロトコルを実行すると、VXLAN コントロール プレーンの一部となり、MP-BGP EVPN ピア間で MP-BGP EVPN ルートを配布します。デバイスは、MP-iBGP EVPN ピアまたはルート リフレクタ、または MP 外部 BGP（MP-eBGP）EVPN ピアです。自社の OS ソフトウェアは、MP-BGP EVPN
の更新を理解し、標準で定義された構造を使用して他の MP-BGP EVPN ピアに配布できるように、MP-BGP EVPN をサポートする必要があります。データ転送では、IP トランスポート デバイスは VXLAN

カプセル化パケットの外部 IP アドレスのみに基づいて IP ルーティングを実行します。VXLAN データのカプセル化およびカプセル化解除機能をサポートする必要はありません。

MP-BGP EVPN を実行する VTEP

MP-BGP EVPN を実行する VTEP は、コントロール プレーンとデータ プレーンの両方の機能をサポートする必要があります。コントロール プレーンでは、ローカル ホストをアドバタイズするために MP-BGP EVPN ルートを開始します。ピアから MP-BGP EVPN 更新を受信し、転送テーブルに EVPN ルートをインストールします。データ転送では、ユーザ トラフィックを VXLAN でカプセル化し、IP アンダーレイ ネットワーク経由で送信します。逆方向では、他の VTEP から VXLAN カプセル化トラフィックを受信し、カプセル化を解除し、ネイティブ イーサネット カプセル化を使用してトラフィックをホストに転送します。

さまざまなネットワーク ロールに適したスイッチ プラットフォームを選択する必要があります。IP トランスポート デバイスの場合、ソフトウェアは MP-EVPN コントロール プレーンをサポートする必要がありますが、ハードウェアは VXLAN データプレーン機能をサポートする必要はありません。VTEP の場合、スイッチはコントロール プレーンとデータ プレーンの両方の機能をサポートする必要があります。

VXLAN 間ルーティング

MP-BGP EVPN コントロール プレーンは、VXLAN オーバーレイ ネットワーク上のエンドホストにレイヤ 2 とレイヤ 3 の両方の到達可能性情報を配信することで、統合されたルーティングとブリッジングを提供します。異なるサブネットのホスト間の通信には、VXLAN 間ルーティングが必要です。BGP EVPN は、ホスト IP アドレス ルートまたは IP アドレス プレフィックスのいずれかの形式でレイヤ 3 到達可能性情報を配布することで、この通信を可能にします。データ プレーンでは、VTEP は IP アドレス ルート検索をサポートし、検索結果に基づいて VXLAN カプセル化を実行する必要があります。この機能は、VXLAN ルーティング機能と呼ばれます。すべてのスイッチハードウェア プラットフォームが VXLAN ルーティングをサポートしているわけではないため、ハードウェア プラットフォームの選択に影響します。

Cisco Nexus 9000 シリーズ スイッチでの MP-BGP EVPN VXLAN のサポート

VXLAN の MP-BGP EVPN コントロール プレーンは、Cisco Nexus 9000 シリーズ スイッチの Cisco®NX-OS ソフトウェア リリース 7.0(3)I1(1) に導入されました。ソフトウェア機能は、Cisco Nexus 7000 シリーズスイッチなどの他の Cisco Nexus スイッチ プラットフォームの Cisco NX-OS ソフトウェア トレインにも実装されます。

Cisco NX-OS 7.0(3)I1(1) では、Cisco Nexus 9300 プラットフォーム スイッチは MP-BGP EVPN コントロールプレーン機能と VTEP データプレーン機能の両方をサポートします。Cisco Nexus 9500 プラットフォーム スイッチは、MP-BGP EVPN コントロール プレーン機能をサポートします。VTEP データプレーン機能は、Cisco NX-OS 7.0(3)I1(1) のメンテナンス リリースの Cisco Nexus 9500 プラットフォーム スイッチに追加されます。Cisco Nexus 9300 および 9500 プラットフォームはいずれも、ハードウェアでの VXLAN 間ルーティングをサポートします。

このドキュメントの MP-BGP EVPN 機能および設計の説明の多くはプラットフォームに依存しませんが、Cisco Nexus 9000 シリーズはこのプロトコルをサポートする最初のスイッチ プラットフォームであるため、例は Cisco Nexus 9000 シリーズに基づいています。

MP-BGP EVPN のマルチテナンシー

既存の MP-BGP の拡張機能として、MP-BGP EVPN は、Virtual Routing and Forwarding（VRF）構造を使用した VPN によるマルチテナンシーのサポートを継承します。MP-BGP EVPN では、複数のテナントが共有し、共通の IP トランスポート ネットワークを共有する一方で、VXLAN オーバーレイ ネットワーク内に独自の VPN があります。

EVPN VXLAN オーバーレイ ネットワークでは、VXLAN ネットワーク識別子（VNI）がレイヤ 2 ドメインを定義し、レイヤ 2 トラフィックが VNI 境界を通過できないようにすることで、レイヤ 2 セグメンテーションを適用します。同様に、VXLAN テナント間のレイヤ 3 セグメンテーションは、レイヤ 3 VRF テクノロジーを適用し、各 VRF インスタンスにマッピングされた個別のレイヤ 3 VNI を使用して、テナント間にルーティング分離を適用することによって実現されます。各テナントには、独自の VRF ルーティング インスタンスがあります。特定のテナントの VNI の IP サブネットは、レイヤ 3 ルーティング ドメインを他のテナントから分離する同じレイヤ 3 VRF インスタンスにあります。

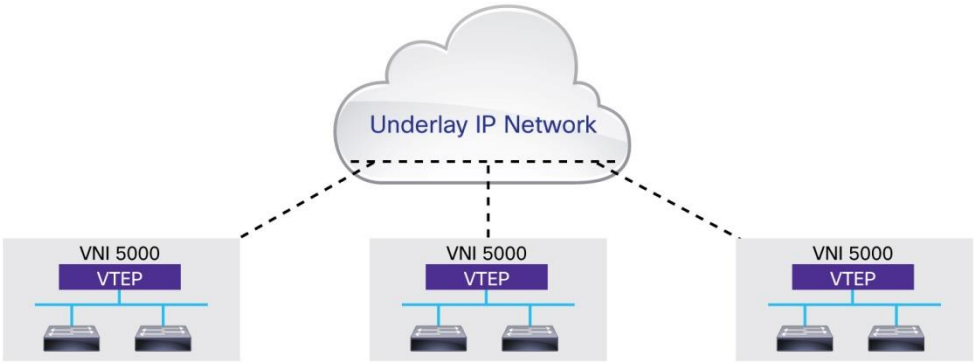
組み込みマルチテナンシーのサポートは、マルチキャストベースのフラッドアンドラーニング VXLAN およびマルチテナンシー機能のない他のレイヤ 2 拡張テクノロジーと比較して、MP-BGP EVPN VXLAN の利点です。VXLAN テクノロジーは、マルチテナント モデルを使用して展開されるクラウド ネットワークにより適しています。

MP-BGP EVPN NLRI および L2VPN EVPN アドレス ファミリ

他のネットワーク ルーティングコントロール プロトコルと同様に、MP-BGP EVPN はネットワークのネットワーク層到達可能性情報（NLRI）を配信するように設計されています。EVPN NLRI の独自の機能は、EVPN VXLAN オーバーレイ ネットワークに存在するエンド ホストのレイヤ 2 とレイヤ 3 の両方の到達可能性情報を含むことです。つまり、EVPN VXLAN エンドホストの MAC アドレスと IP アドレスの両方をアドバタイズします。この機能は、VXLAN 統合ルーティングおよびブリッジングのサポートの基盤となります。

VXLAN はレイヤ 2 拡張テクノロジーであるため、レイヤ 2 MAC アドレスを配布する必要があります。ネットワーク内の特定の場所に限定され、レイヤ 2 およびレイヤ 3 の境界内にある従来の VLAN とは異なり、VNI はオーバーレイ ネットワーク内の仮想レイヤ 2 セグメントです。ただし、アンダーレイ ネットワークの観点からは、アンダーレイ インフラストラクチャのレイヤ 2 およびレイヤ 3 の境界を超えて、複数の非隣接サイトにまたがることがあります（図 1）。同じ VNI 内のエンド ホスト間のトラフィックは、オーバーレイ ネットワークでブリッジする必要があります。つまり、特定の VNI 内の VTEP デバイスは、この VNI 内のエンド ホストの他の MAC アドレスを認識する必要があります。BGP EVPN を介した MAC アドレスの配布により、VXLAN での不明なユニキャスト フラッディングを削減または排除できます。

図 1.　　アンダーレイ IP ネットワーク全体の VNI



レイヤ 3 ホスト IP アドレスは MP-BGP EVPN を介してアドバタイズされるため、VXLAN 間トラフィックは最適なパスを介して宛先エンド ホストにルーティングできます。宛先エンド ホストにルーティングする必要がある VXLAN 間トラフィックの場合、ホストベースの IP ルーティングにより、宛先ホストの正確な場所への最適な転送パスを提供できます。

MP-BGP EVPN は、VNI の IP サブネット プレフィックス ルートもアドバタイズできます。プレフィックス ルートは、ホスト IP ルートがない場合（たとえば、ホスト IP ルートが MP-BGP を介して VTEP によってまだ学習されていない場合）、宛先ホストにトラフィックをルーティングするために使用できます。サブネットがルーティング可能で、VXLAN ネットワークの外部に通知される必要がある場合、VTEP は VXLAN ネットワークの外部にプレフィックス ルートをアドバタイズすることもできます。

EVPN NLRI は、レイヤ 2 VPN（L2VPN）EVPN と呼ばれる新しいアドレス ファミリを持つ BGP マルチプロトコル拡張機能を使用して BGP で伝送されます。BGP MPL ベース IP VPN（RFC 4364）の VPNv4 アドレスファミリと同様に、EVPN の L2VPN EVPN アドレスファミリは、ルート識別子（RD）を使用して、異なる VRF インスタンスの同一ルート間の固有性を維持し、ルート ターゲット（RT）を使用して、さまざまな VRF インスタンスでルートをアドバタイズおよび共有する方法を決定するポリシーを定義します。

ルート識別子は、あるルート セット（1 つの VRF インスタンス）を別のルート セットと区別するために使用される 8 ビットのオクテット番号です。これは、各ルートに付加される固有の番号であるため、複数の異なる VRF インスタンスで同じルートが使用される場合、BGP はそれらを個別のルートとして扱うことができます。EVPN ルートが MP-BGP ピアと交換される場合、ルート識別子は MP-BGP を介してルートとともに送信されます。

ルート ターゲットを VRF インスタンスに適用して、このインスタンスと他の VRF インスタンス間のルートのインポートとエクスポートを制御できます。ルートのルートターゲット属性は BGP 拡張コミュニティ属性の形式で配布されるため、MP-BGP EVPN を実行するデバイスの BGP 設定を有効にして、拡張コミュニティ属性を生成または処理する必要があります。

Cisco NX-OS の実装では、設定を容易にするために BGP ルート識別子とルート ターゲットを自動的に生成できます。BGP ルート識別子は、VTEP スイッチの VNI および BGP ルータ ID から自動的に取得でき、BGP ルート ターゲットは BGP AS：VNI として自動的に生成できます。または、BGP ルート識別子とルート ターゲットを手動で設定することもできます。ネットワーク内のすべての MP-BGP EVPN VTEP が Cisco Nexus スイッチ プラットフォームの場合は、自動生成された route-distinguisher および route-target 値を使用することを推奨します。複数のベンダーの VTEP デバイスが相互運用されている場合は、ベンダーの実装の違いによる問題を回避するために、値を手動で設定す

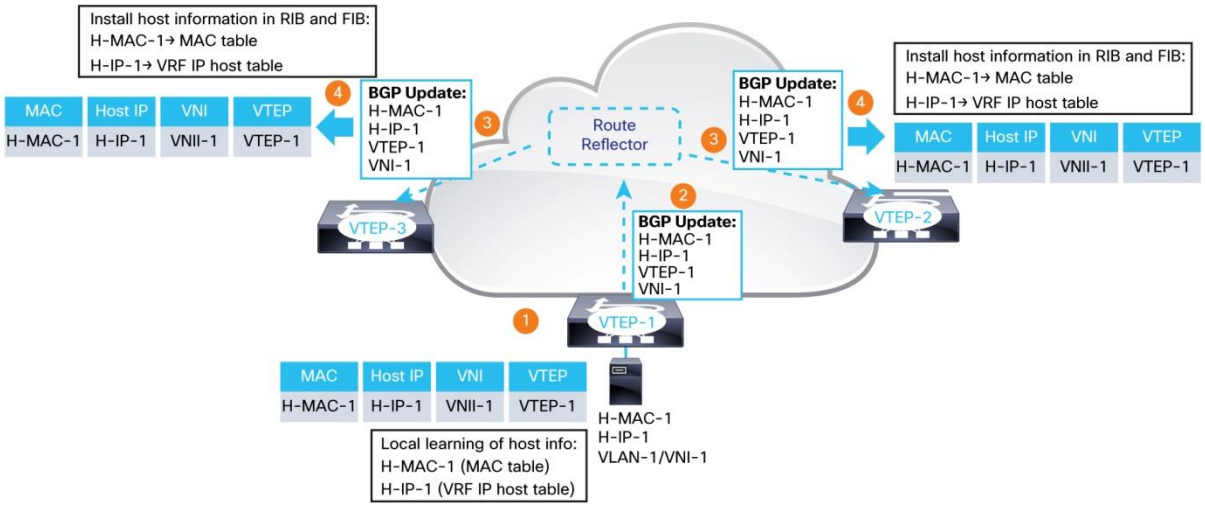
ることを推奨します。VTEP が異なる BGP ドメインに存在する eBGP 展開シナリオでは、BGP ルート ターゲットを手動で割り当てる必要があります。

MP-BGP EVPN コントロール プレーンによる統合ルーティングおよびブリッジング

MP-BGP EVPN コントロール プレーンは、VXLAN オーバーレイ ネットワークに存在するエンドホストにレイヤ 2 およびレイヤ 3 の到達可能性情報を配信することで、統合されたルーティングとブリッジングを提供します。各 VTEP はローカル学習を実行して、ローカルに接続されたホストから MAC および IP アドレス情報を取得し、MP-BGP EVPN コントロール プレーンを介してこの情報を配布します。リモート VTEP に接続されたホストは、MP-BGP コントロール プレーンを介してリモートで学習されます。このアプローチは、エンドホスト学習のネットワーク フラッディングを削減し、エンドホストの到達可能性情報の配信をより適切に制御します。

図 2 に、ルート リフレクタを使用した MP-iBGP EVPN でのエンドホスト NLRI 学習および配布の例を示します。

図 2. MP-BGP EVPN ホスト NLRI ラーニングおよび配布



ローカルホスト ラーニング

MP-BGP EVPN の VTEP は、ローカルに接続されたエンドホストの MAC アドレスと IP アドレスを学習します。この学習は、受信イーサネット フレームからの送信元 MAC アドレス学習や、ホストが Gratuitous ARP（GARP）およびリバース ARP（RARP）パケットまたはVTEP上のゲートウェイ IP アドレスに対する ARP 要求を送信する際の IP アドレス学習など、標準のイーサネットおよび IP 学習手順を使用して、ローカルデータプレーンベースにすることができます。または、コントロール プレーンを使用するか、または VTEP とローカル ホスト間の管理プレーン統合によって学習することもできます。

EVPN ルート アドバタイズメントとリモートホスト ラーニング

ローカルホスト MAC アドレスと IP アドレスを学習した後、VTEP は MP-BGP EVPN コントロール プレーンでホスト情報をアドバタイズし、この情報を他の VTEP に配布できるようにします。このアプローチにより、EVPN VTEPはMP-BGP EVPN コントロール プレーンのリモート エンド ホストを学習できます。

EVPN ルートは、L2VPN EVPN アドレス ファミリを介してアドバタイズされます。BGP L2VPN EVPN ルートには、次の情報が含まれます。

- RD：Route Distinguisher（ルート識別子）
- MAC アドレス長：6 バイト
- MAC アドレス：ホスト MAC アドレス
- IP アドレスの長さ：32 または 128
- IP アドレス：ホスト IP アドレス（IPv4 または IPv6）
- L2 VNI：エンド ホストが属するブリッジ ドメインの VNI
- L3 VNI：テナント VRF ルーティング インスタンスに関連付けられた VNI

MP-BGP EVPN は、BGP 拡張コミュニティ属性を使用して、エクスポートされたルートターゲットを EVPN ルートで送信します。EVPN VTEP は EVPN ルートを受信すると、受信したルートのルート ターゲット属性をローカルに設定されたルートターゲット インポート ポリシーと比較して、ルートをインポートするか無視するかを決定します。このアプローチでは、10 年前の MP-BGP VPN テク ノロジー（RFC 4364）を使用し、VRF をローカルに持たないノードが対応するルートをインポートしないスケーラブルなマルチテナント機能を提供します。VPN スケーリングは、route-target-constrained ルート配布（RFC 4684）などの BGP 構造を使用することでさらに強化できます。

VTEP スイッチは、ローカルで学習されたエンド ホストに対して MP BGP EVPN ルートを発信するときに、自身の VTEP アドレスを BGP ネクストホップとして使用します。この BGP ネクストホップは、ネットワーク全体のルート配布を通じて変更されないようにする必要があります。これは、オーバーレイ ネットワークのパケットを転送するときに、リモート VTEP が発信 VTEP アドレスを VXLAN カプセル化のネクストホップとして学習する必要があるためです。

アンダーレイ ネットワークは、アンダーレイ ネットワークを介して、カプセル化された VXLAN パケットを出力 VTEP 方向にルーティングするすべての VTEP アドレスに IP 到達可能性を提供します。アンダーレイ ネットワーク内のネットワーク デバイスは、VTEP アドレスに対してのみルーティング情報を維持する必要があります。EVPN ルートを学習する必要はありません。このアプローチにより、アンダーレイ ネットワークの動作が簡素化され、安定性と拡張性が向上します。

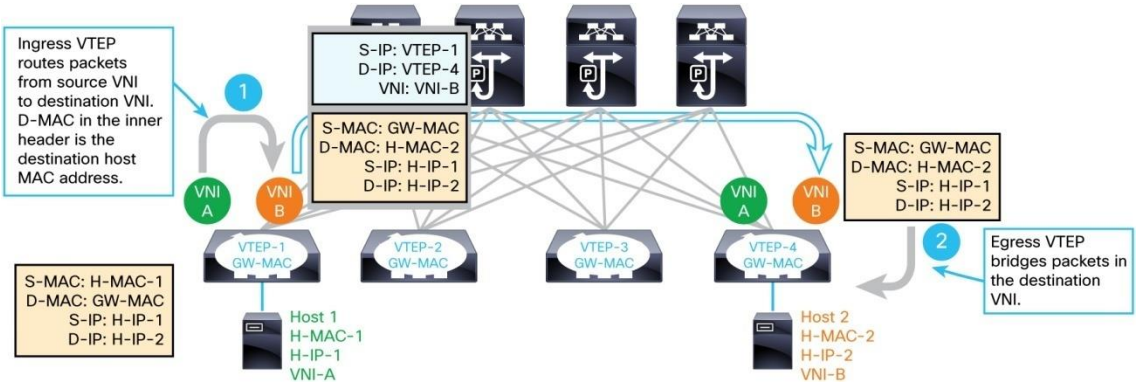
対称および非対称統合ルーティングおよびブリッジング

IETF EVPN ドラフトでは、非対称 IRB と対称 IRB の 2 つの Integrated Routing and Bridging（IRB）セマンティクスが定義されています。Cisco Nexus スイッチ プラットフォーム用の Cisco NX-OS は、拡張性の利点と、レイヤ 2 およびレイヤ 3 のマルチテナンシー サポートを簡素化するために、対称 IRB を実装しています。

非対称 IRB

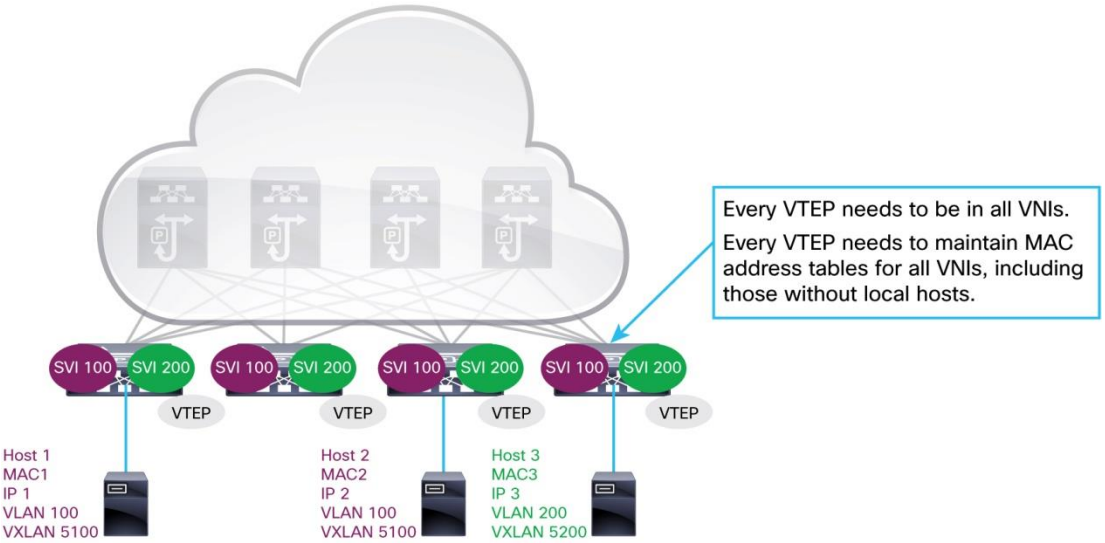
非対称 IRB では、入力 VTEP はレイヤ 2 ブリッジングとレイヤ 3 ルーティング検索の両方を実行しますが、出力 VTEP はレイヤ 2 ブリッジング検索のみを実行します。図 3 に示すように、非対称 IRB ではパケットが 2 つの VNI 間を移動すると、入力 VTEP が送信元 VNI から宛先 VNI にパケットをルーティングします。出力 VTEP は、パケットを宛先 VNI 内の宛先ポイントにブリッジします。

図 3. 非対称 IRB を使用した VXLAN ルーティング



非対称 IRB では、レイヤ 2 およびレイヤ 3 転送の両方に対し、送信元と宛先の両方の VNI で入力 VTEP を設定する必要があります。基本的に、各 VTEP を VXLAN ネットワークのすべての VNI で設定し、それらの VNI に接続されているすべてのエンド ホストの ARP エントリと MAC アドレスを学習する必要があります（図 4）。この動作により、エンド ホストの密度やオーバーレイ ネットワーク内の VXLAN VNI の数が増えると、スケーラビリティの問題が発生する可能性があります。

図 4. 非対称 IRB の VTEP VNI メンバーシップ



対称 IRB

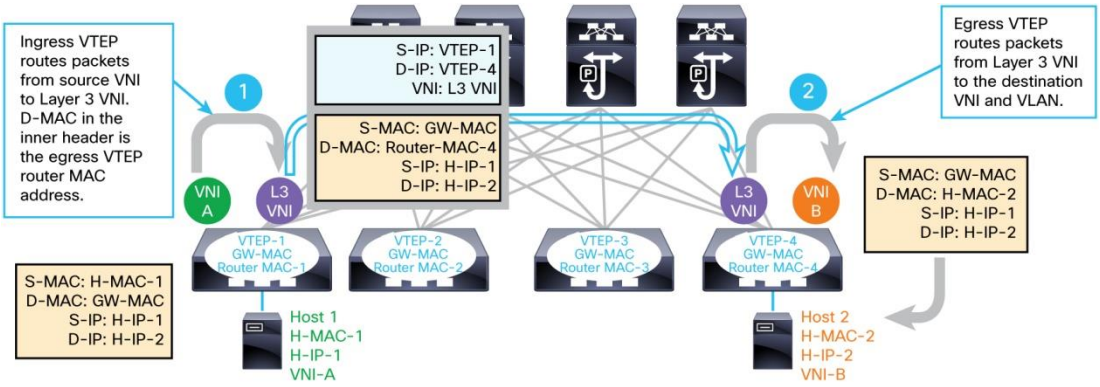
対称 IRB では、入力 VTEP と出力 VTEP の両方がレイヤ 2 およびレイヤ 3 検索を実行します。対称 IRB では、いくつかの新しい論理構造が導入されています。

- **レイヤ 3 VNI：**各テナント VRF インスタンスは、ネットワーク内の一意のレイヤ 3 VNI にマッピングされます。このマッピングは、ネットワーク内のすべての VTEP で一貫している必要があります。すべての VXLAN 間ルーテッド トラフィックは、VXLAN ヘッダー内のレイヤ 3 VNI でカプセル化され、受信 VTEP に VRF コンテキストを提供します。受信側の VTEP はこの VNI を使用して、内部 IP パケットを転送する必要がある VRF コンテキストを決定します。この VNI は、データ プレーンでレイヤ 3 セグメンテーションを実行するための基盤にもなります。

- **VTEP ルータの MAC アドレス：**各 VTEP には、他の VTEP が VNI 間ルーティングに使用できる固有のシステム MAC アドレスがあります。この MAC アドレスは、ここではルータ MAC アドレスと呼ばれます。ルータの MAC アドレスは、ルーテッド VXLAN パケットの内部宛先 MAC アドレスとして使用されます。

図 5 に示すように、パケットが VNI A から VNI B に送信されると、入力 VTEP はパケットをレイヤ 3 VNI にルーティングします。内部の宛先 MAC アドレスを出力 VTEP のルータ MAC アドレスに書き換え、レイヤ 3 VNI を VXLAN ヘッダーにエンコードします。出力 VTEPは、カプセル化された VXLAN パケットを受信すると、最初に VXLAN ヘッダーを削除してパケットのカプセル化を解除します。次に、内部パケット ヘッダーを調べます。内部パケット ヘッダーの宛先 MAC アドレスは独自の MAC アドレスであるため、レイヤ 3 ルーティング検索を実行します。VXLAN ヘッダーのレイヤ 3 VNI は、このルーティング検索が実行される VRF コンテキストを提供します。

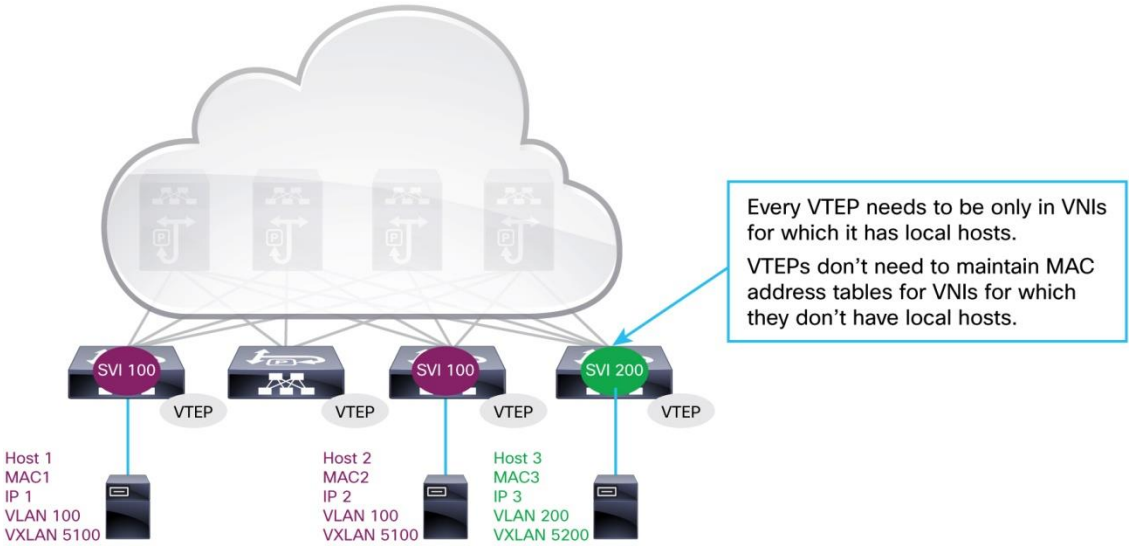
図 5. 対称 IRB を使用した VXLAN ルーティング



対称 IRB の利点

対称 IRB を使用すると、入力 VTEP は VNI 間ルーティングの宛先 VNI を認識する必要がありません。したがって、VTEP は、ローカル ホストを持たない出力 VNI に接続されたリモート ホストの MAC アドレス情報を学習および維持する必要はありません。このアプローチにより、VTEP での MAC アドレス テーブルと ARP 隣接の使用率が向上します。たとえば、図 6では VNI-B のすべてのホスト MAC アドレスと ARP 隣接が VTEP-1 に存在する必要はありません。その結果、ルーティングとブリッジングは、非対称 IRB よりも拡張性が高くなります。Cisco NX-OS は対称 IRB を実装して、最適な学習とスケーリングを実現します。

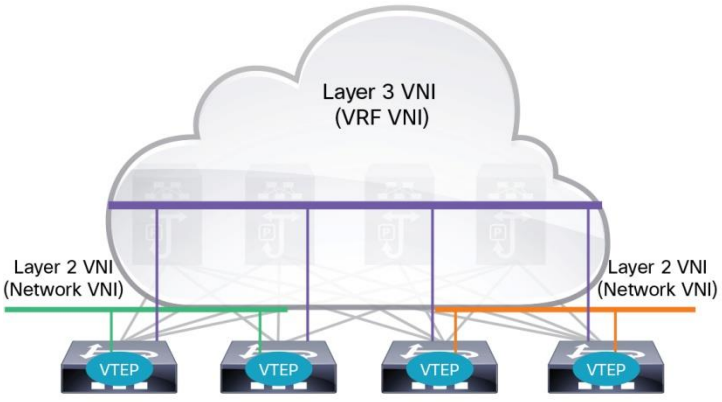
図 6.　　対称 IRB を使用した VTEP VNI メンバーシップ



ブリッジ ドメインおよび IP VRF インスタンスの VNI

EVPN VXLAN テナントは、それぞれが対応する VNI を持つ複数のレイヤ 2 ネットワークを持つことができます。これらのレイヤ 2 ネットワークは、オーバーレイ ネットワークのブリッジ ドメインです。これらに関連付けられている VNI は、多くの場合レイヤ 2（L2）VNI と呼ばれます。VXLAN 間ルーティングが必要な場合、各テナントには対称 IRB のレイヤ 3（L3）VNI も必要です。VTEP には、VXLAN EVPN のレイヤ 2 VNI のすべてまたはサブセットを含めることができますが、VXLAN 間ルーティングにはレイヤ 3 VNI が必要です。EVPN 内のすべての VTEP に同じレイヤ 3 VNI が必要です（図 7）。

図 7.　　ブリッジ ドメインおよび IP VRF インスタンスの VNI



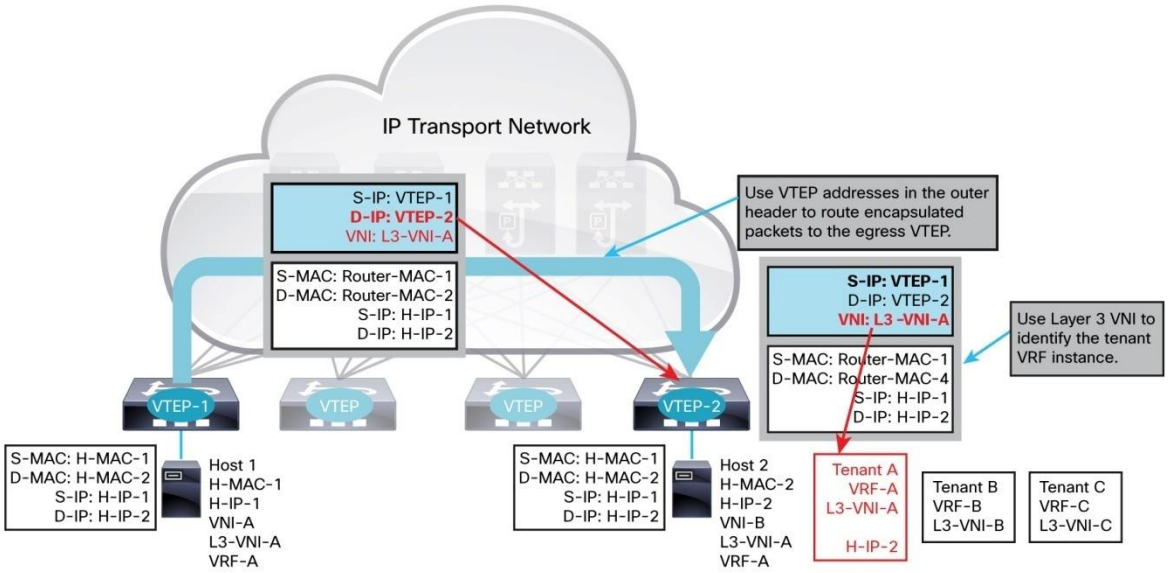

```
vlan 200
  vn-segment 20000
vlan 201
  vn-segment 20100

vlan 3900
  name l3-vni-vlan-for-tenant-1
  vn-segment 39000

vrf context evpn-tenant-1
  vni 39000
  rd auto
  address-family ipv4 unicast
    route-target import 39000:39000
    route-target export 39000:39000
    route-target both auto evpn

interface Vlan200
  no shutdown
  vrf member evpn-tenant-1
  ip address 20.0.0.1/24
  fabric forwarding mode anycast-gateway

interface Vlan201
  no shutdown
  vrf member evpn-tenant-1
  ip address 20.1.0.1/24
  fabric forwarding mode anycast-gateway

interface Vlan3900
  description l3-vni-for-tenant-1-routing
  no shutdown
  vrf member evpn-tenant-1
  ip forward
```

EVPN VTEPは、ローカルエンドホストから受信したパケットに対して転送ルックアップとVXLANカプセル化を実行する場合、パケットをブリッジする必要があるかどうかに応じて、VXLANヘッダーでレイヤ2 VNIまたはレイヤ3 VNIを使用します。ルーティングされます。元のパケット ヘッダーの宛先 MAC アドレスがローカル VTEP に属していない場合、ローカル VTEP はレイヤ 2 検索を実行し、送信元ホストと同じレイヤ 2 VNI

にある宛先エンド ホストにパケットをブリッジします。 ローカル VTEP は、このレイヤ 2 VNI を VXLAN ヘッダーに埋め込みます。この場合、送信元ホストと宛先ホストの両方が同じレイヤ 2 ブロードキャスト ドメインにあります。宛先 MAC アドレスがローカル VTEP スイッチに属する場合、つまり、ローカル VTEP が送信元ホストの IP ゲートウェイであり、送信元ホストと宛先ホストが異なる IP サブネットにある場合、パケットはローカル VTEP によってルーティングされます。この場合、レイヤ 3 ルーティング検索を実行します。次に、レイヤ3 VNIを含むパケットを VXLAN ヘッダーにカプセル化し、内部宛先 MAC アドレスをリモート VTEP のルータ MAC アドレスに書き換えます。カプセル化された VXLAN パケットを受信すると、リモート VTEP は内部 IP ヘッダーに基づいて別のルーティング検索を実行します。これは、受信したパケットの内部宛先 MAC アドレスがリモート VTEP 自体に属するためです。

VXLAN パケットの外部 IP ヘッダー内の宛先 VTEP アドレスは、アンダーレイ ネットワーク内の宛先ホストの場所を識別します。VXLAN パケットは、外部宛先 IP アドレスに基づいて、アンダーレイ ネットワークを介して出力 VTEP にルーティングされます。パケットが出力 VTEP に到着すると、VXLAN ヘッダー内の VNI が検査され、パケットがブリッジされる VLAN、またはパケットがルーティングされるテナント VRF インスタンスが決定されます。後者の場合、VXLAN ヘッダーはレイヤ 3 VNI でエンコードされます。レイヤ 3 VNI はテナント VRF ルーティング インスタンスに関連付けられているため、出力 VTEP は、ルーティングされた VXLAN パケットを適切なテナント ルーティング インスタンスに直接マッピングできます。図 8 に、対称 IRB でのこの転送の概念を示します。このアプローチにより、レイヤ 2 とレイヤ 3 の両方のセグメンテーションのサポートが容易になります。

図 8. 対称 IRB ルーティングを使用した VXLAN パケット転送



MP-BGP EVPN での VTEP ピア検出および認証

MP-BGP EVPN 以前は、VXLAN には制御プロトコルベースの VTEP ピア検出メカニズムや VTEP ピアを認証する方法がありませんでした。これらの制限は、VXLAN トラフィックを送受信するために不正な VTEP を VNI セグメントに簡単に挿入できるため、実際の VXLAN 展開では重大なセキュリティ リスクをもたらします。

MP-BGP EVPN コントロール プレーンでは、VTEP デバイスは最初に他の VTEP または内部 BGP（iBGP）ルート リフレクタとの BGP ネイバー隣接関係を確立する必要があります。エンドホスト NLRI の BGP 更新に加えて、VTEP は BGP を介して自身に関する次の情報を交換します。

- レイヤ 3 VNI
- VTEP アドレス
- ルータ MAC アドレス

リモート VTEP BGP ネイバーから BGP EVPN ルート更新を受信するとすぐに、VTEP はそのルート アドバタイズメントからの VTEP アドレスを VTEP ピア リストに追加します。この VTEP ピア リストは、有効な VTEP ピアの許可リストとして使用されます。この許可リストにない VTEP は、無効な送信元または許可されていない送信元と見なされます。これらの無効な VTEP からの VXLAN カプセル化トラフィックは、他の VTEP によって廃棄されます。

データプレーン転送では、BGP EVPN VTEP は、許可リストにある VTEP ピアからの VXLAN カプセル化パケットのみを受け入れます。したがって、MP-BGP EVPN はプロトコルベースの VTEP ディスカバリと、VXLAN オーバーレイ トラフィックの配信を BGP 学習 VTEP のみに制限する機能を導入しています。

VTEP ピアの学習を促進する VTEP アドレスとともに、BGP EVPN ルートは VTEP ルータの MAC アドレスを伝送します。各 VTEP にはルータの MAC アドレスがあります。VTEP のルータ MAC アドレスが MP-BGP を介して配信され、他の VTEP によって学習されると、他の VTEP はそれを VTEP ピアの属性として使用して、VXLAN 間でルーティングされたパケットをその VTEP ピアにカプセル化します。ルータの MAC アドレスは、ルーテッド VXLAN の内部宛先 MAC アドレスとしてプログラムされます。

セキュリティを強化するために、既存の BGP Message Digest 5（MD5）認証を BGP ネイバーセッションに簡単に適用できるため、スイッチが事前に設定されたMD5 Triple Data Encryption Standard（3DES）で相互に正常に認証されるまで、MP-BGP EVPN ルートを交換するための BGP ネイバーになることができません。MP-BGP EVPN の BGP ネイバー認証は、以前に BGP でサポートされていたのと同じ方法で設定されます。次に例を示します。

On VTEP-1

```
router bgp 100
  router-id 10.1.1.101
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 10.1.1.102 remote-as 100
    password 3 a667d47acc18ea6b
    update-source loopback0
    address-family ipv4 unicast
      send-community both
    address-family l2vpn evpn
      send-community both
```

- Can be configured using the command-line interface (CLI) with a clear password: **password cisco123**. The system will automatically change this password to a 3DES-encrypted password in the running configuration display.
- Both neighbors need to have the exact same password.

On VTEP-2

```
router bgp 100
  router-id 10.1.1.102
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
    retain route-target all
  neighbor 10.1.1.101 remote-as 100
    password 3 a667d47acc18ea6b
    update-source loopback0
    address-family ipv4 unicast
      send-community both
    address-family l2vpn evpn
      send-community both
```

次に、Cisco NX-OS での VNI ピアのステータスと情報の表示例を示します。

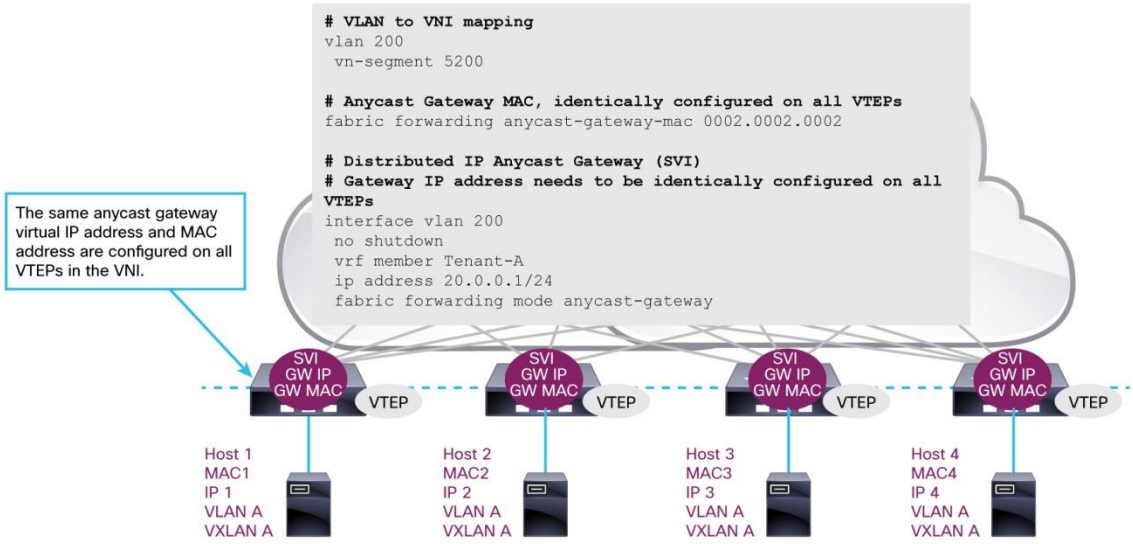
```
VTEP1-1# sh nve peers
Interface Peer-IP          State LearnType Uptime   Router-Mac
--------- ---------------  ----- --------- -------- -----------------
nve1      10.1.1.102        Up    CP        1w3d    6412.2574.6ae7
nve1      10.1.1.134        Up    CP        1w3d     7c69.f6df.e71f

VTEP-1#
```

```
VTEP-1# sh nve peers peer-ip 10.1.1.102 det
Details of nve Peers:
----------------------------------------
Peer-Ip: 10.1.1.102
    NVE Interface       : nve1
    Peer State          : Up
    Peer Uptime         : 1w3d
    Router-Mac          : 6412.2574.6ae7
    Peer First VNI      : 20100
    Configured VNIs     : 20000,20100,21000,21100,39000,39010
    Provision State     : add-complete
    Route-Update        : Yes
    Peer Flags          : DisableLearn
    Learnt CP VNIs      : 20000,20100
----------------------------------------
VTEP-1#
```

MP-BGP EVPN の分散型エニーキャスト ゲートウェイ

MP-BGP EVPN では、同じ仮想ゲートウェイ IP アドレスと仮想ゲートウェイ MAC アドレスをサポートすることで、VNI 内の VTEP を IP サブネット内のエンド ホストの分散型エニーキャスト ゲートウェイにすることができます（図 9）。EVPN のエニーキャスト ゲートウェイ機能を使用すると、VNI 内のエンド ホストは、この VNI のローカル VTEP をデフォルト ゲートウェイとして使用して、IP サブネットの外部にトラフィックを送信できます。この機能により、VXLAN オーバーレイ ネットワークのエンド ホストからのノースバウンド トラフィックの最適な転送が可能になります。分散型エニーキャスト ゲートウェイは、VXLAN オーバーレイ ネットワークでのシームレス ホスト モビリティの利点も提供します。ゲートウェイ IP と仮想 MAC アドレスは VNI 内のすべての VTEP で同じようにプロビジョニングされるため、エンド ホストが VTEP から別の VTEP に移動する場合、ゲートウェイ MAC アドレスを再学習するために別の ARP 要求を送信する必要はありません。

図 9. MP-BGP EVPN の分散型エニーキャスト ゲートウェイ



MP-BGP EVPN での ARP 抑制

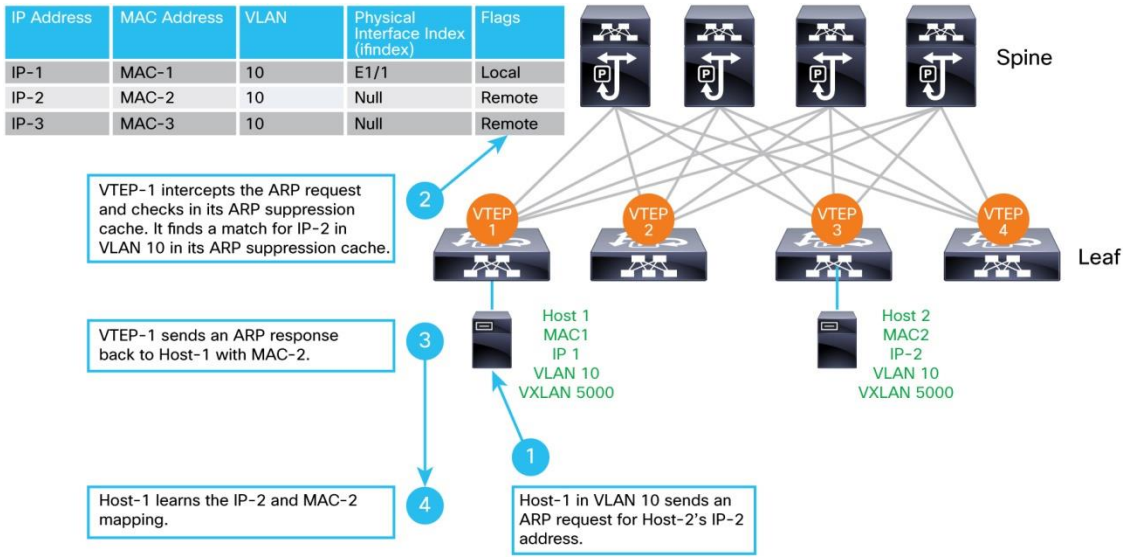
ARP 抑制は MP-BGP EVPN コントロール プレーンによって提供された拡張機能であり、ARP 要求からのブロードキャスト トラフィックによって生じるネットワーク フラッディングを軽減します。

VNI に対して ARP 抑制を有効にすると、その VTEP はそれぞれ、VNI セグメント内の既知の IP ホストとそれらに関連付けられた MAC アドレスの ARP 抑制キャッシュ テーブルを維持します。図 10 に示すように、VNI のエンド ホストが別のエンドホスト IP アドレスの ARP 要求を送信すると、そのローカル VTEP が ARP 要求をインターセプトし、ARP 抑制キャッシュ テーブル内の ARP された IP アドレスをチェックします。一致が見つかった場合、ローカル VTEP は、リモート エンド ホストに代わって ARP 応答を送信します。ローカル ホストは、ARP 応答でリモート ホストの MAC アドレスを学習します。ローカル VTEP の ARP 抑制テーブルに ARP された IP アドレスがない場合、VNI 内の他の VTEP に ARP 要求をフラッディングします。この ARP フラッディングは、ネットワーク内のサイレント ホストへの最初の ARP 要求で発生する可能性があります。ネットワーク内の VTEP は、別のホストがその IP アドレスの ARP 要求を送信し、ARP 応答を返送するまで、サイレント ホストからのトラフィックを認識しません。ローカル VTEP がサイレント ホストの MAC アドレスと IP アドレスを学習すると、その情報は MP-BGP EVPN

コントロール プレーンを介して他のすべての VTEP に配信されます。後続の ARP 要求はフラッディングする必要はありません。

ほとんどのエンドホストは、オンラインになった直後にネットワークにアナウンスするために GARP または RARP 要求を送信するため、ローカル VTEP はすぐに MAC および IP アドレスを学習し、この情報を MP-BGP EVPN コントロール プレーンを介して他の VTE Pに配布できます。したがって、VXLAN EVPN のほとんどのアクティブな IP ホストは、ローカル ラーニングまたはコントロールプレーンベースのリモート ラーニングのいずれかで VTEP によって学習されます。その結果、ARP 抑制により、ホスト ARP 学習動作によって引き起こされるネットワーク フラッディングが減少します。

図 10. MP-BGP EVPN での ARP 抑制



MP-BGP EVPN VTEP の設定

ここでは、MP-BGP EVPN VTEP の設定手順の概要を示します。

ステップ 1. 各 VTEP スイッチの初期設定を実行します。

Enable the VXLAN and MP-BGP EVPN control plane.

```
feature nv overlay
feature vn-segment-vlan-based
feature bgp
nv overlay evpn
```

- Enable VXLAN.
- Enable VLAN-based VXLAN (currently the only mode).
- Enable BGP.
- Enable the EVPN control plane for VXLAN.

Other features may need to be enabled.

```
feature ospf
feature pim
feature interface-vlan
```

- Enable Open Shortest Path First (OSPF) if it's chosen as the underlay Interior Gateway Protocol (IGP) routing protocol.
- Enable IP Protocol-Independent Multicast (PIM) routing in the underlay network.
- Enable VLAN switch virtual interfaces (SVIs) if the VTEP needs to be the IP gateway and route for the VXLAN VLAN IP subnet.

ステップ 2. EVPN テナント VRF インスタンスを設定します。

次に、2 つのテナント VRF インスタンスの設定例を示します。

```
vrf context evpn-tenant-1
  vni 39000
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
```

- Create a VXLAN tenant VRF instance.
- Specify the Layer 3 VNI for VXLAN routing for this tenant VRF instance.
- Define the VRF route distinguisher.
- Define the VRF route target import and export policies in address-family ipv4 unicast. This example uses route-target auto-generation for this VRF.

```
vrf context evpn-tenant-2
  vni 39010
  rd auto
  address-family ipv4 unicast
    route-target import 100:39010
    route-target import 100:39010 evpn
    route-target export 100:39010
    route-target export 100:39010 evpn
```

- Create a second tenant VRF instance following the preceding steps. The example uses manual configuration for route-target import and export policies.

ステップ 3. テナント VRF インスタンスごとにレイヤ 3 VNI を作成します。

```
vlan 3900
  name l3-vni-vlan-for-tenant-1
  vn-segment 39000

interface Vlan3900
  description l3-vni-for-tenant-1-routing
  no shutdown
  vrf member evpn-tenant-1
  ip forward

vrf context evpn-tenant-1
  vni 39000
  rd auto
  address-family ipv4 unicast
    route-target import 39000:39000
    route-target export 39000:39000
    route-target both auto evpn
```

Create the VLAN for the Layer 3 VNI. Create one Layer 3 VNI for each tenant VRF routing instance.

Create the SVI for the Layer 3 VNI. Put this SVI in the tenant VRF context. The command “ip forward” enables prefix-based routing for the VNI IP subnet. It’s needed to complete the initial routing to silent hosts in the VNI network.

Associate the Layer 3 VNI with the tenant VRF routing instance.

```
vlan 3901
  name l3-vni-vlan-for-tenant-2
  vn-segment 39010

interface Vlan3901
  description l3-vni-for-tenant-2-routing
  no shutdown
  vrf member evpn-tenant-2
  ip forward

vrf context evpn-tenant-2
  vni 39010
  rd auto
  address-family ipv4 unicast
    route-target import 39010:39010
    route-target export 39010:39010
    route-target both auto evpn
```

Define the Layer 3 VNI for a second tenant following the preceding steps.

ステップ 4. レイヤ 2 ネットワークの EVPN レイヤ 2 VNI を設定します。

この手順では、VLAN をレイヤ 2 VNI にマッピングし、EVPN パラメータを定義します。

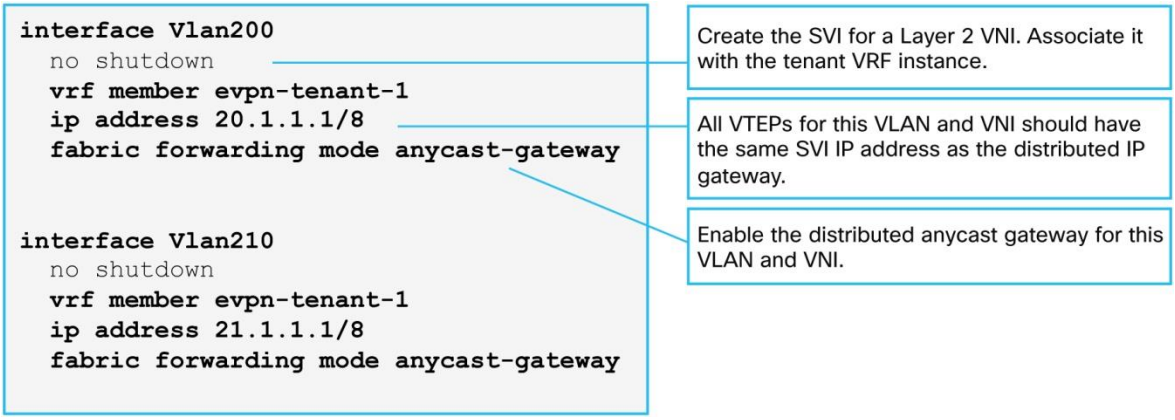
```
vlan 200
  vn-segment 20000
vlan 210
  vn-segment 21000
```

Map the VLAN to the VXLAN VNI.

```
evpn
  vni 20000 l2
    rd auto
    route-target import auto
    route-target export auto
  vni 21000 l2
    rd auto
    route-target import auto
    route-target export auto
```

Under the EVPN configuration, define the route distinguisher and route target import and export policies for each Layer 2 VNI.

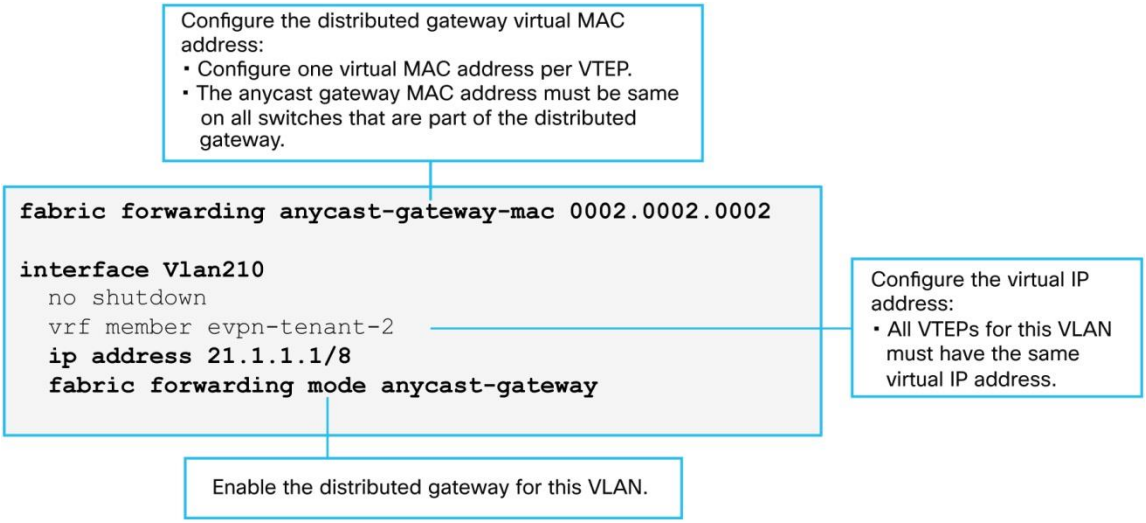
ステップ 5. レイヤ 2 VNI の SVI を設定し、SVI でエニーキャスト ゲートウェイを有効にします。



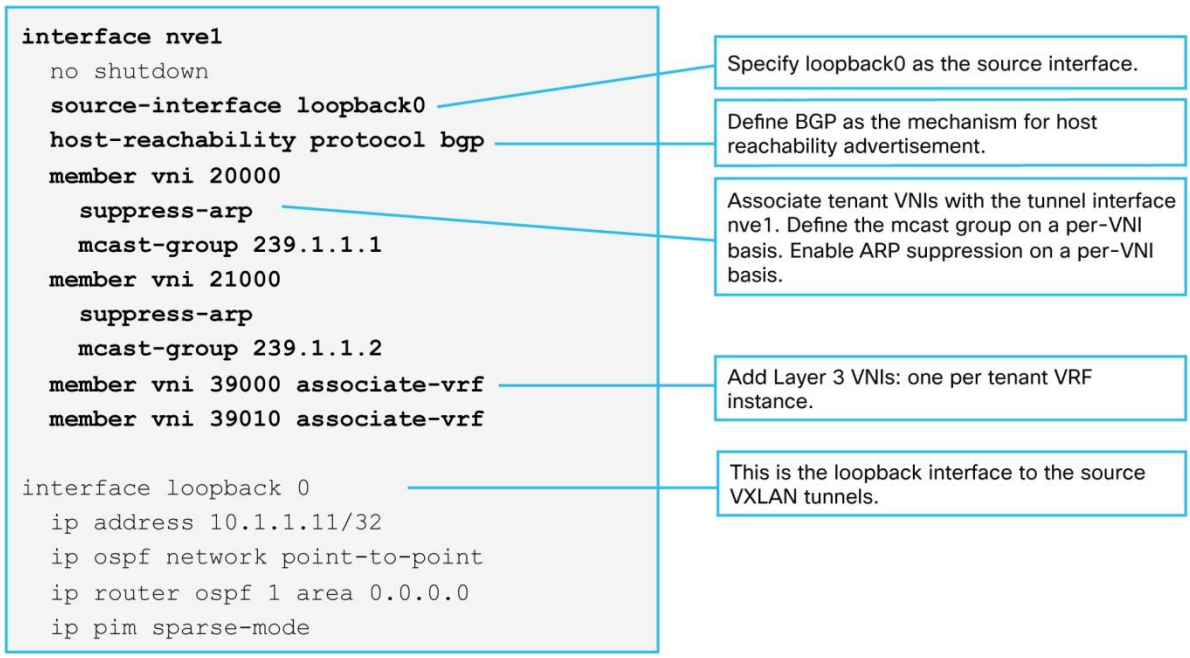
ステップ 6. EVPN 分散型エニーキャスト ゲートウェイを設定します。

この手順には、各 VTEP のエニーキャスト ゲートウェイ仮想 MAC アドレスと各 VNI のエニーキャスト ゲートウェイ IP アドレスの設定が含まれます。

EVPN ドメイン内のすべての VTEP は、デフォルト IP ゲートウェイとして機能する特定の VNI に対して、同じエニーキャスト ゲートウェイ仮想 MAC アドレスと同じエニーキャスト ゲートウェイ IP アドレスを持つ**必要があります**。



ステップ 7. VXLAN トンネル インターフェイス nve1 を設定し、レイヤ 2 VNI とレイヤ 3 VNI を関連付けます。



ステップ 8.　VTEP で MP-BGP を設定します。

```
router bgp 100
  router-id 10.1.1.11
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 10.1.1.1 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  neighbor 10.1.1.2 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended

  vrf evpn-tenant-1
    address-family ipv4 unicast
      advertise l2vpn evpn
  vrf evpn-tenant-2
    address-family ipv4 unicast
      advertise l2vpn evpn
```

Use address-family ipv4 unicast for prefix-based routing.

Use address-family l2vpn evpn for evpn host routes.

Define the MP-BGP neighbors. Under each neighbor, define address-family ipv4 unicast and l2vpn evpn.

Send extended community in address-family l2vpn evpn to distribute EVPN route attributes.

Under address-family ipv4 unicast for each tenant VRF instance, enable advertising for EVPN routes.

ステップ 9.　iBGP ルート リフレクタを設定します。

```
router bgp 100
  router-id 10.1.1.1
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
    retain route-target all
  template peer vtep-peer
    remote-as 100
    update-source loopback0
    address-family ipv4 unicast
      send-community both
      route-reflector-client
    address-family l2vpn evpn
      send-community both
      route-reflector-client
  neighbor 10.1.1.11
    inherit peer vtep-peer
  neighbor 10.1.1.12
    inherit peer vtep-peer
  neighbor 10.1.1.13
    inherit peer vtep-peer
  neighbor 10.1.1.14
    inherit peer vtep-peer
```

Use address-family ipv4 unicast for prefix-based routing.

Use address-family l2vpn evpn for EVPN VXLAN host routes. Retain all the route-target attributes.

Use an iBGP route-reflector client peer template.

Send both standard and extended communities in address-family ipv4 unicast.

Send both standard and extended communities in address-family l2vpn evpn.

MP-BGP EVPN VXLAN の仮想ポート チャネル VTEP

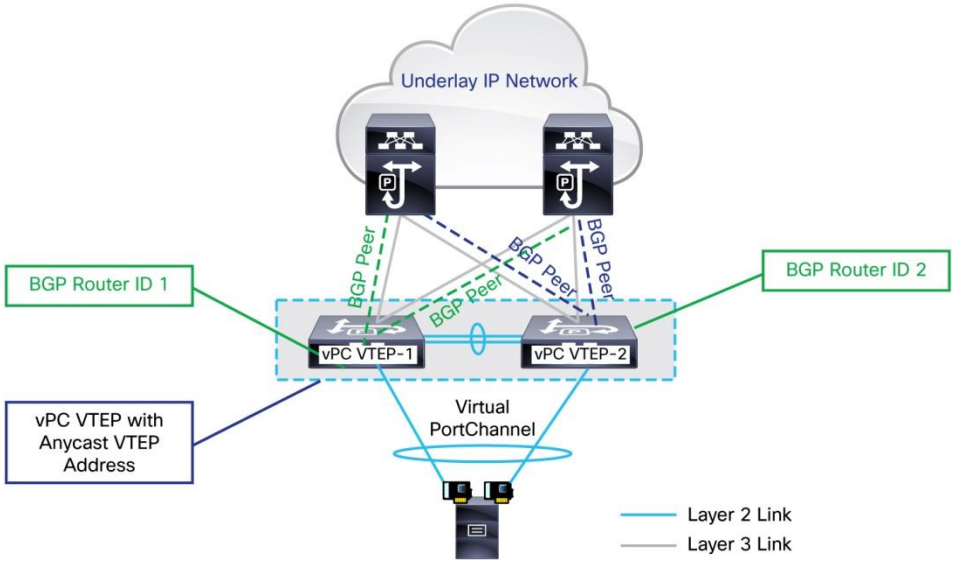
仮想ポートチャネル（vPC）VTEP は、vPC と VXLAN の 2 つのテクノロジーを組み合わせて、VTEP にデバイスレベルの冗長性を提供します。vPC スイッチのペアは、同じ VTEP アドレス（エニーキャスト VTEP アドレスとも呼ばれる）を共有し、論理 VTEP として機能します。ネットワーク内の他の VTEP は、エニーキャスト VTEP アドレスを持つ単一の VTEP として 2 つのスイッチを認識します。両方の vPC VTEP スイッチが稼働している場合、アクティブ/アクティブ設定でロード シェアします。1 つの vPC スイッチがダウンした場合、他のスイッチがトラフィック負荷全体を引き継ぐため、障害発生によって vPC ペアに接続されたデバイスの接続が失われることはありません。

Cisco NX-OS の MP-BGP EVPN コントロール プレーンは、vPC VTEP と透過的に動作するように実装されています。MP-BGP EVPN コントロール プレーンでは、vPC VTEP は VTEP 機能のエニーキャスト VTEP アドレスを持つ単一の論理 VTEP として機能し続けますが、MP-BGP の観点からは 2 つの個別のエンティティとして動作します。これらは BGP に対して異なるルータ ID を持ち、BGP ピアとの BGP ネイバー隣接関係を個別に形成し、EVPN ルートを個別にアドバタイズします。EVPN ルートでは、両方ともエニーキャスト VTEP アドレスをネクスト ホップとして使用するため、リモート VTEP は学習された EVPN ルートを使用し、カプセル化されたパケットの外部 IP ヘッダーの宛先としてエニーキャスト VTEP アドレスを使用してパケットをカプセル化します。

EVPN vPC VTEP 設定

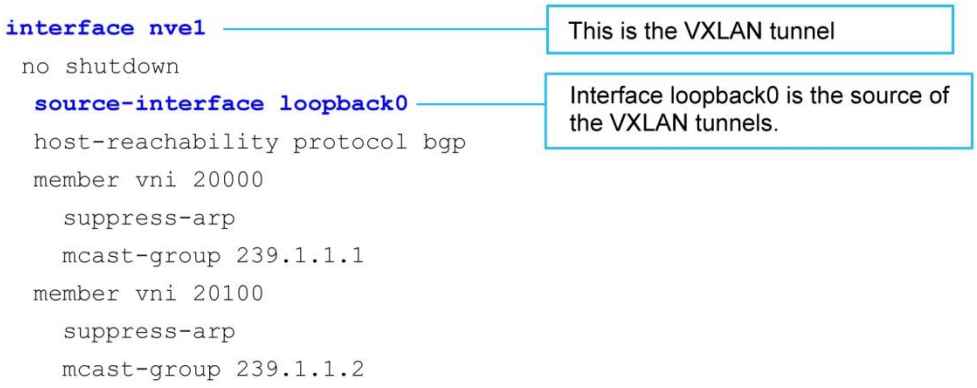
vPC VTEP スイッチは、VXLAN トンネル（interface nve1）の送信元の VTEP アドレスとして、ループバック インターフェイスのセカンダリ IP アドレスを使用するように設定されます。EVPN VXLAN 設定の残りの部分は、標準の単一 VTEP の場合と同じです。両方のスイッチに、固有のルータ ID を使用した独自の BGP 設定が必要です。図 11 に、MP-BGP EVPN vPC VTEP の概念を示します。MP-BGP は、EVPN ルートの BGP 更新を構築するときに、ネクスト ホップとしてエニーキャスト VTEP アドレスを使用します。

図 11. MP-BGP EVPN vPC VTEP



vPC VTEP 設定例を次に示します。

vPC VTEP-1 設定



```
  member vni 39000 associate-vrf

interface loopback0
  ip address 10.1.1.13/32
  ip address 10.1.1.134/32 secondary
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode

router bgp 100
  router-id 10.1.1.13
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 10.1.1.1 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  neighbor 10.1.1.2 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  vrf evpn-tenant-1
    address-family ipv4 unicast
      advertise l2vpn evpn
evpn
  vni 20000 l2
    rd auto
    route-target import auto
    route-target export auto
  vni 20100 l2
    rd auto
    route-target import auto
    route-target export auto
vrf context evpn-tenant-1
  rd auto
  address-family ipv4 unicast
    route-target import 39000:39000
    route-target export 39000:39000
    route-target both auto evpn

n9396-vPC-VTEP-1#
```

This secondary IP address is used as the anycast VTEP address. Both vPC VTEPs need to be configured with the exact same anycast VTEP address.

The BGP instance has its own router ID: 10.1.1.13.

vPC VTEP-2 設定

```
interface nve1
  no shutdown
  source-interface loopback0
  host-reachability protocol bgp
  member vni 20000
    suppress-arp
    mcast-group 239.1.1.1
  member vni 20100
    suppress-arp
    mcast-group 239.1.1.2

  member vni 39010 associate-vrf


interface loopback0
  ip address 10.1.1.14/32
  ip address 10.1.1.134/32 secondary
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode

router bgp 100
  router-id 10.1.1.14
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 10.1.1.1 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  neighbor 10.1.1.2 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  vrf evpn-tenant-1
    address-family ipv4 unicast
      advertise l2vpn evpn
evpn
  vni 20000 l2
    rd auto
    route-target import auto
```

This is the VXLAN tunnel interface.

Interface loopback0 is the source of the VXLAN tunnels.

This secondary IP address is used as the anycast VTEP address. Both vPC VTEPs need to be configured with the exact same anycast VTEP address.

The BGP instance has its own router ID: 10.1.1.14.

```
    route-target export auto
  vni 20100 l2
    rd auto
    route-target import auto
    route-target export auto
vrf context evpn-tenant-1
  rd auto
  address-family ipv4 unicast
    route-target import 39000:39000
    route-target export 39000:39000
    route-target both auto evpn


n9396-vPC-VTEP-2#
```

vPC VTEP MP-BGP ステータスと EVPN ルートの更新

MP-BGP ネイバーにとって、vPC VTEP は 2 つの個別のネイバーとして表示されます。次に、vPC VTEP の BGP ネイバーからの **show bgp l2vpn evpn summary** の出力例を示します。

```
spine-9508-1# sh bgp l2vpn evpn summary
BGP summary information for VRF default, address family L2VPN EVPN
BGP router identifier 10.1.1.1, local AS number 100
BGP table version is 75, L2VPN EVPN config peers 4, capable peers 4
13 network entries and 13 paths using 1716 bytes of memory
BGP attribute entries [12/1728], BGP AS path entries [0/0]
BGP community entries [0/0], BGP clusterlist entries [0/0]

Neighbor        V    AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
10.1.1.11       4   100    8247    8262       75    0    0    5d17h 6
10.1.1.12       4   100    8254    8259       75    0    0    1d08h 3
10.1.1.13       4   100    8258    8409       75    0    0    1d16h 2
10.1.1.14       4   100    8257    8455       75    0    0    1d16h 2
```

The two vPC VTEPs are shown as two separate BGP neighbors.

2 つの vPC VTEP は、BGP ネクスト ホップと同じエニーキャスト VTEP アドレスで EVPN ルートをアドバタイズします。2 つの vPC VTEP からのルート アドバタイズメントの例を次に示します。

VTEP-1 の場合

```
n9396-vPC-VTEP-1# sh bgp l2vpn evpn neighbors 10.1.1.1 advertised-routes

Peer 10.1.1.1 routes for address family L2VPN EVPN:
BGP table version is 94, local router ID is 10.1.1.13
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist, I-
injected
```

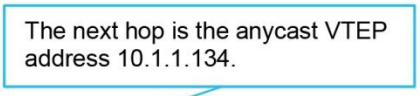
```
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup

  Network      Next Hop      Metric   LocPrf   Weight Path
Route Distinguisher: 10.1.1.11:32967

Route Distinguisher: 10.1.1.11:32968

Route Distinguisher: 10.1.1.11:32977

Route Distinguisher: 10.1.1.12:2

Route Distinguisher: 10.1.1.12:6

Route Distinguisher: 10.1.1.13:32967    (L2VNI 20000)
*>l[2]:[0]:[0]:[48]:[0000.1330.e586]:[0]:[0.0.0.0]/216
                    10.1.1.134                          100      32768 i
*>l[2]:[0]:[0]:[48]:[0000.1330.e586]:[32]:[20.0.0.98]/272
                    10.1.1.134                          100      32768 i

Route Distinguisher: 10.1.1.13:32977    (L2VNI 21000)

Route Distinguisher: 10.1.1.14:32967

Route Distinguisher: 10.1.1.13:3    (L3VNI 39000)

n9396-vPC-VTEP-1#
```

The next hop is the anycast VTEP address 10.1.1.134.

VTEP-2 の場合。

```
n9396-vPC-VTEP-2# sh bgp l2vpn evpn neighbors 10.1.1.1 advertised-routes

Peer 10.1.1.1 routes for address family L2VPN EVPN:
BGP table version is 117, local router ID is 10.1.1.14
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist, I-
injected
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup

   Network            Next Hop            Metric     LocPrf     Weight Path
Route Distinguisher: 10.1.1.11:32967

Route Distinguisher: 10.1.1.11:32968

Route Distinguisher: 10.1.1.11:32977

Route Distinguisher: 10.1.1.12:2
```

```
Route Distinguisher: 10.1.1.12:6

Route Distinguisher: 10.1.1.13:32967

Route Distinguisher: 10.1.1.14:32967    (L2VNI 20000)
*>l[2]:[0]:[0]:[48]:[0000.1330.e586]:[0]:[0.0.0.0]/216
                     10.1.1.134                          100      32768 i
*>l[2]:[0]:[0]:[48]:[0000.1330.e586]:[32]:[20.0.0.98]/272
                     10.1.1.134                          100      32768 i

Route Distinguisher: 10.1.1.14:32977    (L2VNI 21000)

Route Distinguisher: 10.1.1.14:3    (L3VNI 39000)

n9396-vPC-VTEP-2#
```



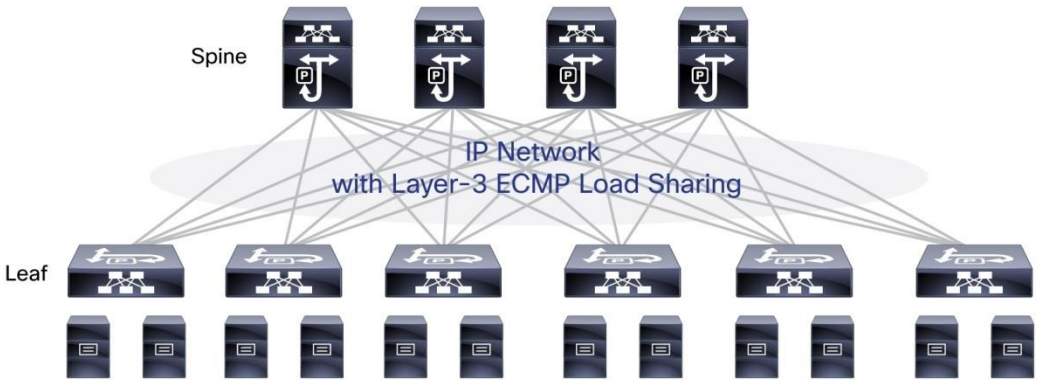
他の VTEP では、ネクストホップとしてエニーキャスト VTEP を使用して EVPN ルートが学習されます。次のスニペットは、前述の例でアドバタイズされたものと同じルートのリモート VTEP で、**show bgp l2vpn evpn** 出力からのものです。

```
Route Distinguisher: 10.1.1.14:32967
* i[2]:[0]:[0]:[48]:[0000.1330.e586]:[0]:[0.0.0.0]/216
                     10.1.1.134                          100          0 i
*>i                  10.1.1.134                          100          0 i
*>i[2]:[0]:[0]:[48]:[0000.1330.e586]:[32]:[20.0.0.98]/272
                     10.1.1.134                          100          0 i
* i                  10.1.1.134                          100          0 i
```

MP-BGP EVPN VXLAN ファブリックの設計

新しいスケーラブルなデータ センター ネットワークを展開する際に、2 層スパインアンドリーフ ファブリック アーキテクチャを検討する組織が増えています（図 12）。2 層ファブリック設計により、ネットワークの拡張に必要な柔軟性が提供され、接続密度と転送容量に対するアプリケーションの増え続ける要件に対応できます。ファブリックはレイヤ 3 ネットワークとして動作し、Open Shortest Path First（OSPF）、BGP、および Intermediate System to Intermediate System（IS-IS）などの既存のレイヤ 3 ルーティング プロトコルの実証済みの安定性と拡張性を利用します。

図 12. 2層スパイン リーフ ファブリック アーキテクチャ



レイヤ 3 ファブリックでは、レイヤ 2 ドメインが各リーフ スイッチの下に含まれます。コンピューティング ノード間の直接的なレイヤ 2
隣接関係を前提とするアプリケーションでは、この設計によってワークロードの配置が制限される可能性があります。VXLAN はレイヤ 3 ファブリック上でレイヤ 2
ドメインを拡張し、ワークロード配置の柔軟性を実現するために展開できます。ここでは、ルート配布とマルチテナンシーのサポートに MP-BGP EVPN コントロール プレーンを使用する VXLAN
ファブリックの一般的な設計オプションについて説明します。

MP-BGP EVPN は、BGP の新しいアドレス ファミリであり、アドレス ファミリに依存しない BGP のメカニズムを使用します。iBGP または eBGP
の使用は必須ではありません。この柔軟性により、組織は現在のデータ センター BGP 設計から MP-BGP EVPN VXLAN 設計に容易に移行できます。このアプローチでは、BGP
自律システム番号（ASN）の割り当ても柔軟に行えます。 このセクションでは、MP-iBGP EVPN および MP-eBGP EVPN 設計の両方を説明します。

MP-iBGP EVPN を使用した VXLAN ファブリック

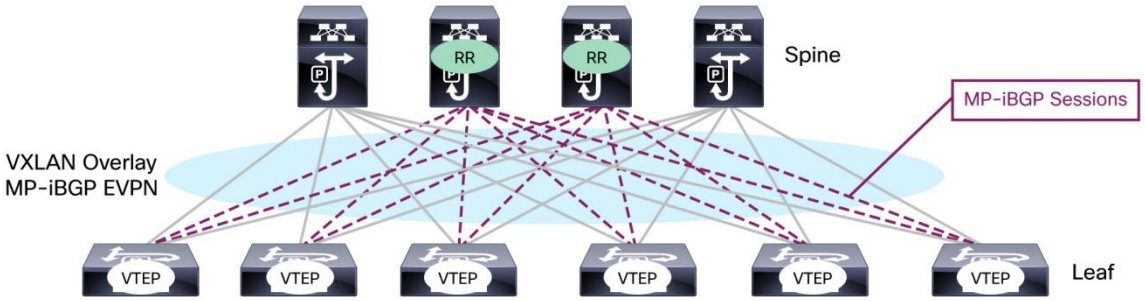
MP-iBGP EVPN 設計では、すべての MP-BGP スピーカーが同じ BGP 自律システム内にあります。iBGP ピアリング トポロジを簡素化するために、iBGP ルート
リフレクタがネットワークに導入されることがよくあります。アンダーレイ ネットワークの VTEP アドレスに IP 到達可能性を提供するために、選択した IGP ルーティング
プロトコルを導入できます。ソフトウェアの機能とスケーラビリティに応じて、iBGP ルート
リフレクタは、スパイン レイヤまたはリーフ
レイヤのいずれかに配置することも、専用デバイスに配置して拡張性を高めることもできます。

スパイン層の MP-iBGP ルート リフレクタ

この設計では、リーフ スイッチは VTEP デバイスです。MP-iBGP と、スパイン
スイッチで実行されているルートリフレクタのペアを持つピアを実行します。この設計では、選択したスパイン デバイスに MP-BGP EVPN ソフトウェア機能が必要ですが、VTEP である必要はありません。

図 13 は、スパイン レイヤに iBGP ルート リフレクタ（RR）を持つMP-iBGP EVPN VXLAN
ファブリックの例を示しています。この設計では、各 VTEP リーフに 2 つのスパイン BGP ルート
リフレクタである 2 つの iBGP ネイバーがあります。各スパイン BGP ルート リフレクタは、ルート リフレクタ クライアントとしてすべての VTEP リーフ ノードを持ち、VTEP リーフ ノードの EVPN ルートを反映します。

図 13. スパイン レイヤのルート リフレクタを使用した MP-iBGP EVPN VXLAN ファブリック設定



次の例は、この設計の VTEP リーフ ノードでの MP-iBGP 設定を示しています。

```
n9396-vtep-1# sh run bgp

!Command: show running-config bgp
!Time: Fri Jan 23 07:38:48 2015

version 7.0(3)I1(1)
feature bgp

router bgp 100
  router-id 10.1.1.11
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 10.1.1.1 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  neighbor 10.1.1.2 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended

  vrf evpn-tenant-1
    address-family ipv4 unicast
      advertise l2vpn evpn
evpn
  vni 20000 l2
    rd auto
    route-target import auto
    route-target export auto
  vni 20100 l2
```

Configure the two spine BGP route reflectors as two iBGP neighbors. Under each neighbor, send extended community in address-family l2vpn evpn. EVPN routes use extended community to carry EVPN attributes.

Advertise EVPN routes to address-family ipv4 unicast. This step is optional. It's needed in case this VTEP routes to an external device such as a WAN edge router and needs to distribute EVPN routes to the outside.

```
    rd auto
    route-target import auto
    route-target export auto
  vni 21000 l2
    rd auto
    route-target import auto
    route-target export auto
  vni 21100 l2
    rd auto
    route-target import auto
    route-target export auto
vrf context evpn-tenant-1
  rd auto
  address-family ipv4 unicast
    route-target import 39000:39000
    route-target export 39000:39000
    route-target both auto evpn
vrf context evpn-tenant-2
  rd auto
  address-family ipv4 unicast
    route-target import 39010:39010
    route-target export 39010:39010
    route-target both auto evpn


n9396-vtep-1#
```

次の例は、スパイン BGP ルート リフレクタの MP-iBGP 設定を示しています。

```
feature bgp
nv overlay evpn
router bgp 100
  router-id 10.1.1.1
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
    retain route-target all

 template peer vtep-peer
    remote-as 100
    update-source loopback0
    address-family ipv4 unicast
      send-community both
      route-reflector-client
    address-family l2vpn evpn
      send-community both
      route-reflector-client

  neighbor 10.1.1.11
    inherit peer vtep-peer
  neighbor 10.1.1.12
    inherit peer vtep-peer
  neighbor 10.1.1.13
    inherit peer vtep-peer
  neighbor 10.1.1.14
    inherit peer vtep-peer
```

Enable MP-BGP l2vpn evpn.

Use address-family l2vpn evpn for VXLAN EVPN routes. The original route-target attributes must be retained while advertising EVPN routes from one iBGP route-reflector client to the others. This requirement is important to allow the routes to be received by the other route-reflector clients.

Use the iBGP RR client peer template.

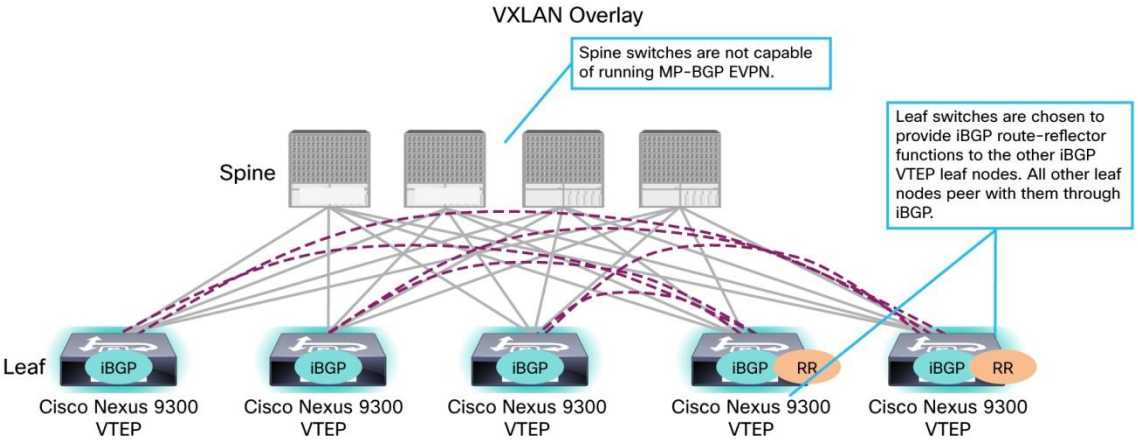
Send both standard and extended community in address-family l2vpn evpn.

VTEP leaf nodes are iBGP route-reflector clients.

リーフ レイヤの MP-iBGP ルート リフレクタ

スパイン レイヤへの BGP ルート リフレクタの配置は、MP-iBGP EVPN の直感的な設計です。選択したスパインデバイスがMP-iBGP EVPNプロトコルのソフトウェア機能をサポートし、EVPNルートのMP-iBGPアップデートを処理および配布できるようにする必要があります。スパイン デバイスが MP-BGP EVPN を実行できない場合は、BGP ルートリフレクタ機能をリーフ レイヤに移動する必要があります。リーフ スイッチでは MP-BGP EVPN および VTEP 機能がサポートされます（図 14）。

図 14. リーフ レイヤでの BGP ルータ リフレクタ機能を使用した MP-iBGP EVPN ファブリック設計



この設計では、スパイン スイッチは MP-BGP EVPN コントロール プレーンにまったく参加しません。アンダーレイ ネットワーク ルーティング プロトコルを実行して、VTEP

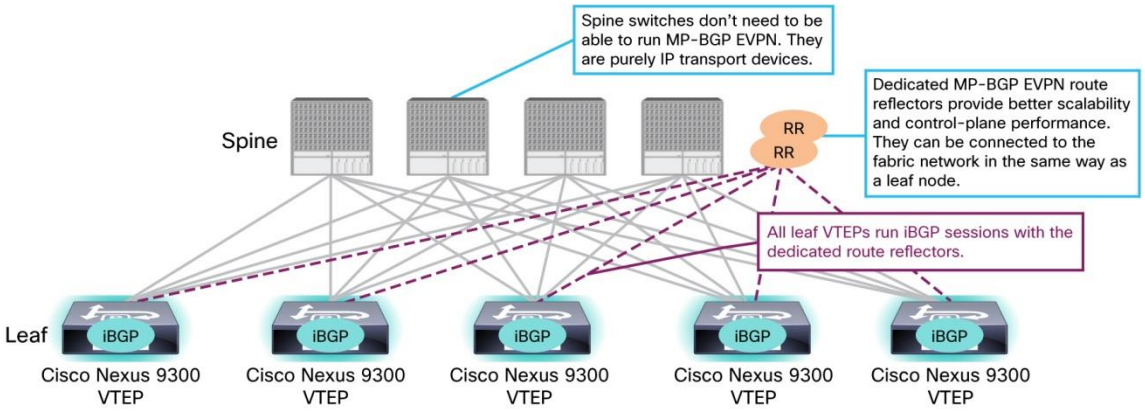
アドレスと iBGP ピアリング アドレスの IP 到達可能性を確立します（vPC VTEP など VTPC アドレスと同じでない場合）。

専用ルート リフレクタを使用した MP-iBGP

EVPN での MP-iBGP ルート リフレクタの役割は、標準的な iBGP ルート リフレクタと同じです。これは、iBGP ピア間の BGP 更新を反映し、完全メッシュの iBGP ピアリング トポロジを形成する必要がないようにすることです。このアプローチにより、iBGP トポロジが大幅に簡素化され、プロトコルのスケーラビリティが向上します。ルート リフレクタ機能は純粋なコントロールプレーン機能であるため、BGP ルート リフレクタをデータプレーン転送パスに配置する必要はありません。この機能により、ルート リフレクタの配置とプラットフォームの選択が非常に柔軟になります。

スケーラブルな設計のオプションは、データ パスからルート リフレクタとして専用デバイスを使用するためです（図 15）。選択したデバイスは MP-BGP EVPN をサポートし、高速コンバージェンスに必要な適切な BGP コントロールプレーン スケーラビリティとコンピューティング能力を備えている必要があります。専用ルート リフレクタを使用すると、スパイン レイヤでの MP-BGP EVPN 機能要件がなくなります。また、データ転送の実行に加えて、BGP ルートリフレクタ機能を実行する必要があるという VTEP リーフ ノードの負担が軽減されます。論理的に VTEP リーフ ノードにはルート リフレクタとの直接的な iBGP ネイバー隣接関係がありますが、ルート リフレクタはリーフ ノードと同じ方法で VXLAN ファブリック ネットワークに物理的に接続でき、VTEP リーフとルート リフレクタ間の iBGP セッションがファブリック アンダーレイ ネットワークの複数のホップ数（通常は 2）に移動できます。VTEP アドレス間のアンダーレイ データ パスがルート リフレクタを通過しないように、ルーティングの考慮事項を適用する必要があります。この要件は、ルート リフレクタがデータ転送パスから外れていることを確認するのに役立ちます。

図 15. 専用ルート リフレクタを使用した MP-iBGP EVPN の設計



MP-eBGP EVPN を使用した VXLAN ファブリック

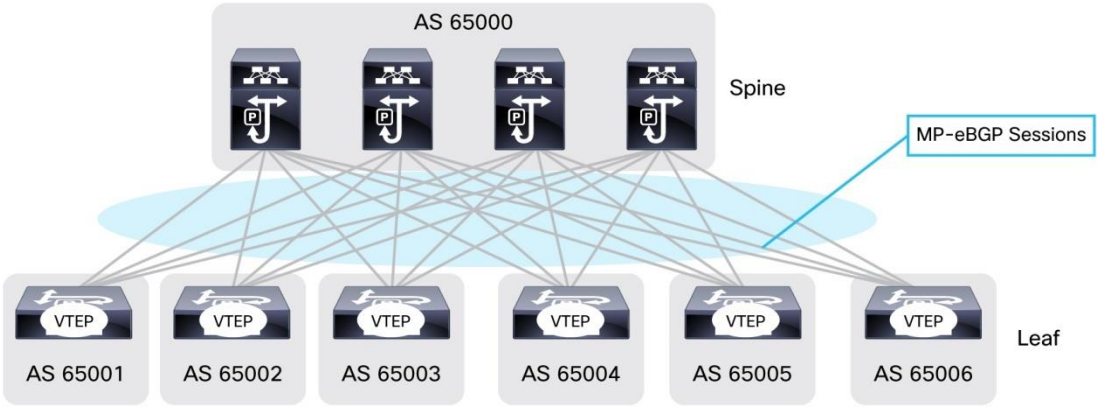
MP-iBGP EVPN 設計は一般的な方法ですが、一部の組織ではリーフ レイヤとスパイン レイヤ間で eBGP を実行することを選択します。MP-BGP EVPN は、iBGP と eBGP の両方で柔軟に動作します。MP-eBGP ピアリングを使用した EVPN は、実行可能な設計オプションです。eBGP 設計には、BGP 自律システム（AS）の割り当てにいくつかのオプションがあります。図 16 は、各 VTEP リーフが独自の BGP

AS 内にある設計を示しています。図 17 は、すべての VTEP リーフ ノードが同じ AS 内にあり、すべてがスパイン スイッチと eBGP を介してピアリングする別の設計を示しています。

MP-BGP EVPN は BGP の拡張であるため、標準の BGP 動作を継承します。MP-BGP EVPN ネットワークでは、一部のデフォルト動作は望ましくありません。たとえば、BGP ルータが eBGP ピアに BGP ルートをアドバタイズする場合、デフォルトでは BGP ネクストホップを自身の IP アドレスに変更します。MP-BGP EVPN では、VTEP が EVPN ルートをアドバタイズするために BGP 更新を開始すると、自身の VTEP アドレスを BGP ネクスト ホップとして使用します。このネクスト ホップは、他の VTEP が元の VTEP アドレスを持つ EVPN ルートをネクスト ホップとして受信し、このルートを使用してデータ プレーンで VXLAN トンネリングを開始できるように、ホップバイホップ BGP ルート配布全体で維持する必要があります。

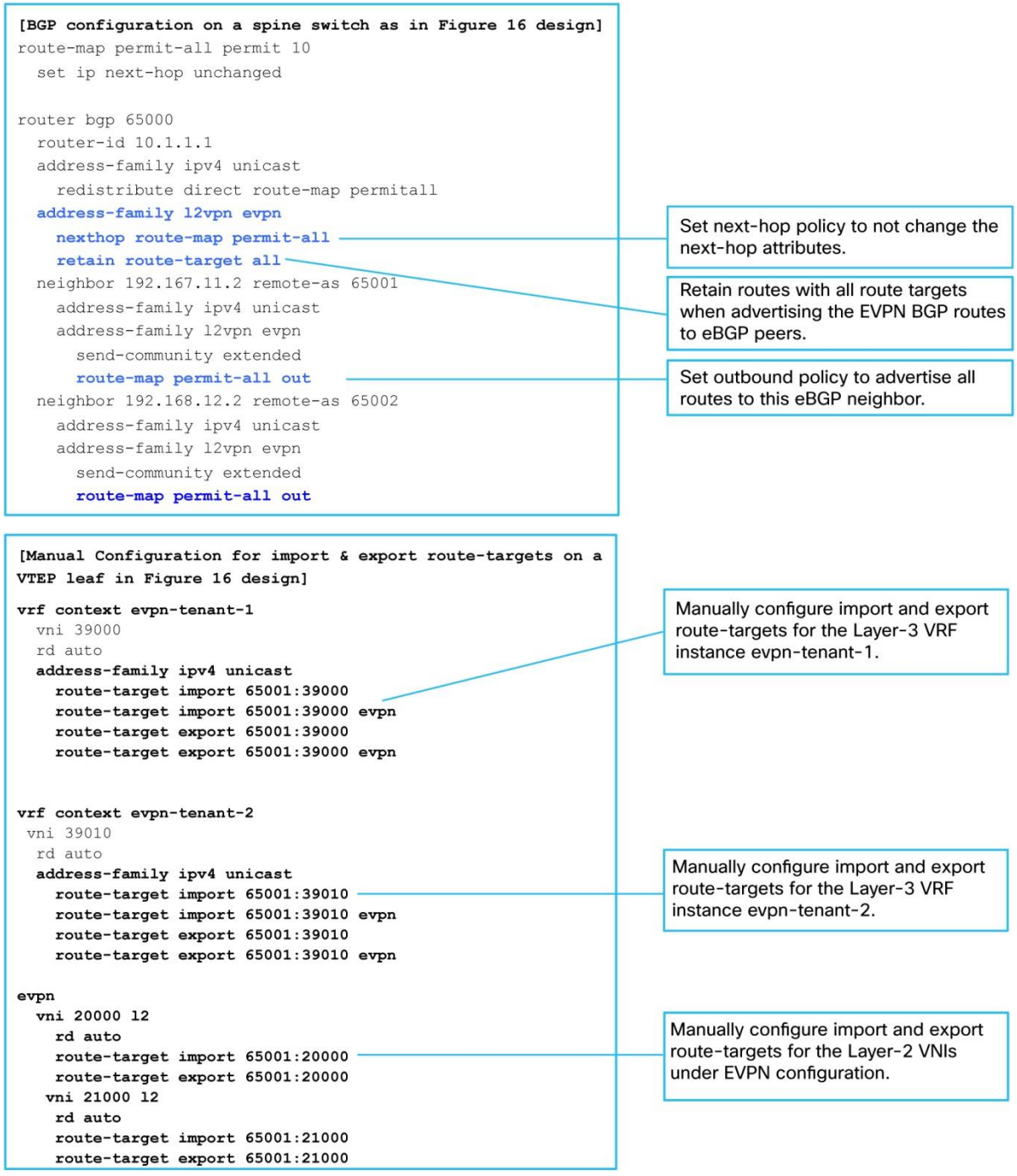
したがって、スパイン スイッチの eBGP は、BGP ネクスト ホップを変更しないように設定する必要があります。BGP ルータは、eBGP ルートを送信するときに BGP コミュニティ属性を変更することもできます。MP-EVPN では、この変更により EVPN ルートのルート ターゲット属性が変更または削除される可能性があります。したがって、すべてのルートターゲット属性を確実に保持するために、追加の設定を中間 eBGP ピアに適用する必要があります。

図 16. 固有の自律システムで VTEP リーフ ノードを使用した MP-eBGP EVPN VXLAN ファブリック



この設計ではすべての VTEP に固有の BGP AS があるため、NX-OS でのルートターゲット自動生成は、同じ VNI の VTEP で異なるルートターゲットになります。インポートおよびエクスポートのルート ターゲットを手動で設定し、VTEP が同じレイヤ 3 VR Fインスタンスと同じ EVPN レイヤ 3 VNI に対して同じルート ターゲット設定にすることを推奨します。

次の例は、図 16 に示すように、スパイン スイッチと VTEP リーフの MP-BGP 設定を示しています。スパイン スイッチの MP-BGP 設定には、eBGP ルート ネクスト ホップを変更しないように、スパイン スイッチのアウトバウンド ポリシーの適用が含まれます。この例は、レイヤ 3 VRF インスタンスと EVPN レイヤ 2 VNI の両方の VTEP リーフでの手動ルートターゲット設定も示しています。



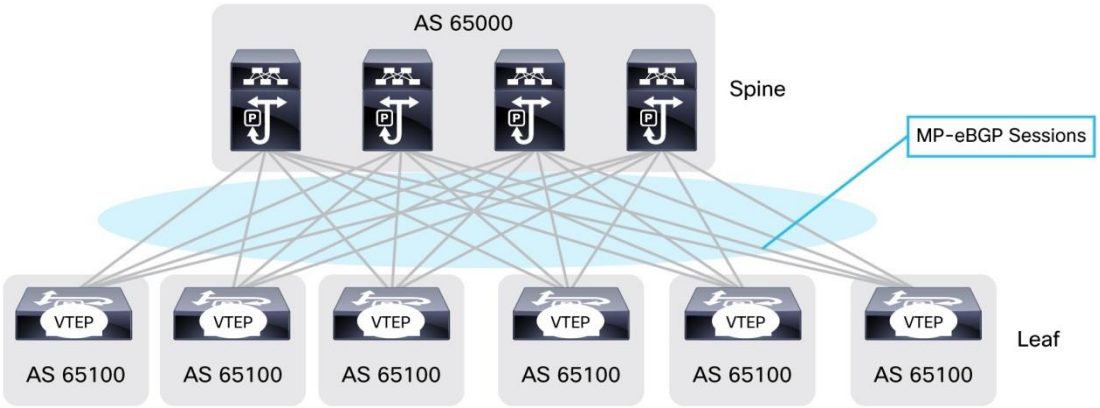
```
[BGP configuration on a leaf switch as in Figure 16
design]
route-map permit-all permit 10
  set ip next-hop unchanged

router bgp 65001
  address-family ipv4 unicast
  neighbor 192.167.11.1 remote-as 65000
    address-family ipv4 unicast
      allowas-in
      send-community extended
    address-family l2vpn evpn
      send-community extended
  neighbor 192.168.11.1 remote-as 65000
    address-family ipv4 unicast
      send-community extended
    address-family l2vpn evpn
      send-community extended
  vrf evpn-tenant-1
    address-family ipv4 unicast
      advertise l2vpn evpn
```

Spine switch 1 is an eBGP neighbor.

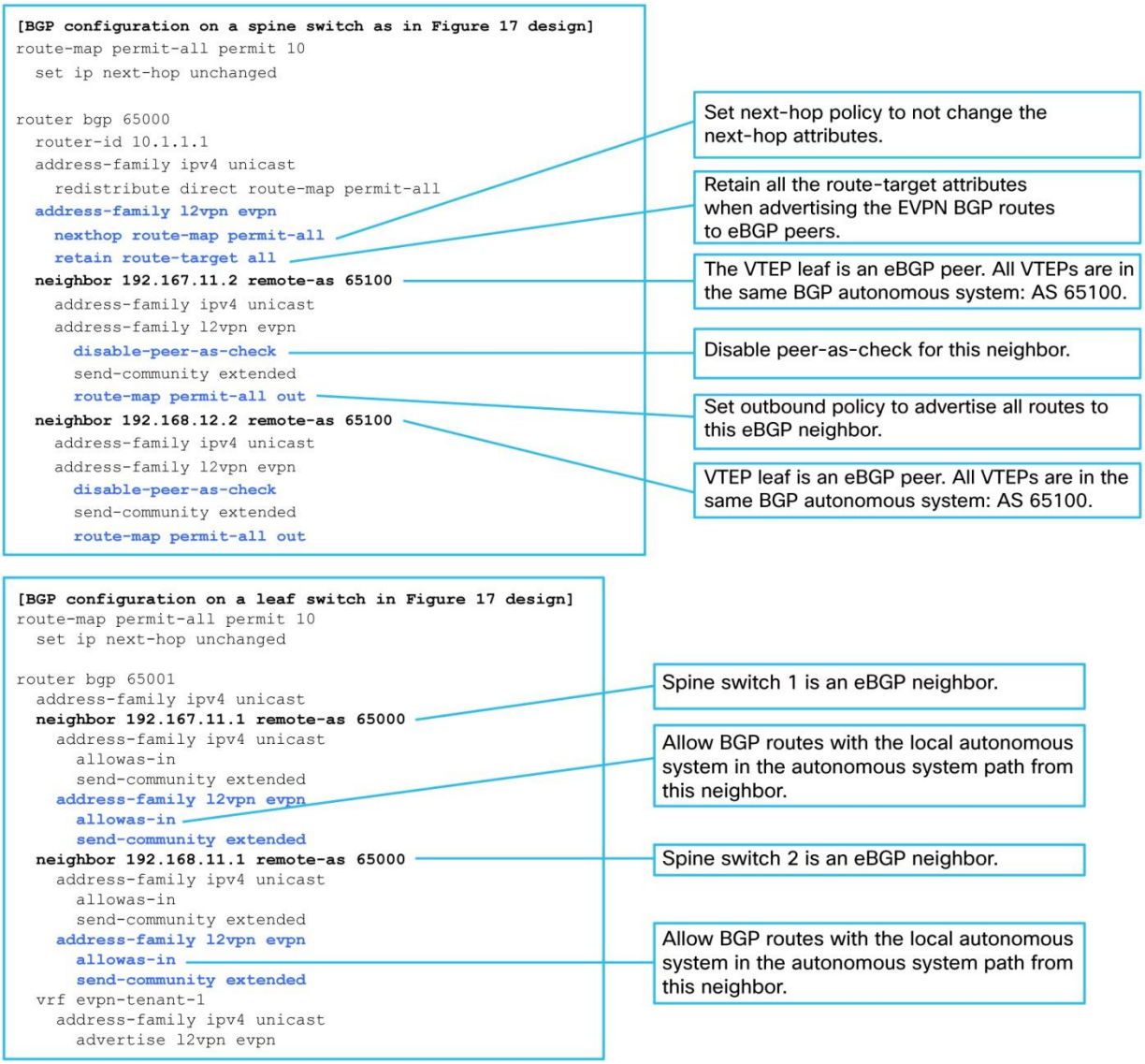
Spine switch 2 is an eBGP neighbor.

図 17 は、すべてのリーフ ノードが同じ自律システム内にあるが、それぞれが MP-eBGP を介してスパイン ノードとピアリングする MP-eBGP 設計を示しています。

図 17. 同じ BGP 自律システムでの VTEP リーフ ノードを使用した MP-eBGP の設計



次の例は、VTEP リーフおよびスパイン スイッチ設計の設定を示しています（図 17を参照）。図 16 の設計の設定に加えて、図 17 のスパイン スイッチは、同じ BGP **自律システム**内の 2 つの eBGP ネイバー間で MP-BGP EVPN ルートをパスする必要があるため、peer-as-check を無効にする必要があります。図 17 の VTEP リーフ ノードでは、同じ BGP **自律システム**内にある他の VTEP から BGP ルートを受け入れるように、allowas-in を有効にする必要があります。この設計では、すべての VTEP リーフが同じ BGP 自律システム内にあるため、レイヤ 3 VRF インスタンスおよび EVPN レイヤ 2 VNI のシステム自動生成インポートおよびエクスポート ルート ターゲットを使用するのに適しています。

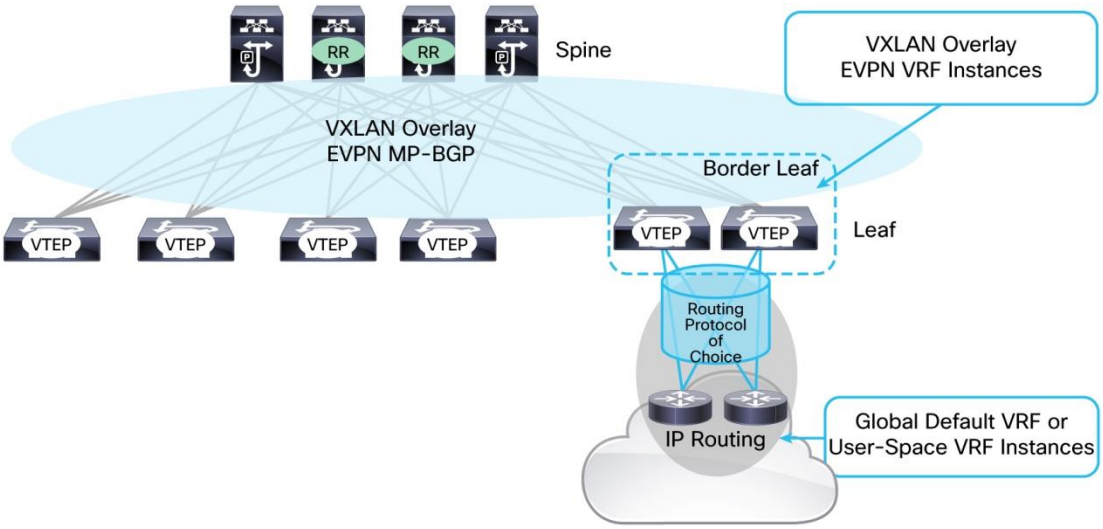


MP-BGP EVPN VXLAN の外部ルーティング

ほとんどの組織では、データ センターは、キャンパス
ネットワーク、WAN、インターネットなど、ネットワークの他の部分から分離されていません。EVPN VXLAN
ファブリックをデータ センターに展開する場合、VXLAN
ファブリックの外部にあるこれらのネットワークとの接続を維持する必要があります。

標準のスパインアンドリーフ ファブリック アーキテクチャでは、境界リーフ
ノードを使用して外部のルーティング デバイスに接続することで外部接続を実現できます。図 18 は、境界リーフ
スイッチのペアを使用したこのような設計を示しています。

図 18. MP-BGP EVPN VXLAN ファブリックの外部ルーティング用境界リーフ スイッチ



境界リーフ スイッチは、内部で VXLAN ファブリック内の他の VTEP と MP-BGP EVPN を実行し、それらと EVPN ルートを交換します。同時に、外部のルーティング デバイスを使用して、テナント VRF インスタンスで通常の IPv4 または IPv6 ユニキャスト ルーティングを実行します。ルーティング プロトコルには、通常の eBGP または任意の IGP を使用できます。設計上、MP-BGP EVPN は IPv4ま たは IPv6 ユニキャスト アドレス ファミリで学習された BGP ルートを L2VPN EVPN アドレス ファミリに自動的にインポートします。

したがって、境界リーフ スイッチは外部ルートを学習した後、それらを EVPN ドメインに EVPN ルートとしてアドバタイズできます。これにより、他の VTEP リーフ ノードも送信トラフィックを送信するために外部ルートを学習できます。
境界リーフ スイッチは、L2VPN EVPN アドレス ファミリで学習された EVPN ルートを IPv4 または IPv6 ユニキャスト アドレス ファミリに送信し、それらを外部ルーティング デバイスにアドバタイズするように設定することもできます。したがって、VXLAN ファブリックにパブリック サブネットが存在する場合は、それらを外部にアドバタイズして、外部からこれらのパブリック サブネットへの受信トラフィックを VXLAN ファブリックにルーティングできるようにします。

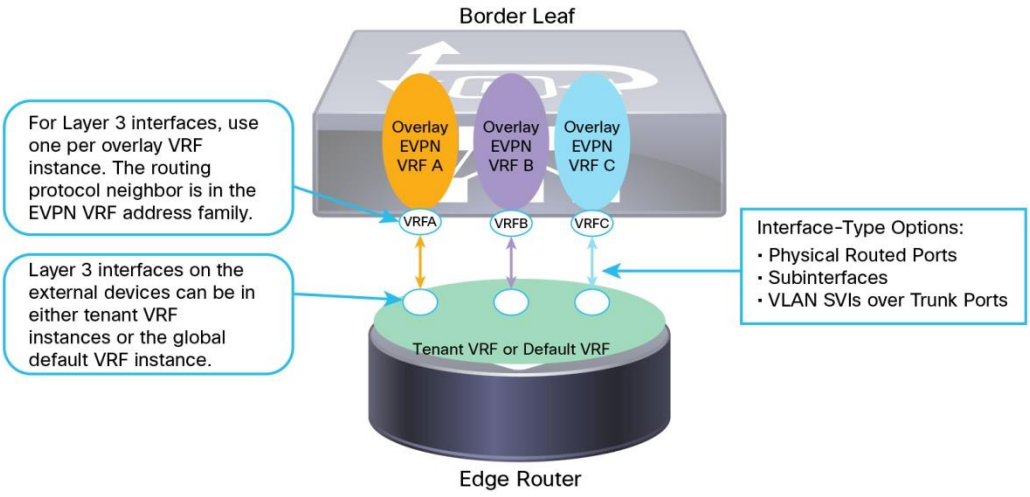
MP-BGP EVPN にはマルチテナント機能が組み込まれているため、VXLAN オーバーレイ ネットワーク内のレイヤ 3 サブネットはテナント VRF ルーティング インスタンス内にあります。異なるテナントは、デフォルトで個別のレイヤ 3 ルーティング インスタンスを維持できます。したがって、異なるテナントの外部ルーティングを個別に提供する必要があります。境界リーフには、外部ルーティングを実行するテナント VRF インスタンスごとに、外部へのレイヤ 3 インターフェイスが必要です（図 19）。

異なるテナント間のこのようなレイヤ 3 ルーティング セグメンテーションを外部ネットワークに拡張するために、外部ルータはテナント VRF インスタンスに境界リーフのレイヤ 3 インターフェイスを配置することもできます。境界リーフと外部ルータ間のルーティング セッションは、両側の VRF-Lite で実行されます。

VXLAN 境界リーフでレイヤ 3 セグメンテーションを終端する設計では、外部ルータはデフォルト ルーティング テーブル内のすべてのルーティング セッションを実行できます。この場合、VXLAN

ファブリック内の異なるテナント ルーティング インスタンスからのルートは、外部の同じデフォルト ルーティング
テーブルにマージされます。このタイプの設計では、テナントは基本的に外部ルーティングを共有するため、VXLAN テナントの IP アドレスは重複できません。

図 19. マルチテナンシーを使用した MP-BGP EVPN VXLAN ファブリック外部ルーティング



VXLAN EVPN 境界リーフと外部ルータ間の eBGP の設定例

次に、VXLAN 境界リーフと外部ルータ間の eBGP ルーティングの設定例を示します。eBGP セッションは境界リーフ上のテナント VRF
インスタンスにありますが、共有外部ルーティングの外部ルータのデフォルト ルーティング テーブルにあります。

境界リーフでは、VXLAN IP サブネット プレフィックスをアドバタイズするように BGP が設定されます。デフォルトでは、BGP は MP-BGP EVPN IP ホスト
ルートをアドバタイズします。プレフィックス ルートのみが外部ルータにアドバタイズされるように、/32 IP ホスト ルートをブロックするために、サンプル設定でルート
フィルタリングが適用されます。外部は受信トラフィック用の特定のホスト
ルートを必要としないため、このアプローチにより外部ルーティングのルータ スケーラビリティが向上します。

VXLAN 境界リーフ：

```
router bgp 100
  router-id 10.1.1.16
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 10.1.1.1 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  neighbor 10.1.1.2 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  vrf evpn-tenant-1
    address-family ipv4 unicast
      network 20.0.0.0/24
    neighbor 30.10.1.2 remote-as 200
      address-family ipv4 unicast
       prefix-list outbound-no-hosts out

ip prefix-list outbound-no-hosts seq 5 deny 0.0.0.0/0 eq 32
ip prefix-list outbound-no-hosts seq 10 permit 0.0.0.0/0 le 32
```

The eBGP neighbor is on the outside. It's in address-family ipv4 unicast of the tenant VRF routing instance.

For better scalability, apply prefix-list to filter out /32 IP host routes. Advertise prefix routes only to the external eBGP neighbor.

外部ルータでの BCP 設定：

```
router bgp 200
  router-id 10.1.1.18
  log-neighbor-changes
  address-family ipv4 unicast
    network 100.0.0.0/24
    network 100.0.1.0/24
  neighbor 30.10.1.1 remote-as 100
    address-family ipv4 unicast
```

前述の例では、VNI サブネット ルート 20.0.0.0/24 は、次のグローバル ルーティング テーブルように VRF-lite eBGP を介して外部ルータにアドバタイズされます。

```
N9372TX-2-ext# sh ip bgp 20.0.0.0/24
BGP routing table information for VRF default, address family IPv4 Unicast
BGP routing table entry for 20.0.0.0/24, version 36
Paths: (1 available, best #1)
Flags: (0x00001a) on xmit-list, is in urib, is best urib route

  Advertised path-id 1
  Path type: external, path is valid, is best path, no labeled nexthop
  AS-Path: 100 , path sourced external to AS
    30.10.1.1 (metric 0) from 30.10.1.1 (20.0.0.1)
      Origin IGP, MED not set, localpref 100, weight 0

  Path-id 1 not advertised to any peer
N9372TX-2-ext#
N9372TX-2-ext# sh ip route 20.0.0.0/24
IP Route Table for VRF "default"

20.0.0.0/24, ubest/mbest: 1/0
    *via 30.10.1.1, [20/0], 1w2d, bgp-200, external, tag 100
N9372TX-2-ext#
```

外部ルータから学習したルートは、MP-BGP EVPN プロトコルを介して境界リーフによって VXLAN ファブリックに配布されます。次の例は、内部 VTEP での外部ルートのキャプチャを示しています。VTEP は、ルート リフレクタを介して境界リーフから外部ルートを学習します。ルートは MP-BGP EVPN を介して配布されます。

```
n9396-vtep-1# sh vrf evpn-tenant-1 detail
VRF-Name: evpn-tenant-1, VRF-ID: 3, State: Up
    VPNID: unknown
    RD: 10.1.1.11:3
    VNI: 39000
    Max Routes: 0  Mid-Threshold: 0
    Table-ID: 0x80000003, AF: IPv6, Fwd-ID: 0x80000003, State: Up
    Table-ID: 0x00000003, AF: IPv4, Fwd-ID: 0x00000003, State: Up

n9396-vtep-1#

n9396-vtep-1# sh bgp l2vpn evpn rd 10.1.1.11:3 100.0.0.0
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 10.1.1.11:3    (L3VNI 39000)
BGP routing table entry for [5]:[0]:[0]:[24]:[100.0.0.0]:[0.0.0.0]/224, version 324
Paths: (1 available, best #1)
Flags: (0x00001a) on xmit-list, is in l2rib/evpn

  Advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop
             Imported from 10.1.1.16:3:[5]:[0]:[0]:[24]:[100.0.0.0]:[0.0.0.0]/120
  AS-Path: NONE, path sourced internal to AS
    10.1.1.16 (metric 3) from 10.1.1.1 (10.1.1.1)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 39000
      Extcommunity:  RT:100:39000 ENCAP:8 Router MAC:6412.2574.6ae7
      Originator: 10.1.1.16 Cluster list: 10.1.1.1

  Path-id 1 not advertised to any peer

n9396-vtep-1#
```

The external route is distributed through EVPN and imported into the tenant VRF instance.

```
n9396-vtep-1# sh ip bgp vrf evpn-tenant-1 100.0.0.0
BGP routing table information for VRF evpn-tenant-1, address family IPv4 Unicast
BGP routing table entry for 100.0.0.0/24, version 70
Paths: (1 available, best #1)
Flags: (0x08041a) on xmit-list, is in urib, is best urib route
  vpn: version 75, (0x100002) on xmit-list

  Advertised path-id 1, VPN AF advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop
             Imported from unknown dest
  AS-Path: NONE, path sourced internal to AS
    10.1.1.16 (metric 3) from 10.1.1.1 (10.1.1.1)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 39000
      Extcommunity:  RT:100:39000 ENCAP:8 Router MAC:6412.2574.6ae7
      Originator: 10.1.1.16 Cluster list: 10.1.1.1

  VRF advertise information:
  Path-id 1 not advertised to any peer

  VPN AF advertise information:
  Path-id 1 not advertised to any peer
n9396-vtep-1#
n9396-vtep-1# sh ip route vrf evpn-tenant-1 100.0.0.0/24
IP Route Table for VRF "evpn-tenant-1"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

100.0.0.0/24, ubest/mbest: 1/0
    *via 10.1.1.16%default, [200/0], 01:01:14, bgp-100, internal, tag 100 (evpn)segid: 0x9858 tunnelid:
0xa010110 encap: 1

n9396-vtep-1#
```

This is the external route.

The next hop is the VTEP address of the border leaf.

The tenant is VRF L3 VNI.

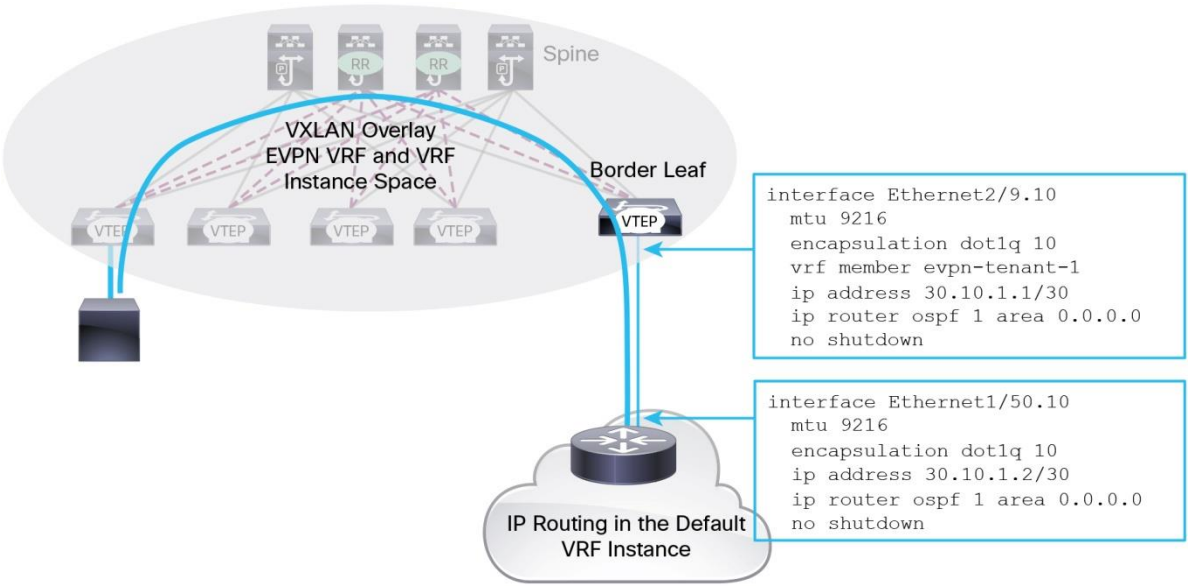
10.1.1.16 is the BGP router ID of the border leaf. 10.1.1.1 is the spine route reflector.

This is the iBGP route. The next hop is the VTEP address of the border leaf.

VXLAN EVPN 境界リーフと外部ルータ間の OSPF の設定例

図 20 の例では、EVPN VXLAN 境界リーフの外部ルーティング プロトコルとして OSPF を使用して、外部とルートを交換します。マルチテナンシーの場合、例では境界リーフと外部ルータ間のルーティングにサブインターフェイスを使用します。サブインターフェイスを使用すると、複数のテナントが外部ルーティング用に同じ物理リンクを共有でき、境界リーフ上のテナント VRF ルーティング インスタンスごとに 1 つのサブインターフェイスを使用できます。この例では、外部ルータのルーティングはデフォルト VRF インスタンスにあります。また、VRF-Lite サブインターフェイスを設定することで、外部デバイスのテナント VRF インスタンスを拡張することもできます。

図 20. OSPF を使用した EVPN VXLAN 外部ルーティング



境界リーフの関連する設定を次に示します。

```
ip prefix-list bgp-ospf-no-hosts seq 5 permit 0.0.0.0/0 eq 32
route-map permit-bgp-ospf deny 5
  match ip address prefix-list bgp-ospf-no-hosts
route-map permit-bgp-ospf permit 10
route-map permit-ospf-bgp permit 10

router ospf 1
  router-id 10.1.1.16
  vrf evpn-tenant-1
    redistribute bgp 100 route-map permit-bgp-ospf

router bgp 100
  router-id 10.1.1.16
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
    retain route-target all
  neighbor 10.1.1.1 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  neighbor 10.1.1.2 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  vrf evpn-tenant-1
    address-family ipv4 unicast
      advertise l2vpn evpn
      redistribute ospf 1 route-map permit-ospf-bgp
```

Redistribute BGP routes to OSPF. Filter out /32 IP host routes.

A BGP router will modify route targets in l2vpn evpn routes when it is an autonomous system boundary router. The original route target must be retained.

Redistribute OSPF to BGP. Advertise the redistributed routes to L2VPN EVPN.

この設計では、境界リーフはテナント VRF インスタンスの OSPF を介して外部ルートを学習します。VRF インスタンス内の MP-BGP にルートを再配布し、MP-BGP L2VPN EVPN を介して内部 VTEP にアドバタイズします。

次に、境界リーフでの外部ルート配布の例を示します。

```
n9396-border-leaf# sh ip route 100.0.0.0/24 vrf evpn-tenant-1
IP Route Table for VRF "evpn-tenant-1"

100.0.0.0/24, ubest/mbest: 1/0
    *via 30.10.1.2, Eth2/9.10, [110/2], 01:43:07, ospf-1, intra

n9396-border-leaf# sh bgp l2vpn evpn 100.0.0.0 vrf evpn-tenant-1
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 10.1.1.16:3    (L3VNI 39000)
BGP routing table entry for [5]:[0]:[0]:[24]:[100.0.0.0]:[0.0.0.0]/224, version 325
Paths: (1 available, best #1)
Flags: (0x00000a) on xmit-list, is not in l2rib/evpn

  Advertised path-id 1
  Path type: local, path is valid, is best path, no labeled nexthop
  AS-Path: NONE, path locally originated
    10.1.1.16 (metric 0) from 0.0.0.0 (10.1.1.16)
      Origin IGP, MED not set, localpref 100, weight 32768
      Received label 39000
      Extcommunity:  RT:100:39000

  Path-id 1 advertised to peers:
    10.1.1.1          10.1.1.2

n9396-border-leaf#
```

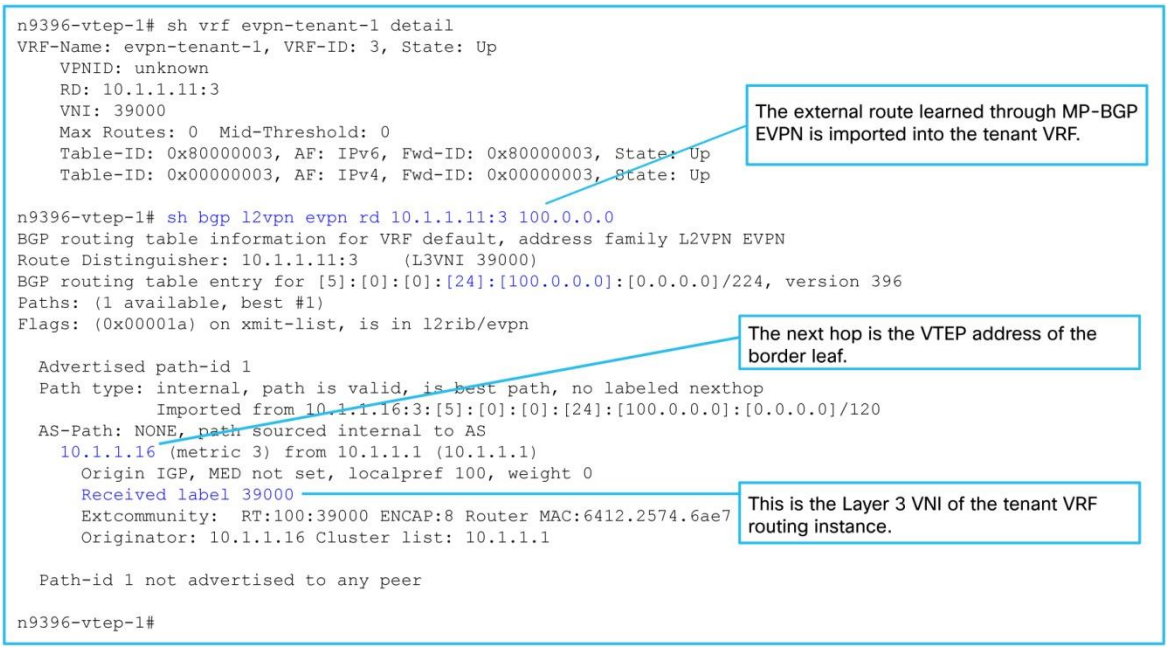
This is an external route learned through OSPF in the tenant VRF.

The external OSPF route is redistributed to BGP and distributed to other VTEPs through MP-BGP L2VPN EVPN.

The BGP next hop is the VTEP address of the border leaf.

The MP-BGP EVPN route is advertised to the BGP peers.

内部 VTEP は MP-BGP EVPN を介して外部ルートを学習します。



EVPN VXLAN 境界リーフ ノードのスケーラビリティに関する考慮事項

VXLAN 境界リーフノードは、外部への VXLAN ファブリック
ネットワークの接続ポイントです。外部ルートを学習し、MP-BGP EVPN を介して他の VTEP
に再配布します。同時に、VXLAN ファブリック上にあるパブリック サブネットの外部にアドバタイズします。

EVPN VXLAN ファブリックへの外部ルートの配布

境界リーフは、外部から多数の外部ルートを受信する場合があります。境界リーフ
ノードは通常ファブリックの内部デバイスの出口ゲートウェイであるため、すべての外部ルートをファブリック
に配布する必要はありません。代わりに、MP-BGP EVPN
にアドバタイズする前にルートを集約することもできます。場合によっては、テナントごとにファブリックへの
デフォルト ルートをアドバタイズするだけで十分です。分散外部ルートの数を減らすことで、内部 VTEP
デバイスで最長プレフィックス一致（LPM）ルーティング テーブル
リソースが不足しないようにすることができます。このアプローチでは、内部 VTEP での MP-BGP EVPN
コントロール プレーンの負荷も軽減されるため、コントロール プレーンのパフォーマンスが向上します。

外部への EVPN VXLAN ファブリック内部ネットワーク アドバタイズメント

EVPN VXLAN オーバーレイ ネットワークの一部であるレイヤ 3
サブネットは、外部から到達可能である必要があります。境界リーフノードは、これらのパブリック
サブネットのレイヤ 3 到達可能性情報をアドバタイズする必要があります。MP-BGP EVPN は、外部で IP ホスト
ルートと内部サブネット プレフィックス
ルートの両方を配布できます。境界リーフと外部ルータ間のルーティング プロトコル
セッションでは、フィルタを適用して、内部 IP ホスト
ルートを外部に送信しないようにすることができます。ほとんどの場合、外部サブネットが VXLAN
ファブリックにトラフィックを送信するために必要なのは、パブリック サブネットの LPM プレフィックス
ルートです。

境界リーフ ノードでの EVPN テナントのスケーラビリティ

境界リーフは、VXLAN オーバーレイ ネットワークのテナントに外部接続を提供します。境界リーフ ノードとして機能するすべてのテナント VRF ルーティング インスタンスに参加する必要があります。大規模なマルチテナント設計を構築する場合は、境界リーフがサポートできる EVPN レイヤ 3 VRF インスタンスの最大数の要件に従います。

境界リーフ ノードでの IP ホスト ルートのスケーラビリティ

内部エンド ホストを宛先とする着信トラフィックの最適な転送を実現するために、境界リーフはテナント パブリック サブネット内のエンド ホストに対して IP ホストベースのルーティングを実行する必要があります。この要件は、境界リーフが IP ホスト ルートのハードウェア転送テーブルでホスト ルートを学習およびプログラムする必要があることを意味します。IP ホスト テーブル サイズによって、テナントのパブリック サブネットに存在できるエンド ホストの総数が決まります。

MP-BGP EVPN VXLAN のデータ センター インターコネクト

オーバーレイ トランスポート仮想化（OTV）と仮想プライベート LAN サービス（VPLS）は、最も実績のあるレイヤ 2 データ センターインターコネクト（DCI）ソリューションですが、MP-BGP EVPN コントロール プレーンを備えた VXLANは、特定の展開条件下で代替手段を提供できます。VXLAN をデータ センター内に導入すると、データ センター間の相互接続に VXLAN を使用することで、ネットワーク設計全体が簡素化され、運用の複雑さが軽減され、データ センター内およびデータ センター間のトラフィックにユニファイド ネットワーク オーバーレイ ソリューションが提供されます。

図 21 は、MP-BGP EVPN VXLAN を使用したシンプルなデータ センターと DCI 設計を示しています。この設計では、各データ センターが独自の BGP 自律システムを維持し、簡素化とスケーラビリティのため、ルート リフレクタとともに MP-iBGP を実行する EVPN VXLAN ファブリックを展開します。データ センター間では、DCI 境界リーフ ノードが相互にマルチホップ MP-eBGP EVPN を実行します。その結果、2 つのデータ センターが結合され、1 つの統合 MP-BGP EVPN ルーティング ドメインが形成されます。コントロール プレーンでは、EVPN ルートはデータセンター間の iBGP-eBGP-iBGP パスを介して配布されます。データ プレーンでは、データ センター A のエンド ホストがデータ センター B の別のホストにトラフィックを送信すると、データ パケットは 1 つの VXLAN トンネルを通過し、データ センター A の入力 VTEP によってカプセル化され、データ センター B の出力 VTEP によってカプセル化解除されます。 このアプローチは、オーバーレイ ネットワークで非常に効果的な DCI データ転送を提供します。

図 21. Unified MP-BGP EVPN 管理ドメインを使用した DCI ソリューション

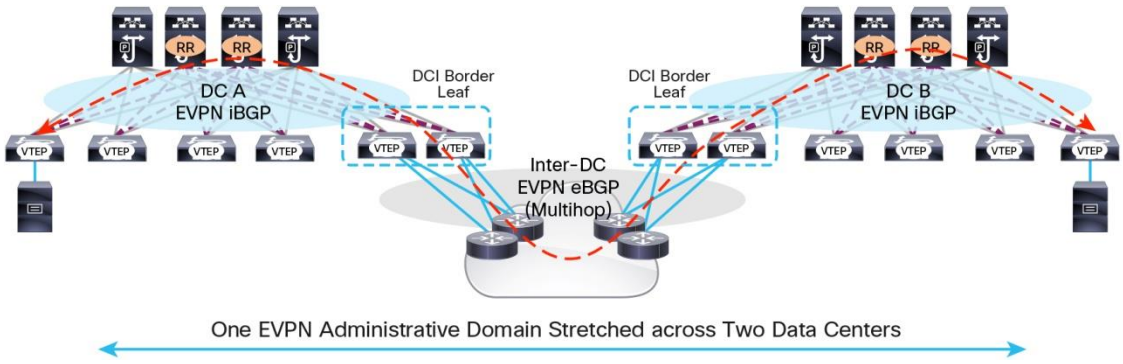
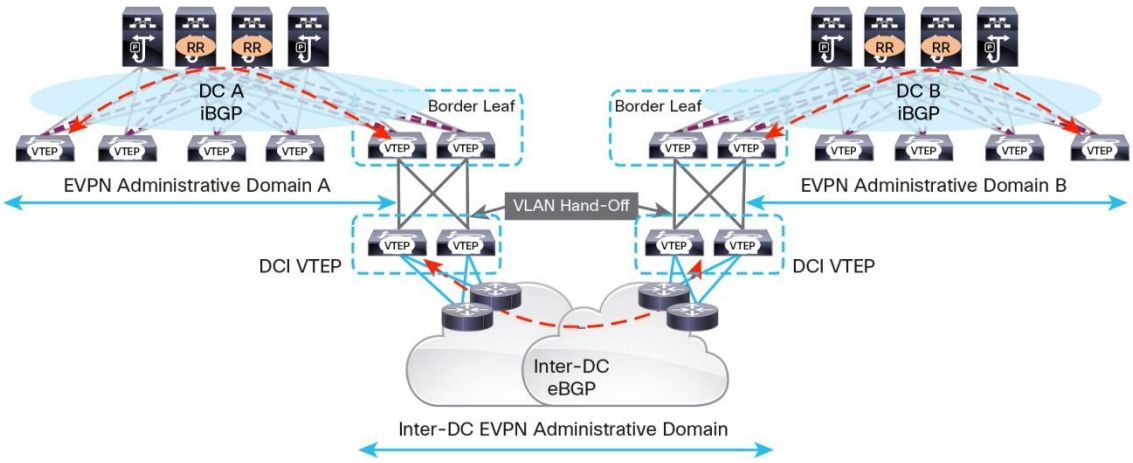


図 22 に、MP-BGP EVPN を使用した別の DCI 設計を示します。各データ センターに個別の MP-iBGP EVPN ドメインがあり、DCI VTEP 間のデータセンター間の MP-eBGP EVPN ドメインを介してそれらを結合します。VTEP が直接接続されていない場合、DCI VTEP 間の MP-eBGP セッションはマルチホップである必要があります。この設計により、各データ センターにさまざまな EVPN 運用モデルと機能モデルを柔軟に導入できます。また、各データ センターに独自のアトミック EVPN ドメインがあるため、データ センター内の VTEP ピアリングに関して、データ センター内のスケーラビリティが向上します。

図 22. 個別の MP-BGP EVPN 管理ドメインを使用した DCI



まとめ

MP-BGP EVPN は、VXLAN オーバーレイ ネットワークのパラダイムを変更します。コントロールプレーン ラーニングを導入して、フラッディングや学習に依存する代わりに、あらゆる規模のネットワークで一貫してシグナリングされる転送データベースを提供します。MP-BGP EVPN は、業界標準のドラフトと、シンプルで相互運用可能なテクノロジーを開発するために協力する複数のベンダーとサービス プロバイダーによる共同作業に基づいています。最適化されたトラフィックの配信のために、オーバーレイ ネットワークに統合されたブリッジングとルーティングを提供します。Cisco NX-OS ソフトウェアの MP-BGP EVPN 機能と Cisco Nexus 9000 シリーズ ハードウェアの VXLAN ルーティング機能を使用すると、Cisco Nexus 9000 シリーズ スイッチを使用して、スケーラブルで堅牢な高性能の VXLAN オーバーレイ ファブリック ネットワークを構築できます。

詳細情報

- IETF Draft-BGP MPLS ベース イーサネット VPN：
  https://tools.ietf.org/html/draft-ietf-l2vpn-evpn-11
- IETF Draft-EVPN を使用したネットワーク仮想化オーバーレイ ソリューション：
  https://tools.ietf.org/html/draft-ietf-bess-evpn-overlay-00
- IETF ドラフト：EVPN の統合ルーティングとブリッジング：
  https://tools.ietf.org/html/draft-ietf-bess-evpn-inter-subnet-forwarding-00
- IETF Draft-IP プレフィックス アドバタイズメント：
  https://tools.ietf.org/html/draft-rabadan-l2vpn-evpn-prefix-advertisement-02
- RFC 4271-Border Gateway Protocol 4（BGP-4）：
  https：//tools.ietf.org/html/rfc4271
- RFC 4760 - BGP-4 マルチプロトコル拡張機能：
  https://tools.ietf.org/html/rfc4760
- RFC 4364 - BGP/MPLS IP VPN：
  https://tools.ietf.org/html/rfc4364#page-15
- VXLAN の概要：Cisco Nexus 9000 シリーズ スイッチ：
  http://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-729383.html
- Cisco Nexus 9300 プラットフォーム スイッチを使用した VXLAN の設計：
  http://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-732453.html

この資料の記載内容は2021年2月現在のものです。
この資料に記載された仕様は予告なく変更する場合があります。

お問い合せ先

シスコシステムズ合同会社
〒107 - 6227 東京都港区赤坂9-7-1 ミッドタウン・タワー
http://www.cisco.com/jp

Printed in USA C07-734107-01 03/16