BGP EVPN VXLAN の概要

BGP EVPN VXLAN

BGP EVPN VXLAN は、Cisco IOS XE ソフトウェアを実行する Cisco Catalyst 9000 シリーズスイッチ用のキャンパス ネットワーク ソリューションです。このソリューションは、BGP Enabled ServiceS（BESS[1]）ワークグループによって提案された IETF 標準規格と Internet Draft の成果です。統合型オーバーレイ ネットワーク ソリューションを提供し、既存のテクノロジーの課題と短所に対処するように設計されています。

この章では、ソリューションの進化の背景について説明し、BGP EVPN VXLAN を理解するために必要な概念情報と基本的な用語を示します。このコンフィギュレーション ガイドの後半の章では、BGP EVPN VXLAN の設定、実装、機能、およびトラブルシューティングについて説明します。

BGP EVPN VXLAN の進化

従来、キャンパスネットワークでネットワーク セグメンテーションを提供する標準的な方法は VLAN でした。VLAN は、スパニングツリープロトコル（STP）などのループ防止技術を使用しているため、ネットワーク設計と復元力が制限されます。さらに、レイヤ 2 セグメントのアドレス指定に使用できる VLAN の数には制限があります（4,094 個の VLAN）。したがって、大規模で複雑なキャンパスネットワークを構築する IT 部門やクラウドプロバイダにとって、VLAN は制限要因となります。

VXLAN は、VLAN および STP の固有の制限を打破するように設計されています。これは、提案されている IETF 標準規格（RFC 7348）であり、VLAN と同じイーサネットレイヤ 2 のネットワークサービスを提供しますが、柔軟性が向上します。機能的には、VXLAN は既存のレイ

ヤ3ネットワーク上で仮想オーバーレイとして動作するMAC-in-UDPのカプセル化プロトコルです。

ただし、VXLAN自体は拡張性を制限する「フラッディングと学習」メカニズムを使用するため、ネットワークに最適なスイッチングとルーティングの機能を提供しません。「フラッディングと学習」メカニズムは、ホストの情報が到達可能なようにネットワーク全体にフラッディングされます。最適なスイッチングとルーティングの機能を提供するには、VXLANオーバーレイに次が必要です。

- ファブリックに接続されたエンドポイント間のユニキャスト通信を可能にするためにデータプレーン転送を実行する基盤のトランスポートネットワーク。
- レイヤ2とレイヤ3のホスト到達可能性情報をネットワーク全体に配布できるコントロールプレーン。

これらの追加要件を満たすために、BESSワークグループによって提出されたInternet Draft（draft-ietf-bess-evpn-overlay-12）では、レイヤ2のMAC情報とレイヤ3のIP情報を同時に伝送するMP-BGPが提案されていました。これを実現するために、MP-BGPではネットワーク層到達可能性情報（NLRI）を組み込みます。転送の決定にMACとIPの情報を一緒に使用できるため、ネットワーク内のルーティングとスイッチングが最適化されます。これにより、VXLANで使用される従来の「フラッディングと学習」メカニズムの使用が最小限に抑えられ、ファブリックの拡張性が確保されます。EVPNは、BGPがレイヤ2のMACとレイヤ3のIP情報を転送できるようにする拡張機能です。この展開をBGP EVPN VXLANファブリックと呼びます（VXLANファブリックとも呼びまる）。

BGP EVPN VXLAN を使用したオーバーレイ/アンダーレイアーキテクチャの展開の利点

BGP EVPN VXLANを使用したオーバーレイ/アンダーレイアーキテクチャの展開には次の利点があります。

- 拡張性：VXLANは1,600万のテナントネットワークに拡張可能なインフラストラクチャを可能にするレイヤ2接続を提供します。VLANの4094セグメントの制限を打破します。これは今日のマルチテナントクラウドの要件に対応するために必要です。
- 柔軟性：VXLANを使用すると、マルチテナント環境で、必要なトラフィックの分離とともにワークロードを任意の場所に配置できます。トラフィックの分離は、VXLANセグメントIDまたはVXLANネットワーク識別子（VNI）を使用したネットワークセグメンテーションによって行われます。テナントのワークロードは異なる物理デバイスに分散できますが、それぞれのレイヤ2 VNIまたはレイヤ3 VNIによって識別されます。
- モビリティ：仮想マシンはスパインスイッチテーブルを更新せずにある場所から別の場所に移動できます。これは、同じテナントVXLANネットワーク内のエンティティは、場所に関係なく同じVXLANセグメントIDを保持するためです。

BGP EVPN VXLAN の基本概念

この項では、BGP EVPN VXLAN の動作に関連するさまざまな基本概念と用語について説明します。

VXLAN オーバーレイ

オーバーレイネットワークは、物理ネットワークインフラストラクチャの最上部で動作する静的トンネルまたはダイナミックを形成することにより、既存のレイヤ2ネットワークまたはレイヤ3ネットワーク上に構築される仮想ネットワークです。既存のレイヤ2ネットワークまたはレイヤ3ネットワークは、アンダーレイを形成するものであり、この章で後述します。

データパケットがオーバーレイを介して送信される場合、元のパケットまたはフレームは外部ヘッダーを持つ送信元エッジデバイスでパッケージ化またはカプセル化され、適切な宛先エッジデバイスに向けて発信されます。中間ネットワークデバイスは、外部ヘッダーに基づいてパケットを転送しますが、元のパケットのデータを認識しません。宛先エッジデバイスではオーバーレイヘッダーを削除することによってパケットのカプセル化が解除され、内部の実際のデータに基づいて転送されます。

BGP EVPN VXLAN のコンテキストでは、データパケットをカプセル化し、トラフィックをレイヤ3ネットワーク上でトンネリングするためのオーバーレイテクノロジーとしてVXLANが使用されます。VXLAN は、MAC-in-UDP カプセル化を使用してレイヤ 2 オーバーレイネットワークを作成します。VXLAN ヘッダーが元のレイヤ 2 フレームに追加された後、UDP-IP パケット内に配置されます。VXLAN オーバーレイネットワークは、VXLAN セグメントとも呼ばれています。同じVXLANセグメント内のホストデバイスと仮想マシンのみが相互に通信できます。

VXLAN ネットワーク識別子

各 VXLAN セグメントは、VXLAN ネットワーク識別子と呼ばれる 24 ビットのセグメント ID で識別されます。これにより、最大 1,600 万の VXLAN セグメントを同じ管理ドメイン内に存在させることができます。

仮想トンネルエンドポイント

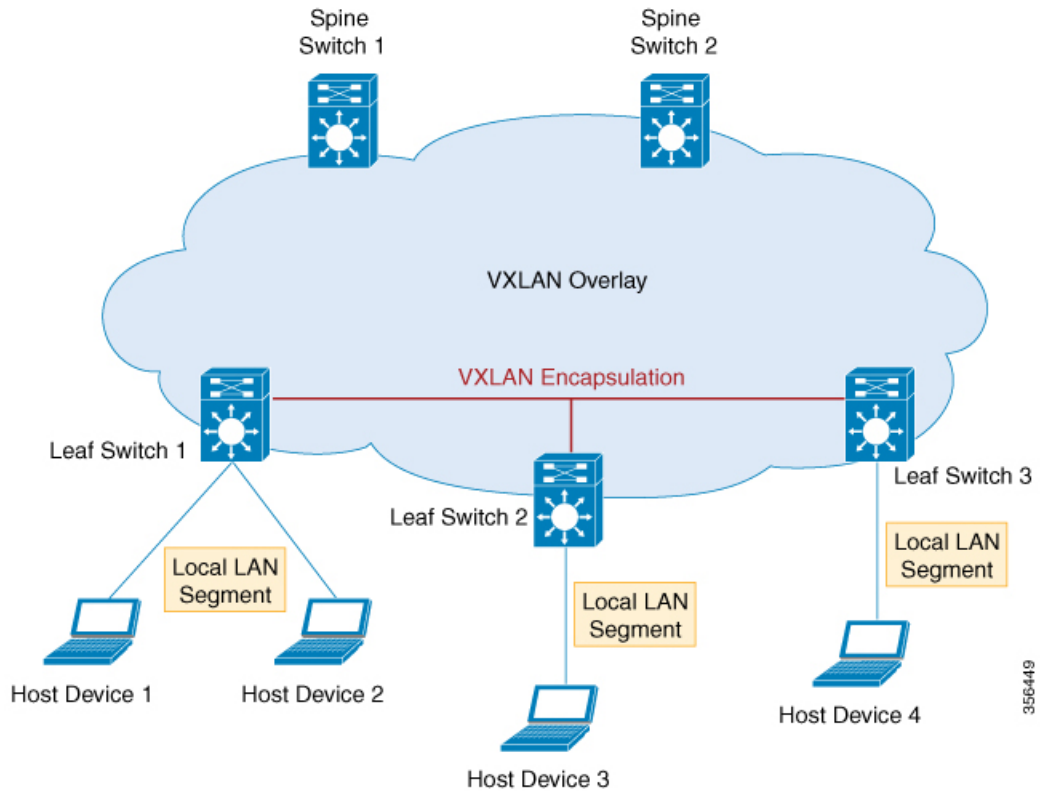
すべての VXLAN セグメントには、仮想トンネルエンドポイント（VTEP）と呼ばれるトンネルエッジデバイスがあります。これらのデバイスは VXLAN ネットワークのエッジにあり、VXLAN トンネルのインスタンスを作成し、VXLAN のカプセル化とカプセル化解除を実行します。

VTEP にはローカル LAN セグメントにスイッチインターフェイスがあり、ブリッジングを介してローカルエンドポイント通信をサポートし、IPインターフェイスでトランスポートIPネットワークと連動します。

IP インターフェイスには、トランスポート IP ネットワークの VTEP を識別する一意の IP アドレスがあります。VTEPはこのIPアドレスを使用してイーサネットフレームをカプセル化し、

カプセル化されたパケットを、IP インターフェイスを介して転送ネットワークへ送信します。また、VTEP デバイスはリモート VTEP で VXLAN セグメントを検出し、IP インターフェイスを介してリモートの MAC アドレスから VTEP へのマッピングについて学習します。

次の図に、さまざまな VTEP を接続するオーバーレイ VXLAN ネットワークの動作を示します。



オーバーレイマルチキャスト

オーバーレイマルチキャストはオーバーレイネットワークがネットワーク内にあるさまざまな VTEP 間でマルチキャストトラフィックを転送する方法です。テナントルーテッドマルチキャスト（TRM）は VXLAN オーバーレイネットワークでマルチキャストトラフィックを効率的に転送するメカニズムを提供します。TRM は、VXLAN ファブリック内の VTEP 上で接続された送信元と受信側間のマルチキャストルーティングを可能にする BGP-EVPN ベースのソリューションです。

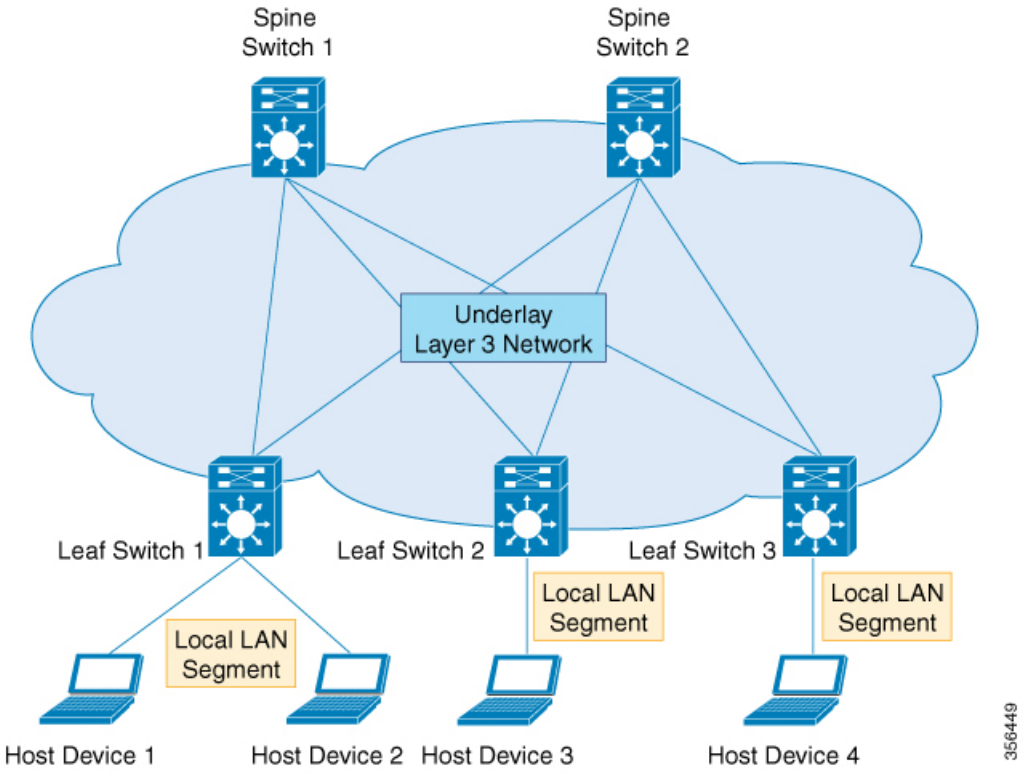
TRM を使用しない場合、マルチキャストトラフィックは、アンダーレイマルチキャストまたはの複製のいずれかの方式を使用し、BUM トラフィックの形式でアンダーレイネットワークの一部として送信されます。この場合は、異なるサブネット上に存在する送信元と受信側が相互に通信できません。TRM を使用すると、マルチキャスト通信は BUM アンダーレイトラフィックから移動されます。これにより、送信元または受信側が存在するサブネットに関係なく、オーバーレイネットワークでマルチキャスト通信が可能になります。

アンダーレイ

アンダーレイネットワークは仮想オーバーレイネットワークが確立される物理ネットワークです。オーバーレイネットワークがデータプレーンのカプセル化とともに定義されると、物理ネットワークの下でデータを移動するためのトランスポート方式が必要になります。このトランスポート方式は、通常、アンダーレイ トランスポート ネットワーク、または単なるアンダーレイです。

BGP EVPN VXLAN ではアンダーレイレイヤ 3 ネットワークが VXLAN でカプセル化されたパケットを送信元と宛先の VTEP 間で転送し、それらの間の到達可能性を実現します。VTEP 間の VXLAN オーバーレイとアンダーレイ IP ネットワークは互いに独立しています。

次の図にアンダーレイネットワークを示します。



EVPN コントロールプレーン

オーバーレイには、どのエンドホストデバイスがどのオーバーレイエッジデバイスの背後にあるかを認識するためのメカニズムが必要です。VXLAN は、特定の VXLAN ネットワーク内のブロードキャスト、不明ユニキャストおよびマルチキャスト（BUM）のトラフィックが IP コアを介してそのネットワーク内のメンバーシップを持つすべての VTEP に送信される、フラッドと学習のメカニズムでネイティブに動作します。IP マルチキャストは、ネットワーク経由でトラフィックを送信するために使用されます。受信側 VTEP はパケットのカプセル化を解除し、内部フレームに基づいてレイヤ 2 MAC の学習を実行します。内部送信元 MAC アドレス

は、送信元 VTEP に対応する外部送信元 IP アドレスと照合して学習されます。このようにして、リバーストラフィックは、以前に学習したエンドホストにユニキャストされます。

フラッディングと学習のメカニズムの欠点は、VXLAN ネットワークで拡張性が得られないことです。この問題に対処するために、コントロールプレーンを使用して MAC アドレスの学習と VTEP の検出を管理します。BGP EVPN VXLAN の展開では、イーサネット仮想プライベートネットワーク（EVPN）がコントロールプレーンとして使用されます。EVPN コントロールプレーンは、MAC アドレスと IP アドレスの両方の情報を交換できます。EVPN は、マルチプロトコル ボーダー ゲートウェイ プロトコル（MP-BGP）をルーティングプロトコルとして使用して、エンドポイントの MAC アドレス、エンドポイントの IP アドレス、およびサブネットの到達可能性情報など、VXLAN オーバーレイネットワークに関連する到達可能性情報を配布します。BGP EVPN 配布プロトコルは、場所とアイデンティティのマッピングデータベース内のトンネルエッジデバイスによるマッピング情報の構築を助長します。

ルートターゲット

ルートターゲットは、マルチテナントネットワークのルート配布を制御する EVPN ルート更新の拡張属性です。EVPN VTEP には、すべての VRF およびレイヤ 2 仮想ネットワークインスタンス（VNI）に対してインポートルートターゲット設定とエクスポート ルート ターゲット設定があります。VTEP が EVPN ルートをアドバタイズする場合、ルート更新でエクスポート ルート ターゲットが付加されます。これらのルートは、ネットワーク内の他の VTEP が受信します。受信側 VTEP は、それ自体のローカル インポート ルート ターゲット設定とそのルートで伝送されたルートターゲット値を比較します。2 つの値が一致した場合、そのルートは受け入れられ、ルーティングテーブルにプログラムされます。それ以外の場合、ルートはインポートされません。

EVPN ルートタイプ

EVPN コントロールプレーンは次のタイプの情報をアドバタイズします。

- ルートタイプ 1：これはイーサネットセグメント識別子、イーサネットタグ ID、および EVPN インスタンス情報をアドバタイズするために使用されるイーサネット自動検出（EAD）ルートタイプです。EAD ルートアドバタイズメントは EVPN インスタンスごとか、またはイーサネットセグメントごとに送信できます。
- ルートタイプ 2：エンドポイントの到達可能性情報（エンドポイントまたは VTEP の MAC アドレスと IP アドレスを含む）をアドバタイズします。
- ルートタイプ 3：マルチキャスト ルータ アドバタイズメントを実行し、特定の VNI に入力の複製を使用する機能と意図を通知します。
- ルートタイプ 4：イーサネットセグメント識別子、IP アドレス長、および発信元ルータの IP アドレスのアドバタイズに使用されるイーサネットセグメントルートです。
- ルートタイプ 5：内部 IP サブネットと外部学習ルートを VXLAN ネットワークのアドバタイズに使用される IP プレフィックスルートです。

EVPN インスタンス

EVPN インスタンス（EVI）は VTEP 上のバーチャル プライベート ネットワーク（VPN）を表します。これは、レイヤ 3 VPNの IP VRF に相当し、MAC VRF とも呼ばれます。

イーサネットセグメント

イーサネットセグメントはVTEPのアクセス側インターフェイスに関連付けられ、ホストデバイスとの接続を表します。各イーサネットセグメントにはイーサネットセグメント識別子（ESI）と呼ばれる一意の値が割り当てられます。ホストデバイスが複数のVTEPに接続されている場合、これらの接続の ESI は同じままです。

EVPN マルチホーミング

EVPN マルチホーミングを使用すると、レイヤ 2 デバイスまたはエンドホストデバイスをVXLANネットワーク内の複数のリーフスイッチに接続できます。これにより冗長性が得られ、カスタマーネットワークが1台のリーフスイッチに接続されているシングルホームトポロジでのネットワーク最適化が可能になります。リーフスイッチとの接続で得られる冗長性によって、ネットワーク障害が発生した場合にトラフィックが中断されることはありません。マルチホームトポロジは、シングルホームトポロジよりも復元力があり、安全で効率的です。EVPN マルチホーミングは、シングルアクティブおよびオールアクティブな冗長モードで動作します。

拡張 VLAN とサブネット

EVPN VXLAN は、既存のネットワーキング インフラストラクチャ上で実行することでレイヤ 2 ネットワークを拡張する手段を提供します。EVPN VXLAN オーバーレイを使用すると、レイヤ2セグメントとブロードキャストドメインをレイヤ3コアネットワーク上のサイトまたはキャンパスビルディング全体に拡張できます。EVPN VXLAN によるレイヤ 2 の拡張は、エンドユーザーの IP アドレス管理を簡素化し、大規模なキャンパスネットワークでのシームレスなモビリティを実現します。

スパイン リーフ アーキテクチャ

スパイン リーフ アーキテクチャは、1つのレイヤがリーフスイッチで構成され、もう1つのレイヤが 1 つ以上のスパインスイッチを持つ 2 レイヤネットワークトポロジです。この設計では、さまざまなスパインスイッチに複数のパスを実装することですべてのリーフスイッチを接続します。

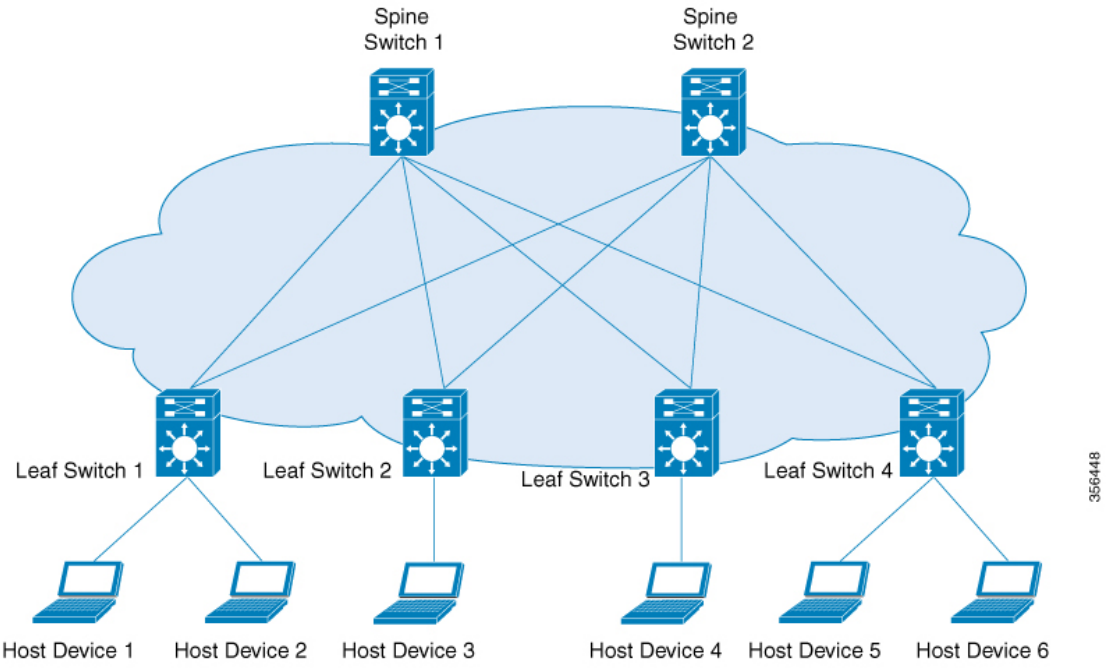
スパイン スイッチ

スパインスイッチはすべてのリーフスイッチ間の接続ノードです。リーフスイッチ間でトラフィックを転送するため、エンドポイントアドレスは認識されません。リーフスイッチを接続する複数のパスを実装することで、スパインスイッチはネットワークの冗長性を実現します。

リーフ スイッチ

リーフスイッチはホストまたはアクセスデバイスに接続されているノードです。リーフスイッチはネットワークのエッジにあるため、エッジまたはネットワーク仮想化エッジ（NVE）とも呼ばれます。あるリーフスイッチ上のホストデバイスが別のリーフスイッチ上のホストデバイスと通信しようとすると、リーフスイッチ間のトラフィックはスパインスイッチを介して送信されます。リーフスイッチは VXLAN ネットワークで VTEP として機能し、カプセル化とカプセル化解除を実行します。

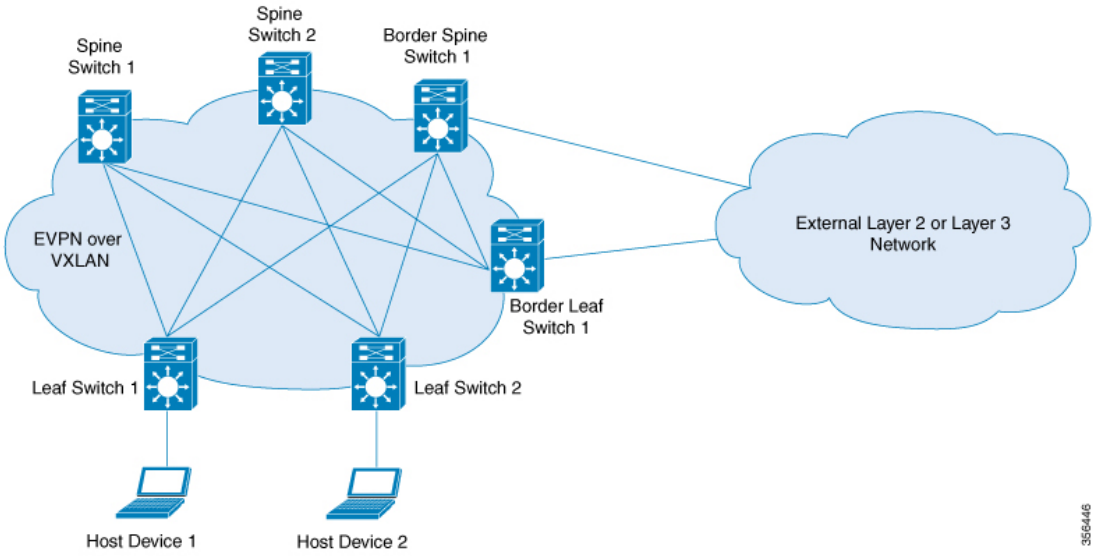
次の図に、4 台のリーフスイッチが 2 台のスパインスイッチを介して接続されている一般的なスパインリーフトポロジを示します。



ボーダースパインスイッチとボーダーリーフスイッチ

VXLAN ファブリックと他のレイヤ 2 ネットワークおよびレイヤ 3 ネットワークとの外部接続は、ボーダーノードと呼ばれるノードを介して助長されます。ボーダー機能がスパインスイッチを介して確立される場合は、ボーダースパインスイッチと呼ばれます。リーフスイッチを介して確立される場合は、ボーダーリーフスイッチと呼ばれます。

次の図に、1 台のボーダーリーフスイッチと 1 台のボーダースパインスイッチでファブリックを外部ネットワークに接続するスパインリーフトポロジを示します。



Integrated Routing and Bridging（IRB）

EVPN VXLAN は VXLAN ネットワーク内の VTEP がレイヤ 2（ブリッジド）トラフィックとレイヤ 3（ルーテッド）トラフィックの両方を転送できるようにする Integrated Routing and Bridging（IRB）機能をサポートしています。VTEP がレイヤ 2 トラフィックを転送するときはブリッジングを実行しています。同様に、VTEP がレイヤ 3 トラフィックを転送するときはルーティングを実行しています。異なるサブネット間のトラフィックは、VXLAN ゲートウェイを介して転送されます。IRB は次の 2 つの方法で実装されます。

- 非対称 IRB
- 対称 IRB

IRB の詳細については、EVPN VXLAN Integrated Routing and Bridging についての項を参照してください。

VXLAN ゲートウェイ

VXLAN ゲートウェイは、VXLAN セグメント間、または VXLAN 環境から非 VXLAN 環境にトラフィックを転送するネットワーク内のエンティティです。VXLAN ネットワークのリーフスイッチは、レイヤ 2 とレイヤ 3 の両方のVXLAN ゲートウェイとして機能できます。

レイヤ 2 VXLAN ゲートウェイは、同じ VLAN 内でトラフィックを転送します。レイヤ 2 VXLAN ゲートウェイでは、VNI セグメントを VLAN にマッピングすることで、VXLAN から VLAN へブリッジングできます。

レイヤ 3 VXLAN ゲートウェイは、トラフィックを別の VLAN に転送します。レイヤ 3 VXLAN ゲートウェイでは、VXLAN から VXLAN へのルーティングと VXLAN から VLAN へのルーティングの両方が可能です。VXLAN から VXLAN へのルーティングは、2 つの VNI 間のレイ

ヤ 3 接続を実現します。VXLAN から VLAN へのルーティングは、VNI と VLAN 間を接続します。

レイヤ 2 仮想ネットワークインスタンス

VXLAN オーバーレイネットワークを作成すると、複数のレイヤ 3 ネットワークによって分離されたさまざまなリーフノードに接続されたホストデバイスが、1 つのレイヤ 2 ネットワーク（VXLAN セグメント）に接続されているかのように連携できます。この論理レイヤ 2 セグメントはレイヤ 2 VNI と呼ばれます。同じサブネット内の 2 つの VLAN 間でレイヤ 2 VNI を通過するトラフィックは、ブリッジドトラフィックと呼ばれます。

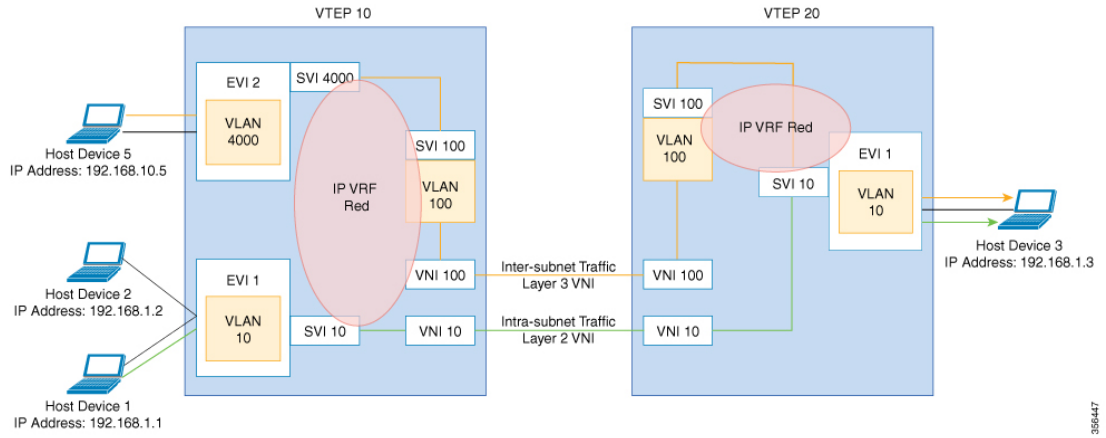
VTEP でローカルに定義された VLAN はレイヤ 2 VNI にマッピングできます。ホストデバイスがレイヤ 2 VNI に接続できるようにするには、接続された VLAN をレイヤ 2 VNI にマッピングし、レイヤ 2 VNI を VTEP 上のネットワーク仮想化エッジ（NVE）の論理インターフェイスに関連付けます。

レイヤ 3 仮想ネットワークインスタンス

レイヤ 2 VNI に接続されたエンドポイントが異なる IP サブネットに属するエンドポイントと通信する必要がある場合、それらのエンドポイントはデフォルトゲートウェイにトラフィックを送信します。異なるレイヤ 2 VNI に属するエンドポイント間の通信は、レイヤ 3 ルーティング機能によってのみ可能です。EVPN VXLAN 展開では、ローカル VLAN とグローバルレイヤ 2 VNI を組み合わせて定義されたさまざまなレイヤ 2 セグメントを VRF に関連付けることで通信できます。

レイヤ 3 VNI は、VTEP 上のすべての VRF のレイヤ 3 セグメンテーションを助長します。これは、各 VRF インスタンスをネットワーク内の一意のレイヤ 3 VNI にマッピングし、VTEP のさまざまなレイヤ 2 VNI を同じVRFに関連付けることによって実行されます。これにより、特定の VRF インスタンス内のレイヤ 3 VNI 全体で VXLAN 間の通信が可能になります。論理レイヤ 3 の分離を可能にするための VRF の使用を、マルチテナントと呼びます。異なるサブネットの 2 つの VLAN 間でレイヤ 3 VNI を通過するトラフィックをルーテッドトラフィックと呼びます。

次の図に、レイヤ 2 とレイヤ 3 の VNI を介した同じサブネットおよび異なるサブネットのホストデバイス間のトラフィックの移動を示します。



モビリティ

BGP EVPN コントロールプレーンのエンドポイントのアイデンティティはその MAC アドレスと IP アドレスから導出され、BGP EVPN は VXLAN オーバーレイ内でエンドポイントモビリティをサポートするメカニズムを提供します。

RFC 7432 は VXLAN ファブリック内のエンドポイントモビリティの範囲を定義します。

MAC モビリティと重複する MAC の検出

あるポートから別のポートにエンドポイント（またはホスト）が移動するときに MAC が移動します。新しいポートは、同じ VTEP か、または同じ VLAN 内の別の VTEP にある場合があります。BGP EVPN コントロールプレーンは、MAC ルートをアドバタイズすることでこのような移動を解決します（EVPN ルートタイプ 2）。エンドポイントの MAC アドレスが新しいポート上で学習されると、そのエンドポイントの新しい VTEP がホストのローカル VTEPであることを（BGP EVPN コントロールプレーンで）アドバタイズします。他のすべての VTEP は新しい MAC ルートを受信します。

ホストが複数回移動すると、対応する VTEP が同じ数の MAC ルートをアドバタイズすることがあります。また、新しい MAC ルートがアドバタイズされてから、古いルートが他の VTEP のルートテーブルから削除されるまでの間に遅延が発生する場合があり、短時間は2つの場所で MAC ルートが同じになります。ここで、MAC モビリティシーケンス番号は、最新の MAC ルートを決定するのに役立ちます。

ホスト MAC アドレスが初めて学習された場合は MAC モビリティシーケンス番号が 0 に設定されます。値 0 は、MAC アドレスにこれまでモビリティイベントがなく、ホストがまだ元の場所にあることを示します。MAC モビリティイベントが検出されると、新しいルートタイプ 2（MAC または IP アドバタイズメント）が、エンドポイントの移動先の新しい VTEP（その新しい場所）によって BGP EVPN コントロールプレーンに追加されます。ホストが移動するたびに、新しい場所を検出した VTEP はシーケンス番号を 1 ずつ増やし、BGP EVPN コントロールプレーンのそのホストの MAC ルートをアドバタイズします。古い場所（VTEP）で MAC ルートを受信すると、古い VTEP は古いルートを取り消します。

同じ MAC アドレスが 2 つの異なるポートで同時に学習される場合があります。EVPN コントロールプレーンはこの状態を検出し、重複する MAC があることをユーザーに警告します。MAC が重複している状態は手動による介入によって、またはいずれかのポートで MAC アドレスが期限切れになったときに自動的にクリアされます。

IP モビリティと 重複 IP の検出

BGP EVPN は MAC モビリティをサポートするのと同様の方法で IP モビリティをサポートします。主な違いは同じポートで学習したか、別のポートで学習したかに関係なく、IP アドレスが異なる MAC アドレスで学習されたときに IP の移動が検出されることです。2 つの異なる MAC アドレスで同じ IP アドレスが同時に学習されると、重複する IP アドレスが検出されます。これが発生すると、ユーザーに警告が表示されます。