vPC ボーダーゲートウェイを使用した VXLAN EVPN マルチサイトでの次世代 DCI

目次

このドキュメントの内容

このドキュメントでは、VXLAN EVPN マルチサイトアーキテクチャの一部である vPC ボーダーゲートウェイ（vPC BGW）の機能と使用例について説明します。使用例の主な目的の 1 つは、クラシック イーサネット ネットワークにデータセンター相互接続（DCI）として VXLAN EVPN マルチサイトを導入することです。vPC BGW の導入は、Cisco NX-OS 9.2(1) 以降でサポートされています。

このドキュメントでは、まず EVPN マルチサイト vPC BGW の具体的な使用例について概要を説明します。次に、レガシーテクノロジーで構築されたデータセンターの相互接続に vPC BGW を使用した EVPN マルチサイト（DCI の使用例）について詳しく解説します。このソリューションを採用することによるアーキテクチャ上の主なメリットを取り上げるとともに、レガシーネットワークを最終的に最新の VXLAN BGP EVPN ファブリックに移行する方法も説明します。

vPC BGW を支えているテクノロジーや導入時に設計面で考慮すべき事項については、このドキュメントの付録を参照してください。

はじめに

VXLAN EVPN マルチサイトは、Cisco NX-OS 7.0(3)I7(1) を搭載した Cisco Nexus® 9000 シリーズ クラウドスケール プラットフォーム（Cisco Nexus 9000 シリーズ EX、FX、FX2 プラットフォーム）で初めて導入された相互接続アーキテクチャを提供します。

一般に VXLAN EVPN マルチサイト導入は、「サイト」と通称される 2 つ以上のデータセンターネットワークで構成され、VXLAN BGP EVPN レイヤ 2 およびレイヤ 3 オーバーレイを介して相互接続されます。

これまで主流だった内部 BGP（iBGP）を使用する手法に対して、EVPN マルチサイトアーキテクチャでは VXLAN BGP EVPN ネットワークに外部 BGP（eBGP）を導入し、オーバーレイネットワークにかつての階層構造を取り入れます。eBGP ネクストホップ動作の導入に続いて、ボーダーゲートウェイ（BGW）に自律システム（AS）が導入され、ネットワーク制御ポイントがオーバーレイネットワークに返されました。このアプローチでは、階層を効率的に使用して複数のオーバーレイネットワークを区分し、相互接続します。また、組織は、単一のデータセンター内外のネットワーク拡張を制御するための制御ポイントも備えています。

ユースケース

VXLAN EVPN マルチサイトアーキテクチャは、VXLAN BGP EVPN ベースのオーバーレイネットワーク向けの設計です。複数の異なる VXLAN BGP EVPN ファブリックまたはオーバーレイドメインの相互接続が可能になり、ファブリックのスケーリング、コンパートメント化、DCI への新しいアプローチが可能になります。VXLAN EVPN マルチサイトの活用例は多く、コンパートメント化、階層型のスケールアウト手法、DCI、レガシーネットワークの統合などに対応できます。このドキュメントでは後に挙げた 2 つの用例に焦点を当てます。

地理的に分散したデータセンターにまたがるネットワークの拡張

EVPN マルチサイトアーキテクチャは、DCI を念頭に構築しました（図 1）。この包括的なアーキテクチャならデータセンター単位で単一または複数のファブリックを配置して、それらをリモートデータセンターの単一または複数のファブリックと相互接続できます。ファブリック内およびファブリック間の VXLAN BGP EVPN を介したシームレスかつ制御されたレイヤ 2 およびレイヤ 3 拡張により、VXLAN BGP EVPN 自体の機能が強化されています。ネットワーク制御、VTEP マスキング、および BUM トラフィック適用に関連する新機能は、EVPN マルチサイトアーキテクチャを最も効率的な DCI テクノロジーにするための機能の一部にすぎません。

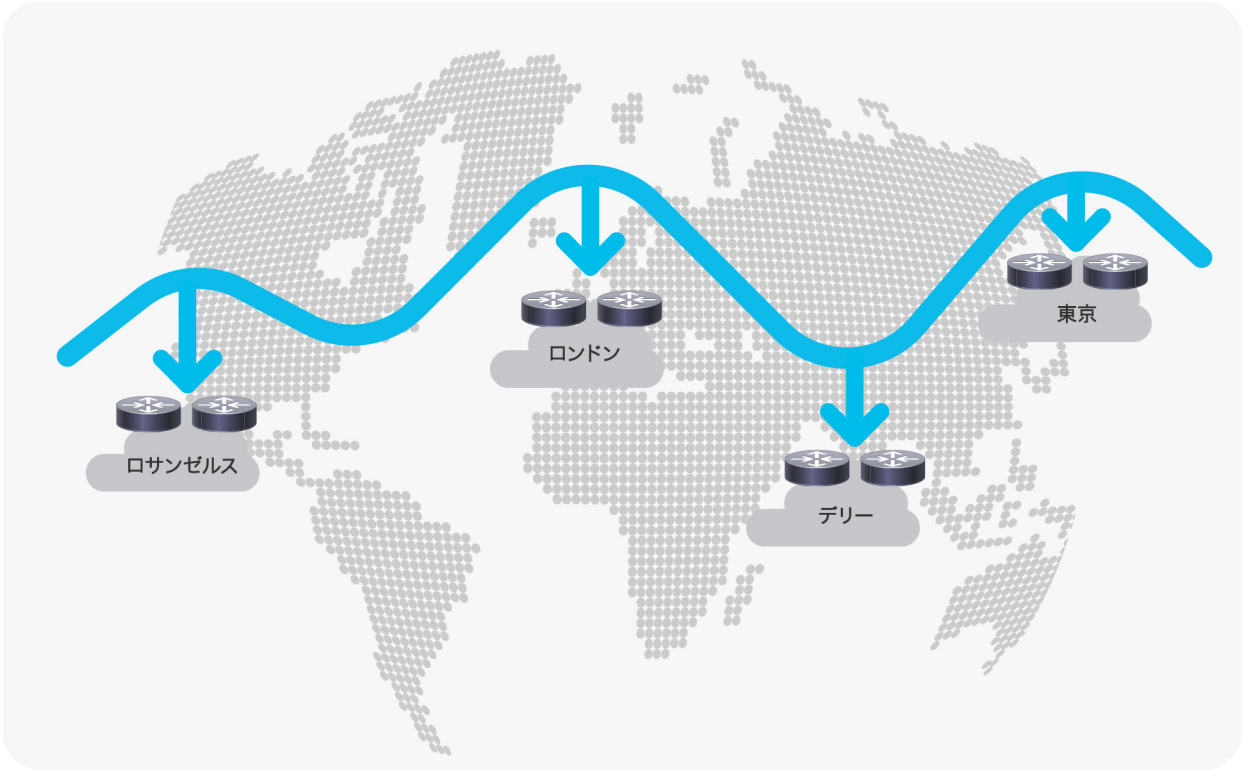

図 1.
地理的に分散したデータセンターを相互接続する VXLAN EVPN マルチサイト

レガシーネットワークとの統合

EVPN マルチサイトソリューションの設計の狙いは、VXLAN BGP EVPN データセンターファブリックを相互接続することだけではありません。共存と移行のシナリオの促進、すなわち旧式の（レガシー）テクノロジーで構築されたデータセンターネットワークを相互接続することも視野に入れて設計されています。スパニングツリープロトコル（STP）、仮想ポートチャネル（vPC）、または Cisco FabricPath を備えたネットワークが複数ある場合、EVPN マルチサイトなら最先端の相互接続機能を使用できます（図 2）。

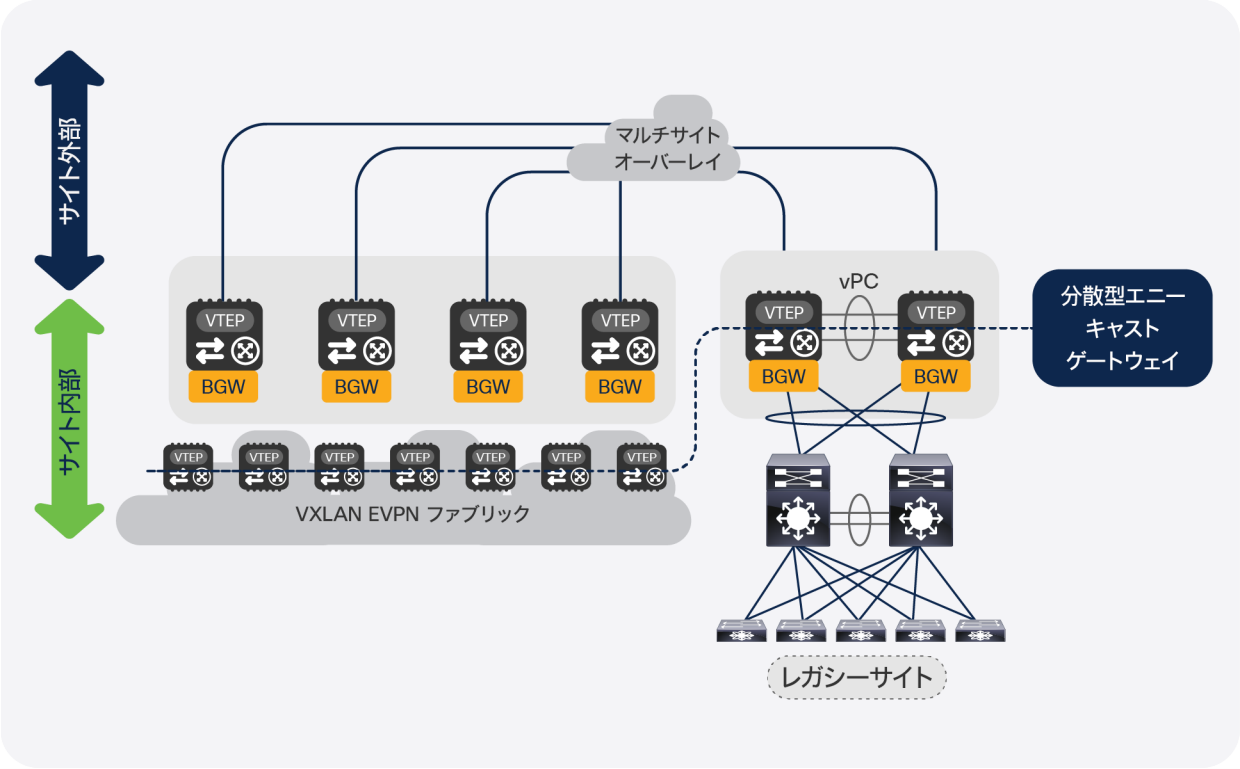

図 2.
レガシーネットワークとの統合を実現する VXLAN EVPN マルチサイト

VXLAN EVPN マルチサイトアーキテクチャは vPC、OTV、VPLS、EoMPLS などの DCI テクノロジーに代わる最新のアーキテクチャです。特にレガシーテクノロジー（STP、vPC、Cisco FabricPath など）のみで構築されているデータセンターネットワークを相互接続するには最適です。

このドキュメントではデータセンター間の相互接続を実現する最新の手法として、VXLAN EVPN マルチサイトによるレガシーネットワークとの相互接続および VXLAN BGP EVPN ファブリックとの共存に焦点を当てて説明します。

vPC BGW を使用した VXLAN EVPN マルチサイト環境の使用例

先ほど触れたとおり、vPC BGW 導入には複数の活用例がありますが、以前は EVPN マルチサイト導入にレガシーネットワークを統合する際の主要なポイントとして考えられていました。vPC BGW は仮想ポートチャネル（vPC）による冗長レイヤ 2 接続と、分散型エニーキャストゲートウェイを使用したファーストホップ ゲートウェイのホスティングを実現します。EVPN マルチサイト機能、レイヤ 2 接続、ファーストホップ ゲートウェイを組み合わせることで、vPC BGW を既存データセンターネットワークのアグリゲーションレイヤの拡張（図 3）として使用したり、VXLAN BGP EVPN ネットワーク内でエンドポイントをローカル接続（図 4）したりできます。

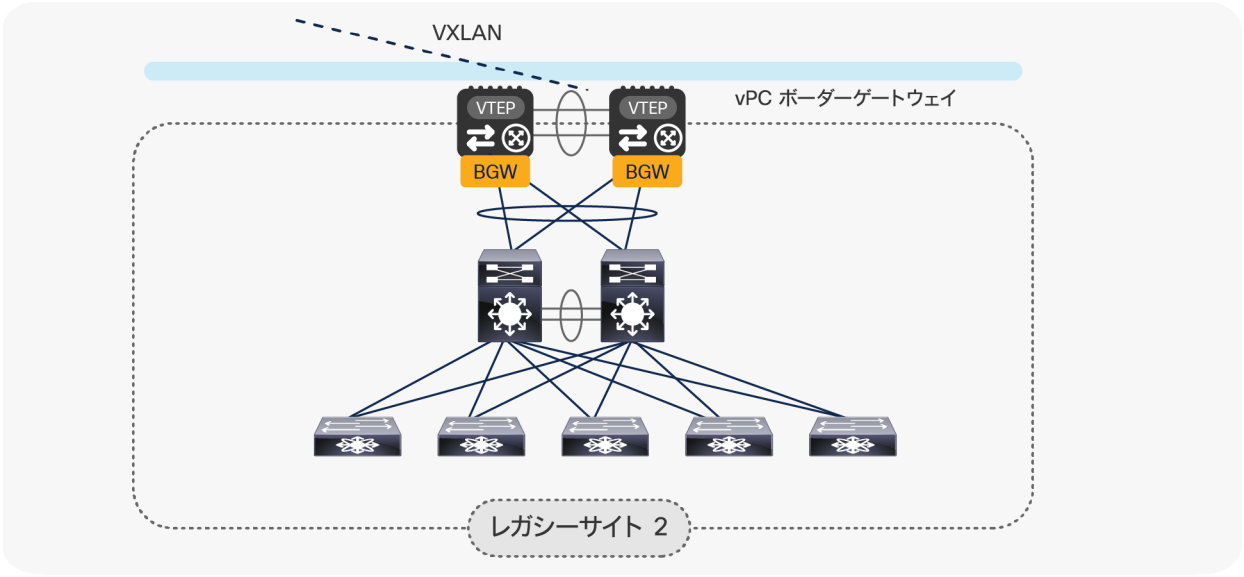

図 3.
vPC BGW の接続

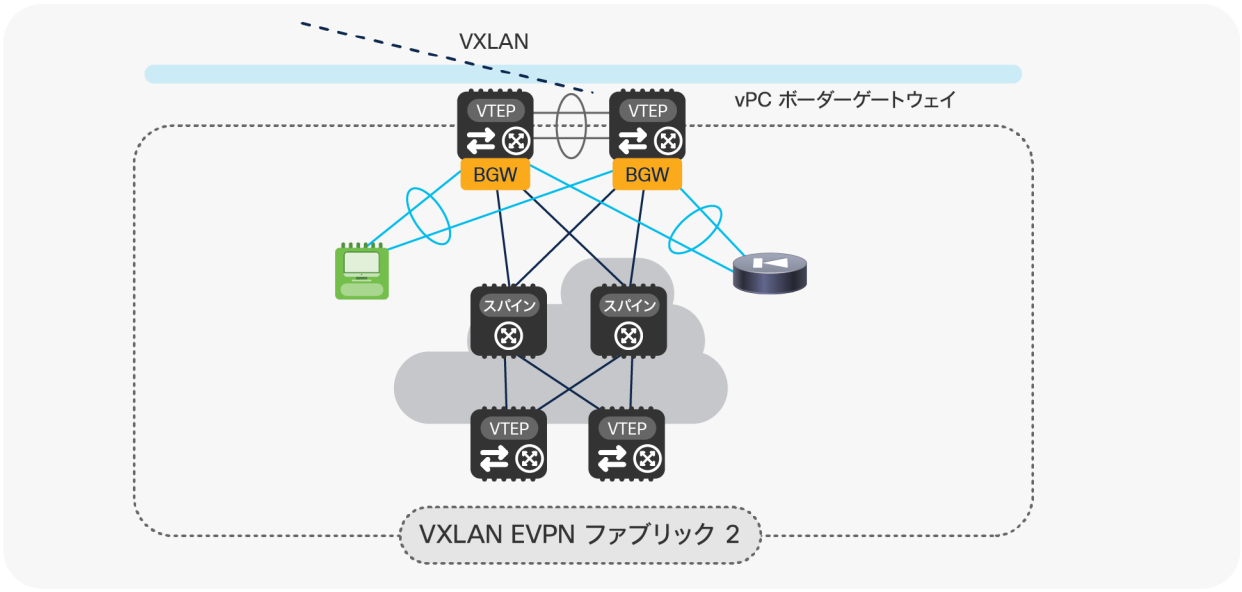

図 4.
vPC BGW とエンドポイント

vPC BGW を既存のレガシーネットワークに接続する必要がある環境では、たいていの場合リモートネットワークと相互接続（DCI の使用例）するか、VXLAN EVPN テクノロジーで構築された最新のファブリックに移行するかのどちらかを選択することになります。どちらを選んでも VXLAN BGP EVPN ファブリックとレガシーネットワークの共存は考慮されています。これらの例では、EVPN マルチサイトによってファーストホップ ゲートウェイ機能が提供され、多様なネットワークタイプ間の完全なレイヤ 2 およびレイヤ 3 接続が可能になります。EVPN マルチサイトを導入し、分散型エニーキャストゲートウェイ（DAG）をファーストホップ ゲートウェイに使用することで、ホストのモビリティが実現します（図 5）。

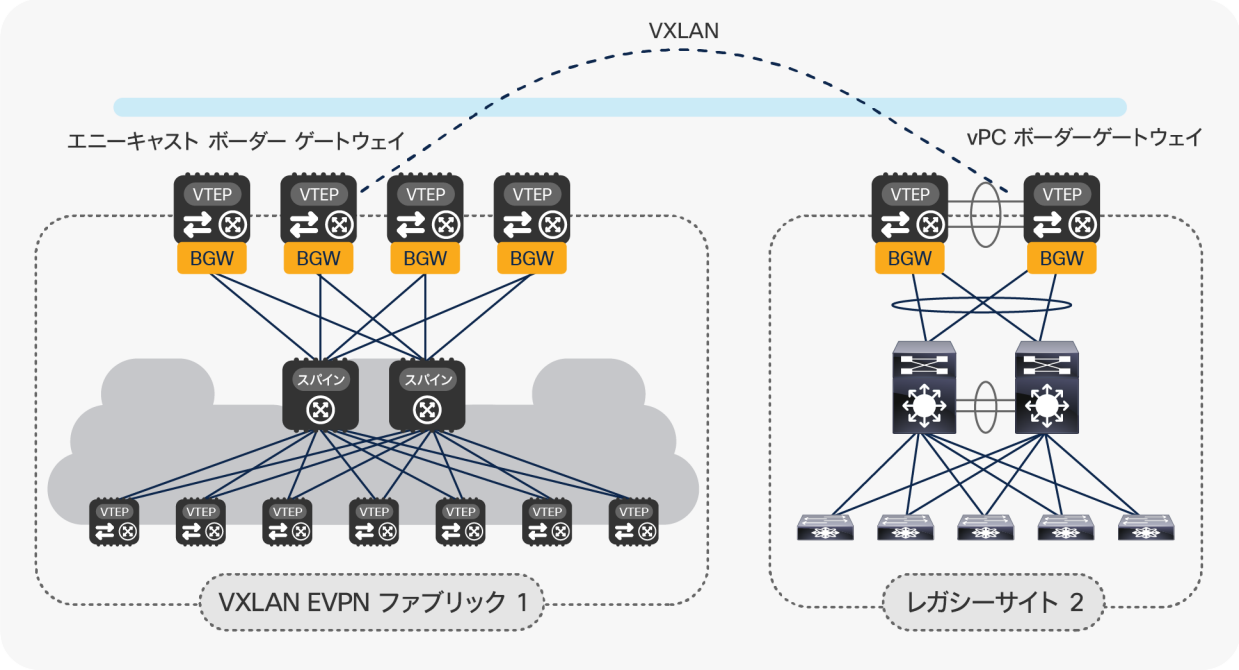

図 5.
EVPN マルチサイトによるレガシーサイトと VXLAN BGP EVPN サイトとの統合と共存

複数のレガシー データセンター サイトを接続する vPC BGW ノード

ドキュメントの後半では、vPC BGW ノードのペアをレガシーネットワークと相互にローカル接続する際に、設計面で考慮すべき事項と、設定に関するベストプラクティスについて詳しく取り上げます。この手法をそのまま流用すれば、VXLAN BGP EVPN をサイト内の接続に利用しないデータセンターサイトとも、同じ導入モデルで相互接続できます（図 6）。

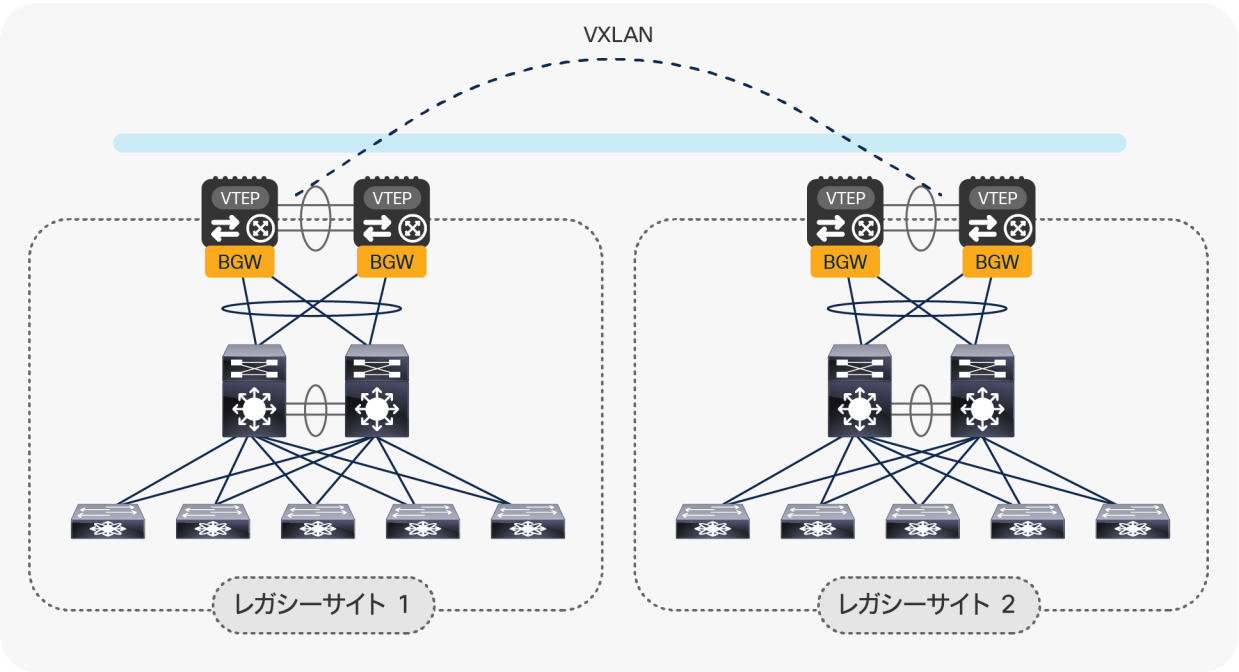

図 6.
vPC BGW ノードによる複数のレガシー データセンター サイトの接続

この使用例では、VXLAN EVPN と vPC BGW は、従来のデータセンター相互接続（DCI）ソリューション（OTV や VPLS など）の代替手段という位置付けになります。「vPC BGW が DCI の使用例にもたらすアーキテクチャ上のメリット」のセクションでは、この手法で上記使用例を実現する場合の主なメリットについて説明します。

「vPC BGW によるレガシーデータセンターの VXLAN EVPN ファブリックへの移行」セクションでは、この具体的な使用例を詳しく取り上げ、上記の導入モデルが、各サイトで使用されているレガシーテクノロジーを刷新して最新の VXLAN EVPN ファブリックに置き換えることを目的とした移行手順の、最初のステップでもあることを明らかにします。

小規模ファブリック導入向けの vPC BGW

vPC BGW ノードの導入が必要になるもう 1 つの使用例は、専用のエニーキャスト BGW ノードを導入できない（または費用対効果が低い）小規模ファブリック間でマルチサイト接続を確立（図 7）するものです。

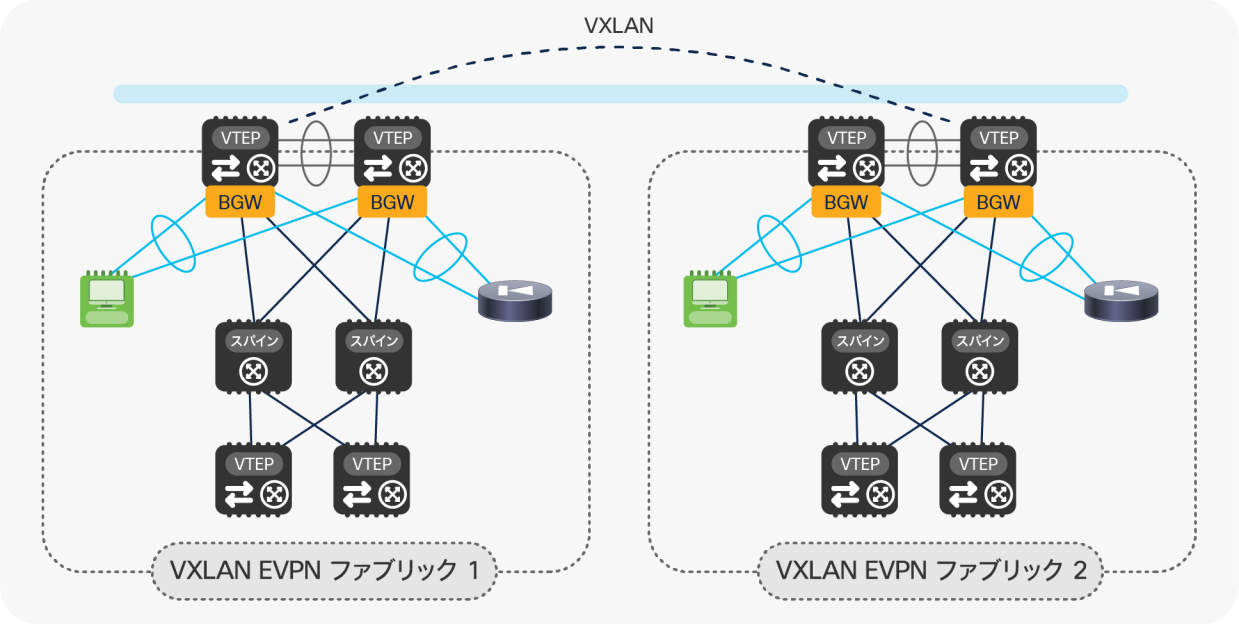

図 7.
小規模ファブリック環境への vPC BGW の導入

この使用例では、BGW の機能を提供するリーフノードのペアが、ローカル接続されたエンドポイントとサービスノードをサポートする役割も果たし、コンピューティングノードとしてもサービスリーフノードとしても効果的に機能します。

この導入モデルは全面的にサポートされてはいますが、これらの機能すべてを同一リーフノードのペアに実装すると、ネットワーク設計や設定、トラフィックフローのデバッグが複雑になります。したがってマルチサイト環境でお勧めする手法は、BGW 機能専用のリーフノードを用意するか、さらに任意でそのリーフノードデバイスから外部ネットワークドメインにレイヤ 3 接続も提供する（ボーダーリーフノードとして機能させる）かです。

注： vPC BGW 機能はスパインノードに実装することを推奨します。この導入モデルは、BGW をエニーキャスト ゲートウェイ モードで導入する場合にサポートされます。詳細については、次のドキュメントの中で説明しています。https://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-739942.html

vPC BGW が DCI の使用例にもたらすアーキテクチャ上のメリット

vPC BGW ノードを導入してレガシー データセンター サイト間に DCI を提供する手法にはアーキテクチャ面でのメリットがあります。以下のセクションではそのメリットについて取り上げます。

コントロールプレーンとデータプレーン

VXLAN EVPN マルチサイトは、BGP EVPN コントロールプレーンと VXLAN データプレーンを使用してサイト間のレイヤ 2 接続とレイヤ 3 接続を拡張します。複数のレガシーデータセンターを接続する場合、vPC BGW ノードを利用することで、コントロールプレーンとデータプレーンの両面から階層的にレガシーデータセンターを切り離すことができます。

コントロールプレーンに関しては、マルチサイト選択的アドバタイズメント機能により、レイヤ 2 およびレイヤ 3 のコントロールプレーン アドバタイズメントの範囲を厳密に制御できます。実際には、レイヤ 2 セグメントの MAC、IP ホスト、IP サブネットのプレフィックス情報と、BGW ノードでローカルに定義されている VRF のみがリモートサイトにアドバタイズされます。この仕組みによりマルチサイトソリューションの全体的な規模が改善され、サイト間のコントロールプレーンの動作が最小限に抑えられます。

データプレーン転送の観点から見ると、vPC BGW ノードはレガシーデータセンター間の接続を VXLAN トンネルによって拡張しています。ローカルなレガシーネットワーク内のエンドポイントから発信されてリモートサイトのエンドポイントに送信されるトラフィックは、標準ベースの VXLAN パケットに動的にカプセル化され、VXLAN トンネルを介して転送ネットワーク全体に配信されます。

レイヤ 2 およびレイヤ 3 統合の拡張

VXLAN BGP EVPN 対称統合ルーティングおよびブリッジング（IRB）機能とマルチテナント機能を利用すれば、同じテクノロジーでレイヤ 2 およびレイヤ 3 の拡張が可能です。これによりレガシーネットワークの統合が大幅に簡素化されます。これは通常レイヤ 3 またはレイヤ 2 どちらか一方の接続モデルを提供する、従来の DCI テクノロジーの場合とは異なります。

たとえば VRF-lite、MPLS L3VPN、LISP などのテクノロジーが提供する接続はレイヤ 3 のみであり、VPLS や Cisco OTV などのテクノロジーが提供する接続はレイヤ 2 拡張のみです。以上をまとめると、vPC BGW を導入すれば、レイヤ 2 およびレイヤ 3 の統合拡張、ワークロードのモビリティ、複数のレガシー データセンター ネットワーク間のマルチテナントを容易に実現できるということです。

障害の抑制

レイヤ 2 拡張を複数のデータセンターサイト間で実装する場合は、レガシーデータセンター間のレイヤ 2 ブロードキャスト、不明なユニキャスト、マルチキャスト（BUM）トラフィックのフラッディングを常に厳密に制御する必要があります。これは、特定のサイトのレガシーネットワークに影響を及ぼす問題（ブロードキャストストームなど）が、他のサイトに伝播するのを防止するという点できわめて重要です。

EVPN マルチサイトストーム制御と呼ばれる特別な機能は、他のレガシーサイトに BUM トラフィックが伝播する量を制御できるように設計されています。図 8 に示すように、レイヤ 2 ブロードキャスト、不明なユニキャスト、マルチキャストは vPC BGW レベルで個別に微調整され、これらのトラフィックタイプが集約されてリモートサイトに伝播するのを制限できます。

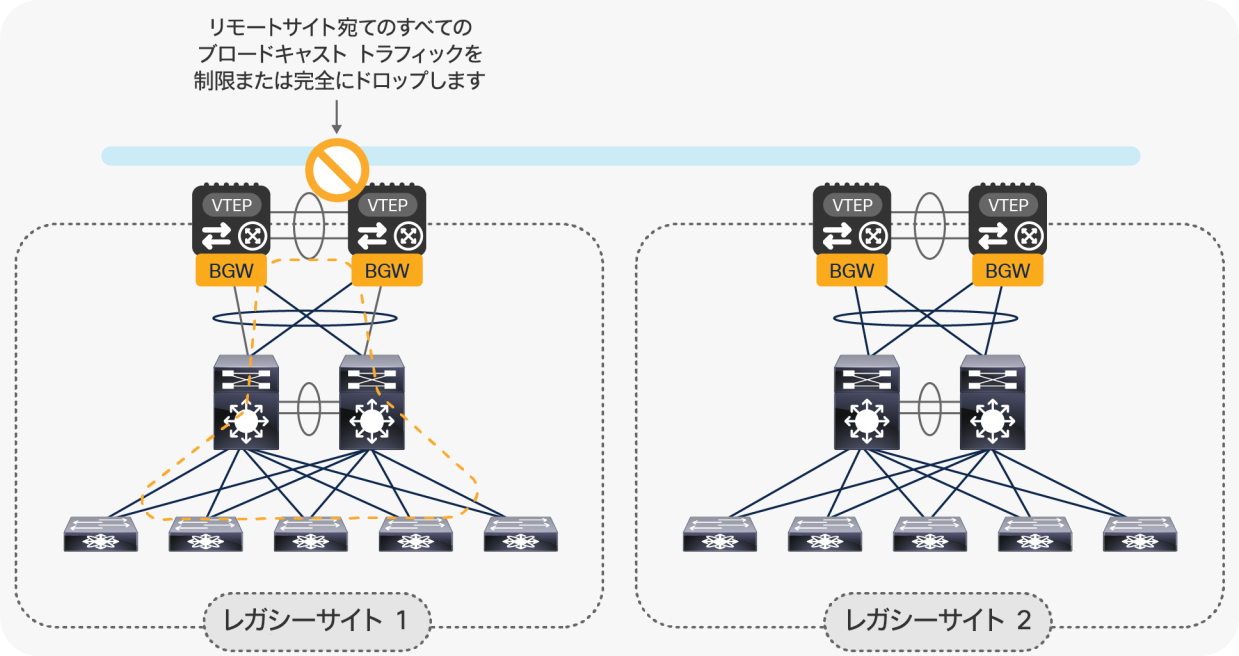

図 8.
サイト間にまたがるレイヤ 2 ブロードキャストストームの伝播の防止

転送に依存しない性質

サイトをまたぐ vPC BGW ペア間で確立されている VXLAN トンネルを使用すれば、データセンターを相互接続する転送ネットワークの機能（および設定）を簡素化できます。このトンネルがリモート vPC BGW ノード間の IP 接続を提供するものである限り、実際にあらゆるタイプの転送インフラストラクチャ上にトンネルを構築できます。また、元のトラフィックに対して VXLAN カプセル化を挿入できるようにするには、転送ネットワークで処理するトラフィックの最大伝送ユニット（MTU）を増やす必要があります。通常は、レガシーネットワークに接続されたエンドポイントを発信元とするトラフィックの MTU に大きく依存する場合でも、50 バイト増やせば十分です。

注： vPC BGW ノードはフラグメンテーションとリアセンブルを行いません。

vPC BGW ノードがレガシーデータセンター間の BUM トラフィックのレプリケーションを、入力レプリケーション（IR）モードでどのように処理しているかにも注意してください。この仕組みのおかげで、アンダーレイ転送インフラストラクチャでマルチキャスト機能をサポートする必要はありません。

マルチホーム機能

前述のように vPC BGW ノードの各ペアは、レイヤ 2 vPC 接続をローカルのレガシーネットワークに使用します。図 9 にあるとおり、リモートレガシーサイトのエンドポイントは、リモート vPC BGW ペアの vPC VIP を介して、到達可能な EVPN のネクストホップアドレスとして学習されます。

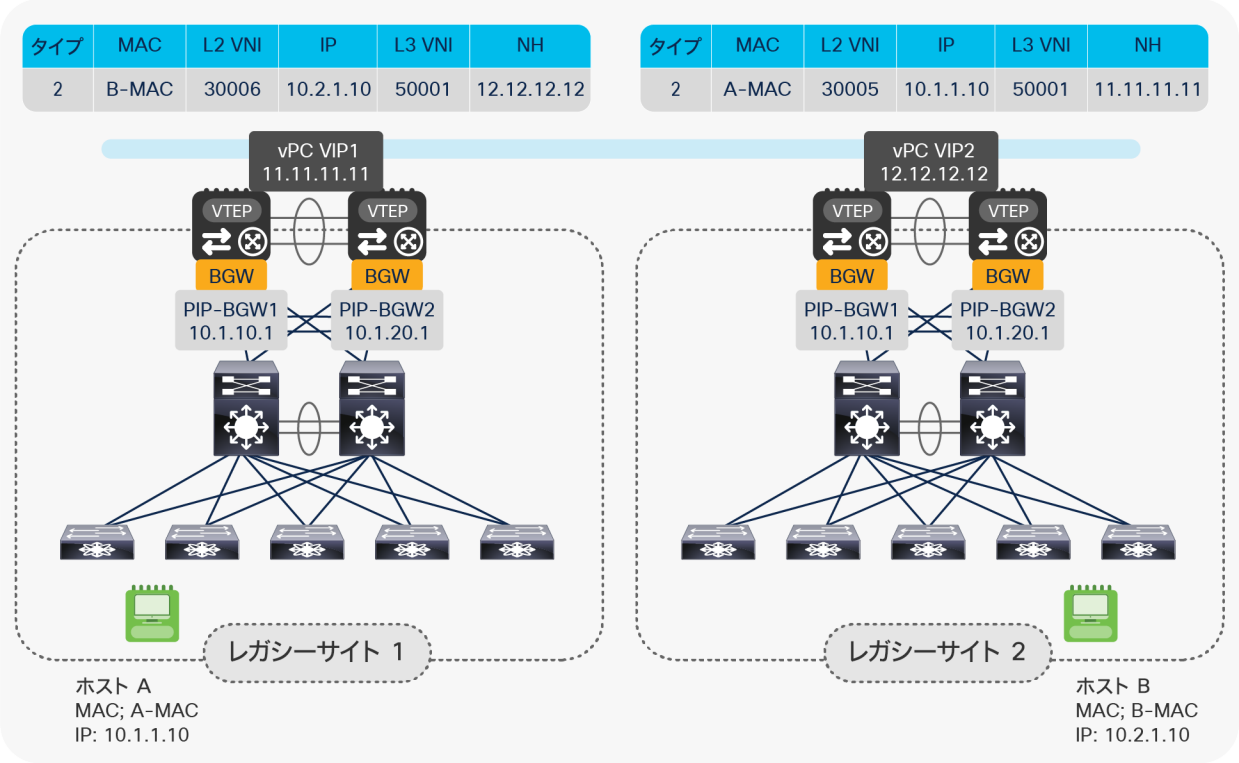

図 9.
リモート vPC VIP を介した EVPN ネクストホップとしてのリモートエンドポイントの学習

注： 前述した DCI の具体的な使用例では、マルチサイト VIP は各サイトの vPC BGW ノードで設定しますが、VXLAN トラフィックの送信や受信には使用されません。これは、ローカル VTEP リーフノードを導入しておらず、レガシーネットワークに存在するすべてのローカルエンドポイントが、直接 BGW ノードに接続されているものとして検出されるためです。

1 つの vPC BGW ノードで障害が発生しても、残りの vPC BGW が同じ vPC VIP アドレスであるため、すべてのデータトラフィックの転送をそのまま引き継げます。これによりオーバーレイ コントロール プレーンの働きによってネットワークをコンバージェンスさせる必要がなくなり、ネットワーク全体の復元力とリカバリ時間が大幅に向上します。

マルチパス ロード シェアリング

前述のように、レガシーネットワークに接続されたエンドポイント間のすべてのサイト間通信（レイヤ 2 またはレイヤ 3 のどちらか）には、サイトをまたぐ 2 組の vPC BGW ノード間で確立された VXLAN トンネルが使用されます。

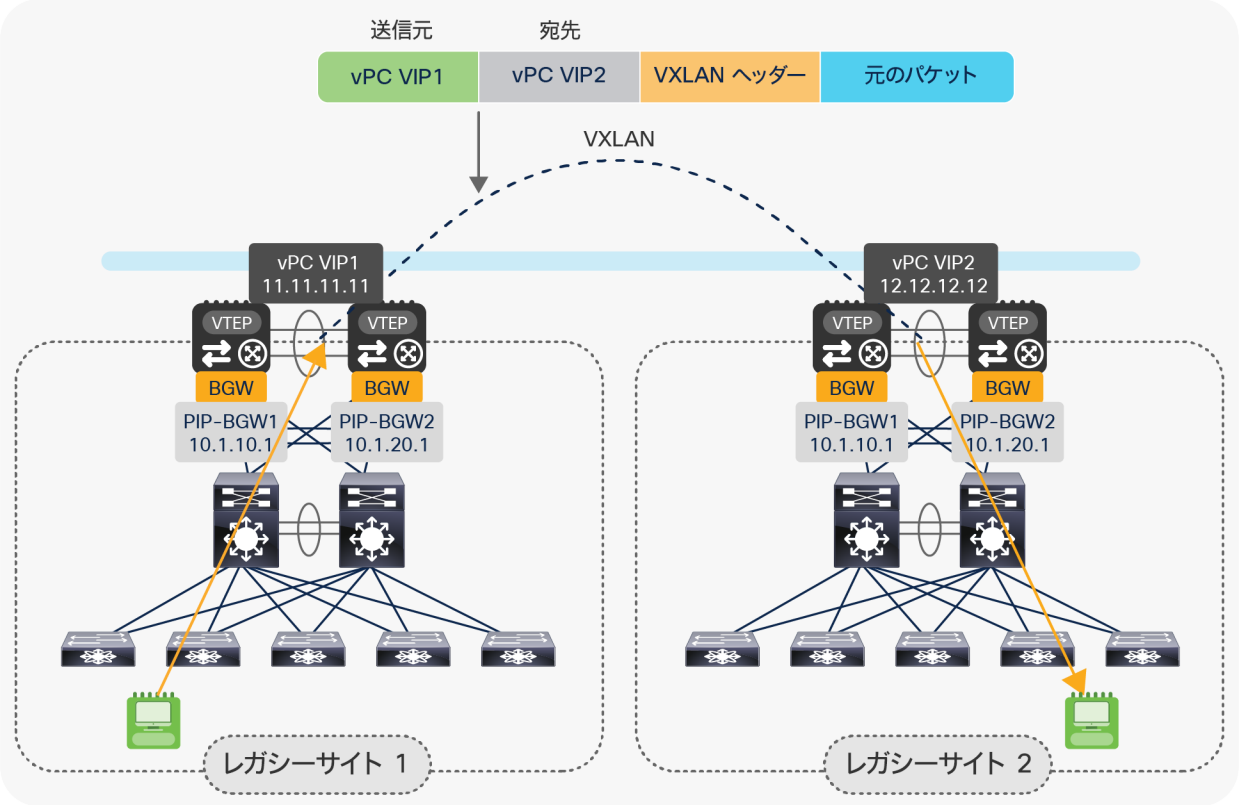

図 10.
サイト間通信での vPC VIP アドレスの使用

図 10 に示すように、VXLAN トラフィックの外部 IP ヘッダーで使用される送信元および宛先 IP アドレスは、各 vPC BGW ペアで定義された vPC VIP アドレスを表します。このことから等コストパスが存在する場合には、サイト間ネットワークでのトラフィック分散に不安があるかもしれません。つまり、同じ IP ヘッダー情報を使用する場合、すべてのサイト間通信の物理パスが同じになると考えるかもしれませんが、それは誤解です。

実際には VXLAN トラフィックは User Datagram Protocol（UDP）でカプセル化され、トラフィックフローごとに UDP の送信元ポート情報を変更することで、パケットに「エントロピー」を組み込むことができます。この仕組みは元のパケットの内部ヘッダーのハッシュを計算し、その値を VXLAN カプセル化トラフィックの送信元 UDP ポートとして使用することで実現されています。このことは、さまざまな（元のヘッダー情報の違いが性質に反映される）フローが、異なる送信元 UDP ポート情報でカプセル化されることを意味します。これにより、アンダーレイ転送 IP ネットワークで使用されるさまざまな ECMP パス間で、トラフィックのロードバランシングが可能になります。

ループの防止と STP の分離

EVPN マルチサイトでは、エンドポイントと IP サブネットの到達可能性に関する情報を交換するため、サイトをまたいで導入している BGW ノードペア間のコントロールプレーンとして、eBGP EVPN を使用する必要があります。MAC アドレスと IP プレフィックスのアドバタイズメントは、vPC VIP をネクストホップアドレスに持つローカル vPC BGW ノードから発信されます。BGP 内蔵の as-path 属性を使用すると、特定のレガシーサイトを発信元とするプレフィックスを同じサイトにインポートして取り込むことができなくなります。この仕組みによってコントロールプレーンのレベルでネイティブにループを防止します。

データプレーンの観点から見ると、vPC 指定フォワーダ選択とスプリットホライズンのルールによって、BUM トラフィックがサイト間でループするのを防止しています。

スパニングツリープロトコル（STP）との統合に関しては、vPC BGW ノードは、レガシー ネットワーク インフラストラクチャと接続されるクラシカルイーサネット（CE）ポートだけが STP に参加します。BPDU パケットはマルチサイト DCI オーバーレイに転送されないため、各レガシーデータセンターは別々の STP ドメインとなります。

vPC BGW で STP ポートのステータスが頻繁に変更されるのを防ぐため、STP ルートをレガシーネットワークからマルチサイト vPC BGW のペアに移動することをお勧めします。この手法は 2 つのレガシーサイト間にレイヤ 2 バックドア接続が確立されている場合にも効果的です。というのは、STP はレイヤ 2 バックドアリンクをブロックしてエンドツーエンドのループを防げるためです。次の図 11 のように動作させるには、vPC BGW ノードに次のような特定の設定を適用しなければならない点に注意してください。

- 両データセンターサイトに導入した vPC BGW で、STP の優先順位が同じになるように設定する必要があります。
- vPC ドメイン番号も同じものにします。そうすることで vPC BGW の両ペアに同じブリッジ ID を割り当てられるようになります。

上記の設定を行ったことで、レガシーネットワーク間にレイヤ 2 バックドアが作成されると vPC BGW の両ペアが同一の STP ルートデバイスとして認識され、STP によってリンクがブロックされます。

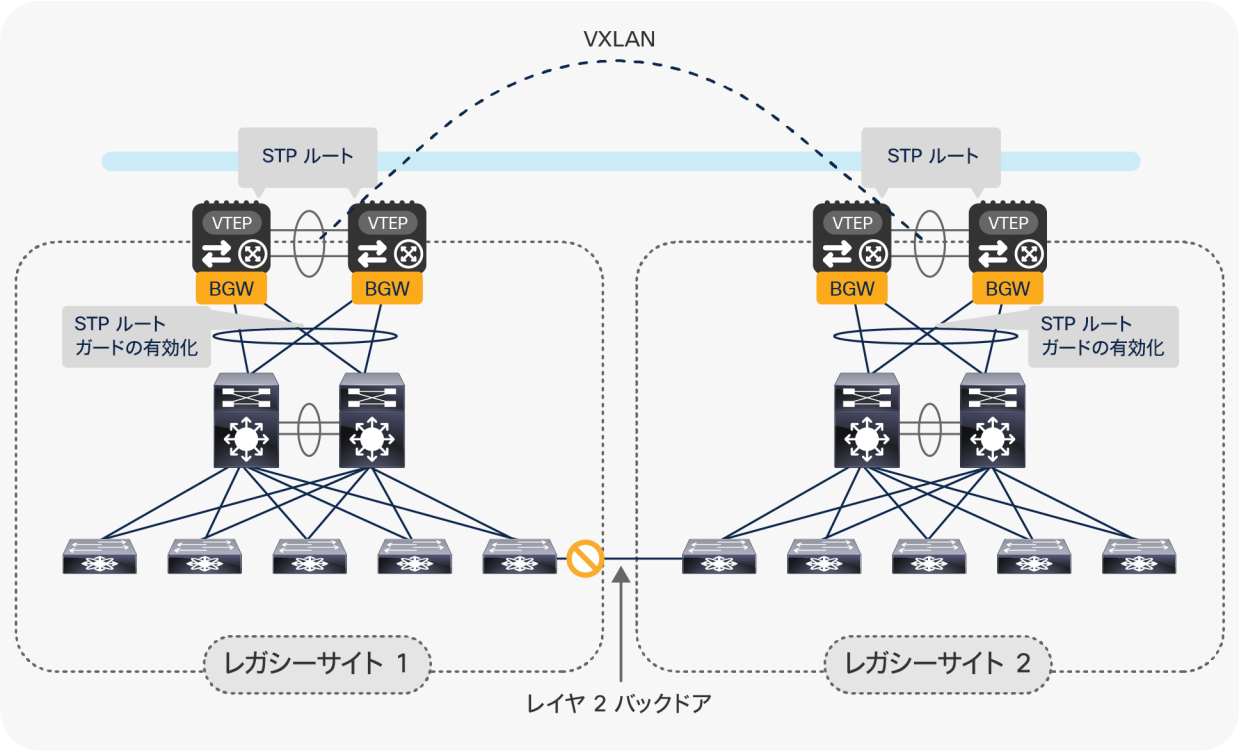

図 11.
エンドツーエンドでのループを防止する STP の使用例

上記の図に示したように、BGW ノードの各ペアとレガシーネットワーク間の論理 vPC 接続に STP ルートガードを設定することをお勧めします。これはレガシー インフラストラクチャ内の誤ったスイッチ設定により、スイッチが STP ルートの役割を主張するのを防止するためです。

注： レガシーネットワーク内のアクセスレイヤスイッチのエッジインターフェイスには、常に Cisco® のベストプラクティスに基づく設定（STP BPDU ガードの有効化など）を適用しておく必要があります。そうすれば STP ルートを vPC BGW デバイスに導入していないシナリオでも、レイヤ 2 バックドアが作成されたことを受けてリンクが無効化されるようになります。

複数サイトのサポート

VXLAN EVPN マルチサイトアーキテクチャは拡張性に考慮して設計されています。Cisco NX-OS リリース 7.0(3)I7(1) では最大 10 サイトをサポートします。Cisco NX-OS リリース 9.2(1) 以降では、レガシーデータセンター（vPC BGW ノードのペアを利用）と、通常複数のエニーキャスト BGW ノードを導入している VXLAN EVPN ファブリックデータセンターの両方が、このサポート数に含まれます。

注： サポート対象のサイトの最大数は、今後の Cisco NX-OS リリースでさらに増える予定です。詳しくは、Cisco.com で提供している拡張性に関する最新情報に目を通してください。

vPC BGW によるレガシーデータセンターの VXLAN EVPN ファブリックへの移行

前のセクションでは、VXLAN EVPN マルチサイトテクノロジーが DCI の使用例を念頭に、どのように設計されているかを説明しました。このセクションでは、vPC BGW ノードを導入しつつレガシーデータセンターを新世代の VXLAN EVPN ファブリックへと移行する手順を詳しく説明します。移行手順の各ステップを詳しく解説し、具体的な設定内容についても紹介していきます。

注： 設定のサンプルはすべて Cisco NX-OS 9.2(1) に基づいています。

前提として、レガシーサイトは従来の DCI 接続（レイヤ 2 の場合は OTV、vPC、VPLS、レイヤ 3 の場合は VRF-Lite または MPLS VPN）を利用して、あらかじめ相互接続（レイヤ 2 およびレイヤ 3）されているものとします。移行目標の 1 つは、この DCI 接続を最新の VXLAN EVPN マルチサイトオプションに置き換えることです。

ステップ 1：レイヤ 2 のダブルサイド vPC を使用して各レガシーサイトに vPC BGW ペアを挿入する

最初の前提として、従来のアグリゲーション設計およびアクセスレイヤ設計によるレガシーネットワークを導入し、アグリゲーションスイッチにデフォルトゲートウェイを導入しているものとします。

注： エンドポイントのファーストホップ ゲートウェイがファイアウォールノードに導入（通常はアグリゲーションレイヤ スイッチに接続）されているシナリオでも、考慮すべき事項はこのステップで説明するものと同じです。

図 12 に示すように、レイヤ 2 ダブルサイド vPC を使用して vPC BGW ノードのペアをアグリゲーションスイッチのペアに接続します。

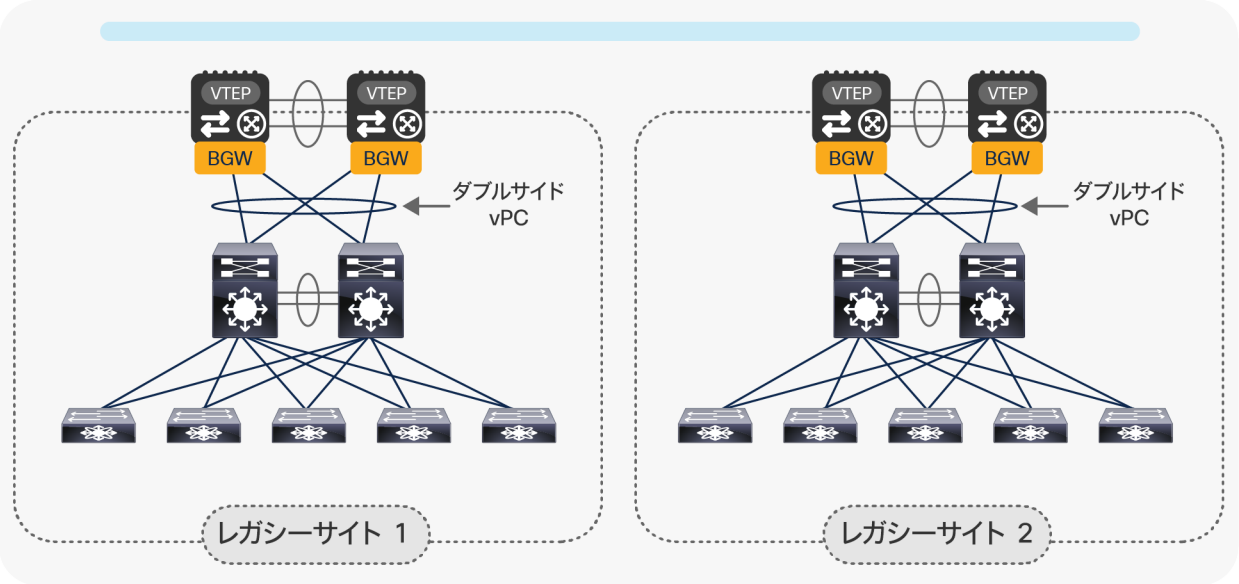

図 12.
レイヤ 2 ダブルサイド vPC を使用した vPC BGW ノードのレガシーネットワークへの接続

ダブルサイド vPC 接続のメリットは、BGW ノードとレガシーネットワークの間に単一のレイヤ 2 論理接続が存在するため、STP でパスをブロックすることなく、使用可能なすべてのリンクによって 2 つのネットワーク間のトラフィックをアクティブに転送できることにあります。

アグリゲーションスイッチが vPC または MLAG をサポートしていない場合は、図 13 に示すように各アグリゲーションスイッチと vPC BGW ノードのペアを元にローカルポートチャネルを作成できます。

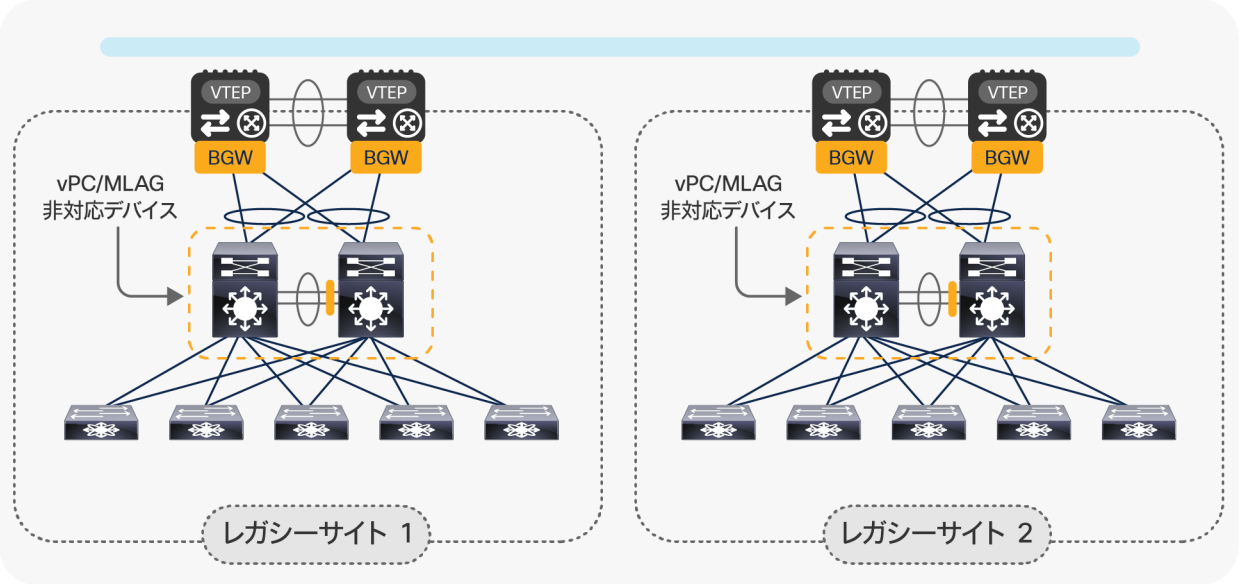

図 13.
ローカルポートチャネルを使用した vPC BGW ノードに接続するアグリゲーションスイッチ

その後、STP はアグリゲーションスイッチと BGW の間に作成されたレイヤ 2 ループを解除する必要があります。これにより、2 つあるローカルポートチャネルの一方が STP ブロッキングステートになります。

残りのページで、最初の推奨オプションであるダブルサイド vPC 接続を利用する手法について説明していきます。BGW ノードは同じ vPC ドメインの一部として、以下のベストプラクティスの設定に沿って設定する必要があります。

feature vpc
 	
vpc domain 1
  peer-switch
  peer-keepalive destination 172.19.217.122 \
 source 172.19.217.123
  delay-restore 150
  peer-gateway
  auto-recovery reload-delay 360
  ipv6 nd synchronize
  ip arp synchronize
 	
interface port-channel10
  vpc peer-link
	vPC ドメインを定義し、delay-restore と reload-delay タイマーを適切な値に調整して、vPC ピアのリロードイベント後にコンバージェンスを最適化します。

vlan 3600
 	
interface Vlan3600
  description VPC-Peer-Link SVI
  no shutdown
  mtu 9216
  no ip redirects
  ip address 10.1.10.49/30
  no ipv6 redirects
  ip ospf network point-to-point
  ip router ospf UNDERLAY area 0.0.0.0
  ip pim sparse-mode
 	
system nve infra-vlans 3600
 	
router bgp 65501
  neighbor 10.1.10.50
  remote-as 65501
  address-family ipv4 unicast
	vPC ピアデバイス間のアンダーレイドメインで iBGP セッションを確立します。非常に具体的な障害シナリオにおけるトラフィックの回復に対応するために、既存の IGP ピアリング（OSPF、IS-IS など）に追加する形でこれを設定する必要があります。

ステップ 2：vPC BGW DCI アンダーレイネットワークを設定する

データセンターサイト（サイト外部ネットワーク）を相互接続するネットワークとは、異なるサイトに導入している vPC BGW ペアの間に、アンダーレイによる到達可能性を提供する転送ネットワークです。vPC BGW ノードは、図 14 に示すように、サイト間ネットワークのファーストホップレイヤ 3 デバイス部とルーティングピアリングを確立する必要があります。

注： ダークファイバ接続や高密度波長分割多重（DWDM）回路をサイト間で使用できるような特定のケースでは、レイヤ 3 のポイントツーポイント インターフェイスを介して、直接 vPC BGW ノードの 2 つのペアを接続できます。

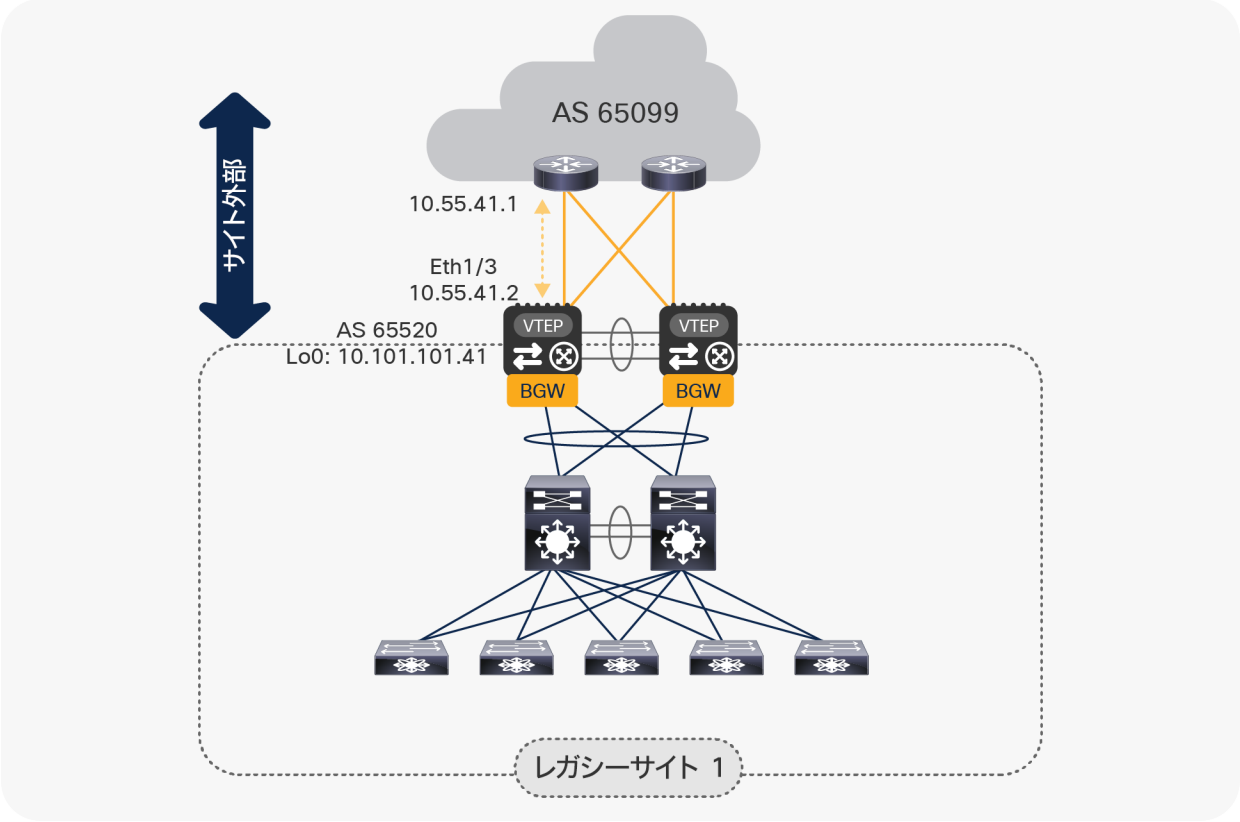

図 14.
vPC BGW ノードのサイト外部アンダーレイネットワークへの接続

サイト外部アンダーレイネットワーク内のファーストホップルータとのルーティングピアリングには、任意のルーティングプロトコル（OSPF、IS-IS、EIGRP など）を使用できます。ただし、次のセクションで説明するように、サイト間のオーバーレイピアリングに EBGP が必要な場合は、アンダーレイプロトコルとして eBGP ピアリングを使用するのが一般的です。

注： このような特有の考慮事項が適用されるのはサイト外部ネットワークに対してであり、サイト内部ネットワークに対してではありません。後者については、アンダーレイプロトコルに IGP（OSPF、IS-IS、または EIGRP）を、オーバーレイプロトコルに iBGP を導入することをベストプラクティスとして常に推奨します。

次の例は、vPC BGW と直接接続されたサイト外部ネットワークのルータとの間で、EBGP ピアリングを確立するために必要な設定を示したものです。図 14 のネットワーク図を参考にしてください。

`interface Ethernet1/3`
`  no switchport`
`  mtu 9216`
`  ip address 10.55.41.2/30 tag 54321`
`  evpn multisite dci-tracking`
	vPC BGW を外部レイヤ 3 コアに接続するサイト外部アンダーレイ インターフェイスを定義します。
	
	インターフェイスの MTU 設定を固有の要件に適した値に調整します（最小値は 1500 バイト + VXLAN カプセル化用の 50 バイト）。
	
	ポイントツーポイント IP アドレッシングが、サイト外部アンダーレイルーティングに使用されます（ここでは、/30 を指定したポイントツーポイント IP アドレッシングを使用）。IP アドレスには関連タグを設定します。このタグがあることで、サイト間アンダーレイ ルーティング プロトコルに再配布する IP アドレスを簡単に選択できるようになります。

	注：特定の vPC BGW ノードが外部ネットワークから分断されているシナリオを検出するには、外部レイヤ 3 コアに接続しているインターフェイスに EVPN マルチサイト インターフェイス トラッキング（**evpn multisite dci-tracking**）を設定する必要があります。

router bgp 65520
  router-id 10.101.101.41
  log-neighbor-changes
  address-family ipv4 unicast
  redistribute direct route-map RMAP-REDIST-DIRECT
  maximum-paths 4
  neighbor 10.55.41.1
  remote-as 65099
  update-source Ethernet1/3
  address-family ipv4 unicast
	サイト固有の自律システムを使用して BGP ルーティングインスタンスを定義します。BGP ルータ ID は loopback0 CP の IP アドレスと一致している必要があります。
	IPv4 ユニキャスト グローバル アドレス ファミリ（VRF デフォルト）をアクティブにして、必須のループバックプレフィックスと（必要に応じて）物理インターフェイスの IP アドレスを BGP に再配布します。
	BGP マルチパスを有効にします（**maximum-paths** コマンド）。
	eBGP ネイバー設定は、具体的には、このシングルホップ eBGP ピアリングの送信元インターフェイスを選択すること（update-source コマンド）によって実行されます。この設定を行うことで、物理リンクに障害が発生した場合にすぐにネイバー関係が切断されます。

route-map RMAP-REDIST-DIRECT permit 10 match tag 54321	ローカルで定義されたインターフェイス（direct）から BGP への再配布は、ルートマップ分類によって実現され、一致するタグを持つ IP アドレスのみが再配布されます。

BGW ノードをサイト外部ネットワークと接続しているすべてのレイヤ 3 インターフェイスに、同じ設定を適用する必要があります。また、サイト外部ネットワークに属するすべてのレイヤ 3 デバイスでも、同様に適切なアンダーレイ設定がされてることが前提となります。

ステップ 3：vPC BGW DCI オーバーレイネットワークを設定する

EVPN マルチサイトでは、別々のサイトに導入している BGW ノード間のオーバーレイ コントロールプレーンとして MP-eBGP EVPN を使用する必要があります。このオーバーレイ コントロールプレーンは、VRF（IP サブネットやホストルート）とレイヤ 2 VNI（MAC アドレス）の到達可能性に関する情報を交換するために使用されます。図 15 の例は、別々のサイトの 2 つの BGW ノード間で EVPN ピアリングが確立された状態を示しています。ベストプラクティスでは、vPC BGW ノードの 2 つのペア間において、EVPN 隣接関係のフルメッシュを構築することが推奨されています。

注： 相互接続されているサイトの数によっては、サイト外部ネットワークに「ルートサーバ」デバイスのペアを導入してルートリフレクタの役割を担わせ、BGW ノード間でフルメッシュの隣接関係を構築しないようにすることでメリットを得られる場合もあります。ルートサーバノードの導入の詳細については、『VXLAN EVPN マルチサイト設計および導入ホワイトペーパー』[1] を参照してください。

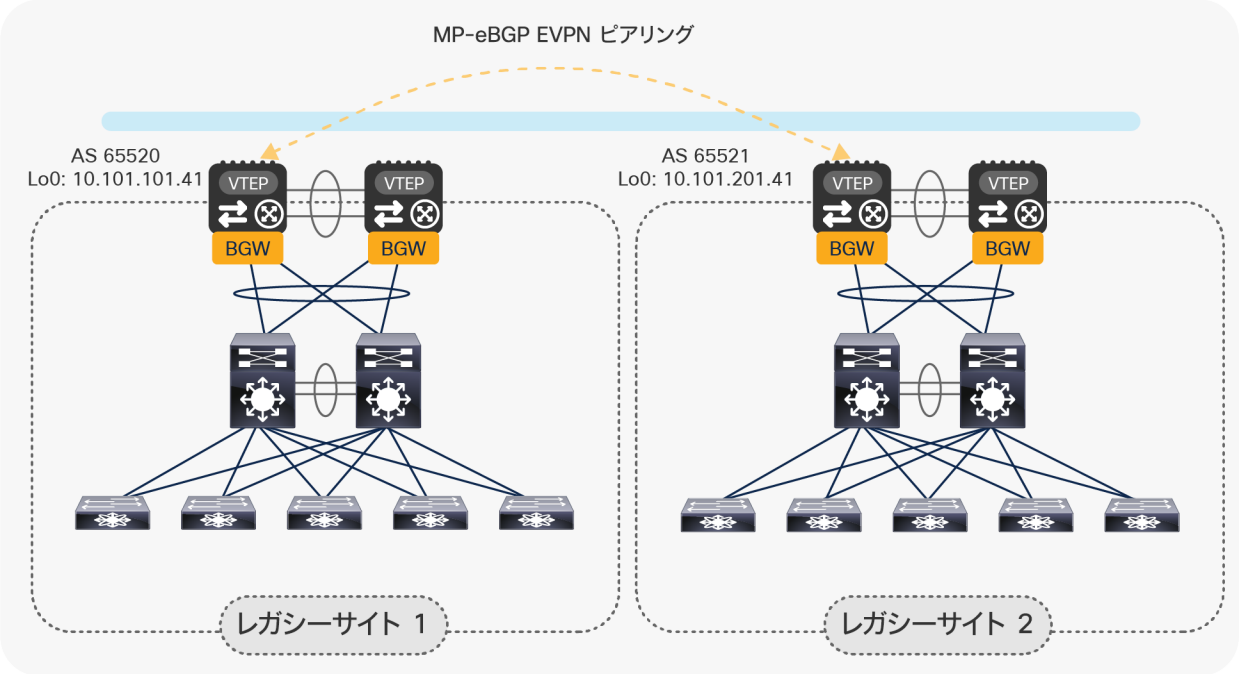

図 15.
別々のサイトの vPC BGW ノード間で MP-eBGP EVPN ピアリングが確立された状態

以下は図 15 に示す EVPN ピアリングの確立に必要な、具体的な設定を示したものです。

```
router bgp 65520
  router-id 10.101.101.41
  log-neighbor-changes
  neighbor 10.101.201.41
    remote-as 65521
    update-source loopback0
    ebgp-multihop 5
    peer-type fabric-external
    address-family l2vpn evpn
      send-community
      send-community extended
      rewrite-evpn-rt-asn
```

EVPN アドレスファミリ（L2VPN EVPN）を有効にしてリモート BGW ネイバーを設定します。ネイバーに指定された IP アドレスは、ネイバーの loopback0 CP IP アドレスを表します。eBGP ネイバーの設定は送信元インターフェイスをローカル loopback0 に指定して実行します。

リモート BGW デバイスはレイヤ 3 で数ホップ離れている可能性があるため、BGP セッションの TTL 設定を適切な値に増やす必要があります（**ebgp-multihop** コマンド）。

リモートサイト BGW にアドバタイズされるすべてのオーバーレイプレフィックスのネクストホップ IP とネクストホップ MAC（RMAC）を書き換えられるようにするため、各リモートマルチサイト BGW ネイバーに **peer-type fabric-external** を設定する必要があります。

最後に、リモート BGW にアドバタイズされるプレフィックスのルートターゲットの値を（BGP ネイバーのリモート ASN に基づいて）書き換えられるようにするため、EVPN アドレスファミリに適用される **rewrite-evpn-rt-asn** 設定を指定する必要があります。

注： ルートサーバを導入しない場合、上記と同じ設定をすべてのリモート BGW ノードに適用する必要があります。

ステップ 4：vPC BGW をサイト間に設定して DCI レイヤ 2 を拡張する

アンダーレイおよびオーバーレイのコントロールプレーンの設定が完了したら、まず vPC BGW ノードを設定して、レガシーサイト間にレイヤ 2 拡張サービスを提供する必要があります。この設定を行うには、BGW ノードとレガシーネットワークの間に確立された vPC 接続（レイヤ 2 トランク）への拡張を必要とする VLAN を許可し、各 VLAN を BGW ノードの L2 VNI セグメントに関連付けます。

EVPN マルチサイト レプリケート オーバーレイ BUM トラフィックは、DCI オーバーレイで入力レプリケーション（IR）モードを使用する点に注意してください。DCI アンダーレイネットワークにはマルチキャスト機能は不要です。また、集約ストーム制御機能を設定してサイト間のレイヤ 2 BUM トラフィックの伝播を制御し、制限することもできます。

EVPN マルチサイト vPC BGW の場合は、vPC BGW で **evpn multisite border-gateway** と **multisite border-gateway interface** コマンドを設定する必要があります。

次の例は、vPC BGW VTEP のレイヤ 2 拡張に必要な設定を示したものです。

`evpn multisite border-gateway 1` `  delay-restore time 300`	サイト ID を定義します。同じサイト上の vPC BGW ペアは同じ値のサイト ID を使用する必要があります。「delay-restore time」コマンドは、BGW が指定秒数の間リロードされた場合に、管理上の目的でマルチサイト VIP をシャットダウン状態のまま維持するために使用します。この例では 300 秒に設定しています。

`interface loopback100`
`  description Multi-Site VIP`
`  ip address 10.10.12.1/32 tag 54321`
`  ip pim sparse-mode`
 	
`interface loopback1`
`  ip address 10.10.10.1/24 tag 54321`
`  ip address 10.10.11.1/24 secondary tag 54321`
	マルチサイト仮想 IP アドレス（マルチサイト VIP）として使用するループバック インターフェイスと、プライマリ IP アドレス（PIP）および vPC 仮想 IP アドレス（vPC VIP）として使用するループバック インターフェイスを定義します。

`vlan 5`
`  vn-segment 30005`
 	
`vlan 6`
`  vn-segment 30006`
	対応するレイヤ 2 VNI に VLAN をマッピングします（マッピング対象の VLAN は、レガシーネットワークで確立されている vPC 接続にトランキングする必要がありますが、その設定はこの画面上には表示されていません）。

注： これらの VLAN を従来の DCI ソリューション（OTV、VPLS など）を使用してすでに拡張している場合は、データセンターサイト間でエンドツーエンドのレイヤ 2 ループが生じないように対処することが重要です。これには以下の 2 つの方法（VLAN 単位）があります。

- 従来の DCI ソリューションで VLAN 拡張機能を無効にし、マルチサイトを利用してデータセンター間のレイヤ 2 接続を提供します。レガシー DCI ソリューションを刷新することが最終目標のため、この手法を利用することを推奨します。
- 従来の DCI ソリューションによる VLAN 拡張機能を残し、レガシーネットワークと vPC BGW ノード間の 2 つの vPC 接続の一方に VLAN をトランキングしないようにします。これは移行の初期段階、具体的には、VLAN に属するエンドポイントのデフォルトゲートウェイを未だレガシーネットワーク内のアグリゲーションレイヤ デバイスによって提供している場合や、各データセンターサイトにローカルのデフォルトゲートウェイを提供する目的で特有の機能（OTV による HSRP フィルタリングなど）を提供している場合に有用な場合があります。

interface nve1
  no shutdown
  host-reachability protocol bgp
  source-interface loopback1
  multisite border-gateway interface loopback100
  global ingress-replication protocol bgp
  member vni 30005
  multisite ingress-replication
  ingress-replication protocol bgp
  member vni 30006
  multisite ingress-replication
  mcast-group 239.1.1.1
	選択的アドバタイズメントを行うために、レイヤ 2 VNI を NVE インターフェイス（VTEP）に関連付けます。これにより関連付けられたレイヤ 2 VNI のみが DCI 全体に拡張されます。
	サイト間 BUM トラフィックのレプリケーションモードを設定します。ここでは ingress-replication にします。
	レイヤ 2 VNI（L2VNI）の BUM レプリケーションは必ず設定します。たとえば、mcast-group や ingress-replication protocol bgp などとします。NVE ごとにグローバルデフォルト（global ingress-replication protocol bgp）を指定して設定を簡素化できます。グローバルの値は VNI ごとの設定で上書きされます。

ステップ 5：vPC BGW でエニーキャストゲートウェイを有効にし、シャットダウン状態を維持する

移行の希望として多いのは、ファーストホップ ゲートウェイ機能をレガシーネットワークのアグリゲーションスイッチから vPC BGW ノードに移行することです。この作業は、特定のサイトでローカルに定義されている IP サブネットと、サイト全体に拡張する必要のある IP サブネットの両方を対象にできます（前述の設定手順を参照）。

レガシーネットワークでは通常、Hot Standby Router Protocol（HSRP）、Virtual Router Redundancy Protocol（VRRP）、Gateway Load-Balancing Protocol（GLBP）などの First-Hop Redundancy Protocol（FHRP）をアグリゲーションスイッチで使用しています。一方 vPC BGW では、安定したファーストホップ ゲートウェイを提供するために分散エニーキャストゲートウェイ（DAG）を使用します。

注： サイト全体に拡張される IP サブネットでは、DAG を使用することで DCI ネットワーク全体でトラフィックのヘアピン化を防ぐ、ローカルで安定性のあるデフォルトゲートウェイ機能をプロビジョニングできます。

これらの異なるファーストホップ ゲートウェイを共存させる手法は、サポートの対象外です。したがって、デフォルトゲートウェイ機能を vPC BGW ノードに移行するにあたっての最初のステップは、図 16 にあるとおりエニーキャストゲートウェイ SVI を作成し、初期状態ではシャットダウンにしておくことです。

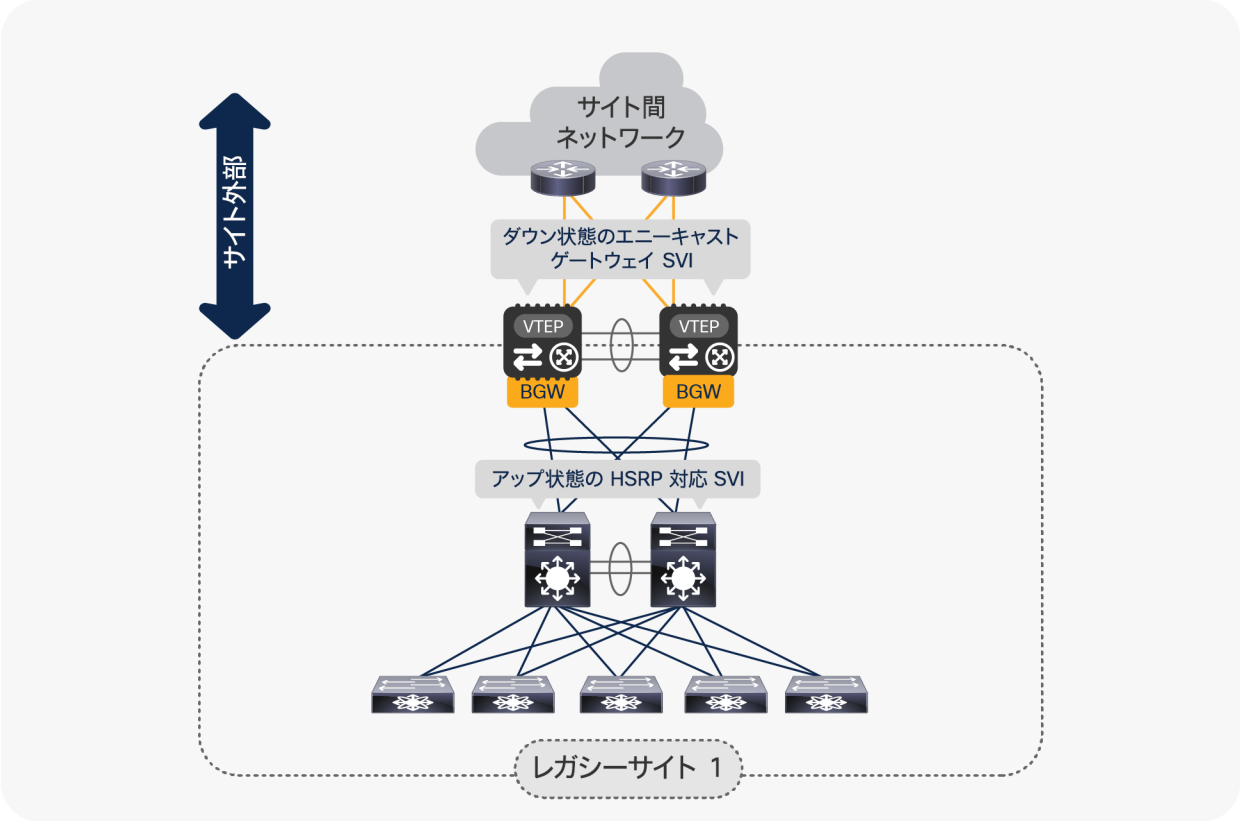

図 16.
エニーキャストゲートウェイ SVI と HSRP ゲートウェイ SVI の共存

次のサンプルは、vPC BGW ノードで必要となるレイヤ 3 設定を作成する手順を示したものです。

`fabric forwarding anycast-gateway-mac 2020.0000.00AA`	定義済みの全テナント SVI にエニーキャストゲートウェイ MAC アドレス（この例では 2020.0000.00AA）を定義します。
`vlan 2001` `  vn-segment 50001`	予約済み VLAN の 1 つを、所定の VRF（tenant-1）で使用する L3 VNI にマッピングします。
`vrf context tenant-1` `    vni 50001`	テナント VRF を定義し、定義済みの L3 VNI に関連付けます。
`interface nve1` `  member vni 50001 associate-vrf`	その L3 VNI を NVE インターフェイスに関連付けます。

`interface Vlan5`
`  shutdown`
`  vrf member tenant1`
	エニーキャストゲートウェイとして使用する SVI を定義し、シャットダウンモードのままにしておきます。ここでは、IP サブネットのプレフィックスをサイト間のオーバーレイ コントロールプレーンに再配布できるように特定のタグを使用しています。

ip address 10.1.5.1/24 tag 12345 fabric forwarding mode anycast-gateway	

router bgp 65520 vrf tenant-1 address-family ipv4 unicast redistribute direct route-map FABRIC-RMAP-REDIST-SUBNET maximum-paths ibgp 2 address-family ipv6 unicast redistribute direct route-map FABRIC-RMAP-REDIST-SUBNET maximum-paths ibgp 2	リモート BGW ノードと L3 プレフィックスのやり取りを開始できるように、BGP プロセスで VRF を設定します。 **注：**max-path はローカルファブリックにのみ必要です。

route-map FABRIC-RMAP-REDIST-SUBNET permit 10 match tag 12345	IP サブネット情報を EVPN コントロールプレーンに再配布するためのルートマップを定義します。

ステップ 6：レガシーサイトのファーストホップ FHRP ゲートウェイを vPC BGW エニーキャストゲートウェイに移行する

次の手順に沿ってレガシーサイトの FHRP ゲートウェイを vPC BGW ペアのエニーキャストゲートウェイに移行します。この作業は IP サブネットごとのアグリゲーションレイヤ スイッチに対して実施できます。

interface vlan 20 vrf member Tenant-A ip address 192.168.20.201/24 hsrp 10 ip 192.168.20.1 mac-address 2020.0000.00aa	すべての FHRP ゲートウェイの MAC アドレスと IP アドレスを、マルチサイト vPC BGW の分散型 IP エニーキャストゲートウェイ（DAG）の設定に合わせます。エニーキャストゲートウェイの仮想 MAC アドレスは VXLAN EVPN VTEP のグローバル設定パラメータであるため、異なる IP サブネットすべてに同じ仮想 MAC アドレスを使用する必要があります。

注： レガシーネットワークのアグリゲーションスイッチが定義済み SVI の静的 MAC 設定に対応していない場合は、すべての SVI が同じ HSRP グループを使用するように設定を変更できます。この設定を行うと、（HSRP グループ番号と直結するように）動的に同一の vMAC が作成され、BGW ノードで定義されたグローバル vMAC 値がその値に一致するように変更できます。

レガシーサイトの vMAC の変更が終わると、状態変更（スタンバイからアクティブへの切り替え）を実施して、FHRP から Gratuitous ARP（GARP）の処理を強制的に行います。GARP の働きにより、エンドポイントの ARP キャッシュ内の MAC アドレスを更新して、新たに作成された vMAC に一致させることができます。

レガシーネットワークのアグリゲーション レイヤ スイッチと vPC BGW の間に、VRF 単位でルーティングピアリングを確立します。この作業は、一度に 1 つの IP サブネットを対象としてデフォルトゲートウェイを vPC BGW に移行する場合に必要です。その目的はアグリゲーションレイヤ スイッチ上でデフォルトゲートウェイが依然機能している IP サブネットと、デフォルトゲートウェイを vPC BGW に移行済みの IP サブネットとの間でトラフィックをルーティングするためです。

注： すべてのサブネットを一度に移行する場合、この作業は必要ありません。

このレイヤ 3 ピアリングを確立するには、図 17 に示すように、専用のレイヤ 3 インターフェイスのペアを個別に使用することをお勧めします。マルチテナント（つまりマルチ VRF）環境では、個別のサブインターフェイスを定義できます。

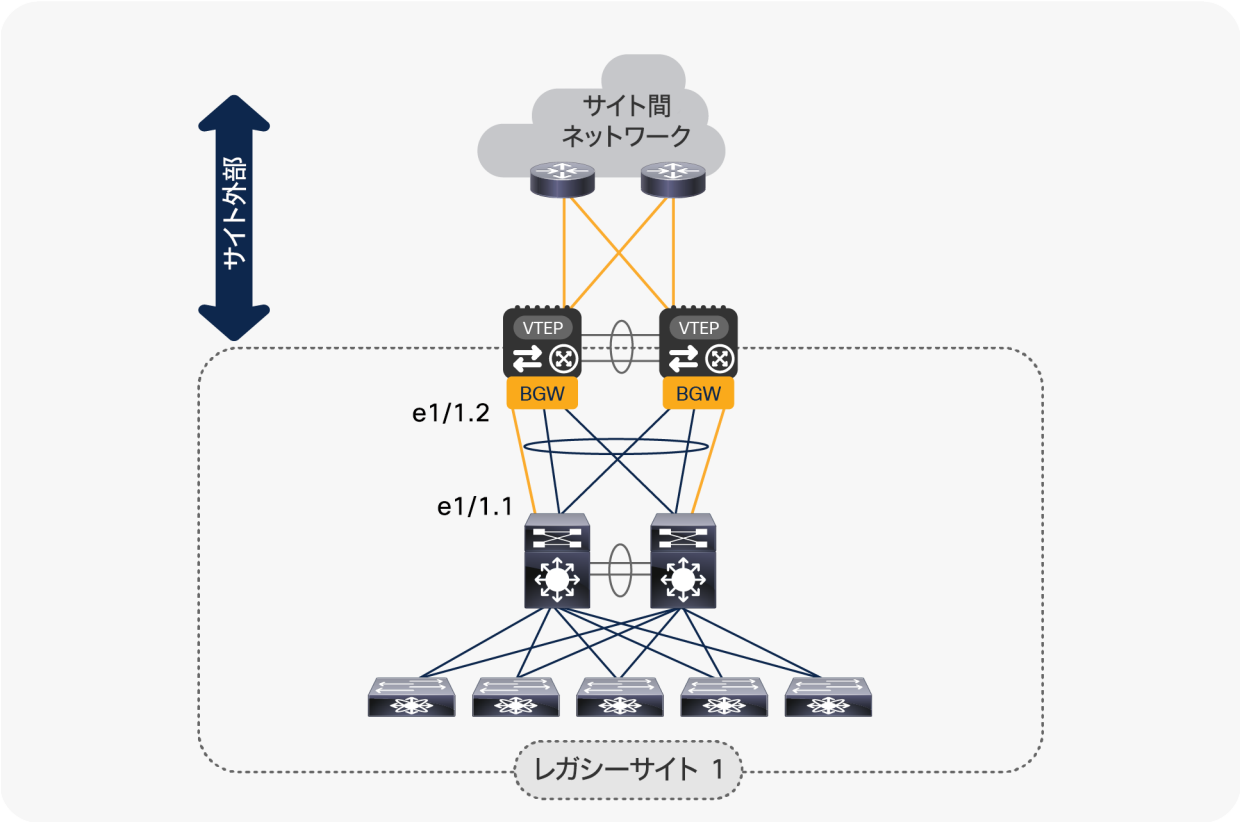

図 17.
vPC BGW ノードとレガシーネットワーク間の専用レイヤ 3 リンクの使用

次に、アグリゲーションスイッチで必要になる設定を示します。言うまでもなく、vPC BGW で同じ設定が必要になります。IPv4 BGP を使用して VRF ごとにピアリングを確立すると、リモート BGW ノードで使用されるオーバーレイ EBGP EVPN コントロールプレーンを使用した自動再配布が可能になります。

`interface Ethernet1/1.10`
`  description L3 Link to vPC BGW1 (T1)`
`  encapsulation dot1q 10`
`  vrf member Tenant-1`
`  ip address 192.168.36.4/31`
 	
`router bgp 65520`
`  router-id 100.100.100.1`
`  vrf Tenant-A`
`  neighbor 192.168.36.5`
`  remote-as 65520`
`  address-family ipv4 unicast`
	テナントごとにサブインターフェイスを作成し、BGP ネイバーとの IPv4 ルートの交換を有効にします。

注： ベストプラクティスではリンクかノードに障害が発生した場合のコンバージェンスを高速化するために、アグリゲーションレイヤ スイッチと vPC BGW の間にフルメッシュの L3 リンクを作成します。

この時点で、アグリゲーションレイヤの FHRP SVI を無効にし、vPC BGW ノードの DAG SVI を有効化できます。このステップではすべてのファーストホップ ゲートウェイの処理を、アグリゲーションレイヤに接続されている vPC BGW に移します。この作業はサブネット単位で行えることは前述のとおりです。

またこれも先に触れましたが、スパニングツリールートをアグリゲーションレイヤから vPC BGW に移すようお勧めします。ここでレガシーサイトにおけるイーサネットのネットワークは、BGW のサウスバウンドになりました。

注： FHRP に対する変更、または BGW、STP ルート、サイト間ルーティングピアリングとの接続に対する変更を行うと、既存ネットワークの動作を短時間ながら中断させるおそれがあります。こうした変更はメンテナンス期間中に実施してください。

この時点で、レガシー データセンター サイト間のレイヤ 2 およびレイヤ 3 接続を拡張するための移行手順は完了となります。以下で説明するおおまかな手順は、1 か所（またはすべて）のレガシーデータセンターにサイト内通信用の VXLAN EVPN を導入する場合にのみ必要な、任意作業です。

ステップ 7：レガシーデータセンターを新しい Cisco Nexus 9000 シリーズ スイッチと最新のファブリックテクノロジーに移行する

vPC BGW ノードのマルチサイト拡張機能を使用したレガシーデータセンターの相互接続が整ったら、レガシーネットワークを段階的に廃止し、最新のファブリックテクノロジー（VXLAN BGP EVPN ファブリックや Cisco ACI™ ファブリックなど）に置き換えていくことができます。以下で説明する手順は、前者（VXLAN BGP EVPN ファブリック）のシナリオに焦点を当てています。

注： 相互接続されたレガシーサイトの 1 つ（または 1 つのサブセット）のみを全面的に VXLAN EVPN ファブリックに移行する場合も、同じ手順を応用できます。

- 各レガシーサイトに VXLAN EVPN スパインと追加の VTEP を導入して新しい VXLAN EVPN ファブリックの構築を開始します。

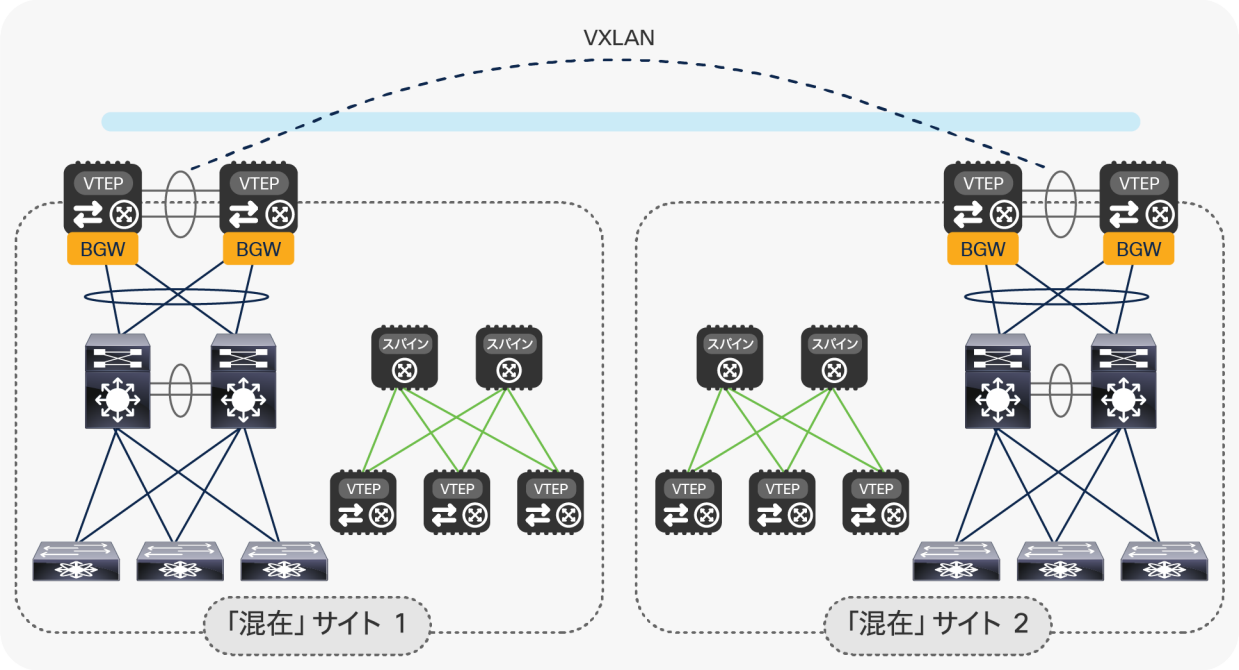

図 18.
各データセンターでの新しい VXLAN EVPN ファブリックの構築開始

- ポイントツーポイントのレイヤ 3 リンクを使用して新しいファブリックスパインを vPC BGW ペアに接続します。vPC BGW の設定を変更し、新しい VXLAN EVPN ファブリックと統合します。この変更を行ってもレガシーネットワーク間の既存の接続は影響を受けません。

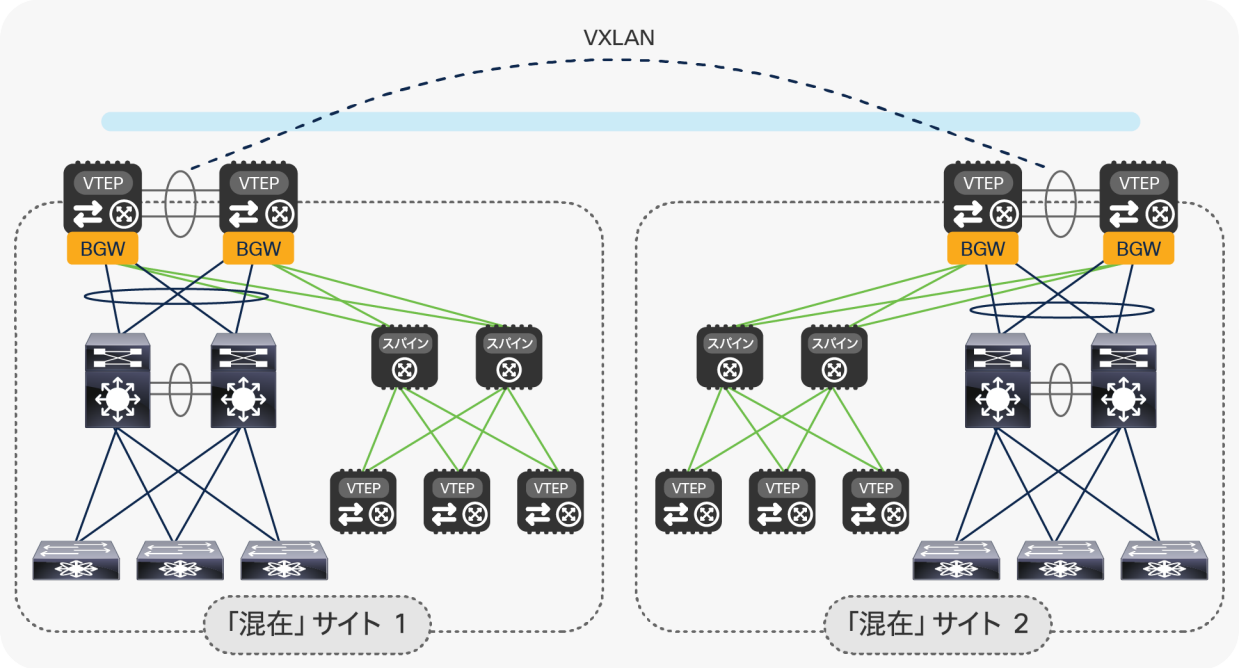

図 19.
vPC BGW ノードを使用してレガシーデータセンターを VXLAN EVPN ファブリックに移行する最初のステップ

- 新しく作成した VXLAN EVPN ファブリックとレガシーネットワークのローカル接続が完了したら、両者の間でアプリケーションとサービスの移行を開始できます。移行手順の最終的な状態は図 20 のようになります。すべてのアプリケーションとサービスが新しい VXLAN EVPN ファブリックに再配置され、古いレガシーネットワークデバイスの使用を停止した状態になっています。

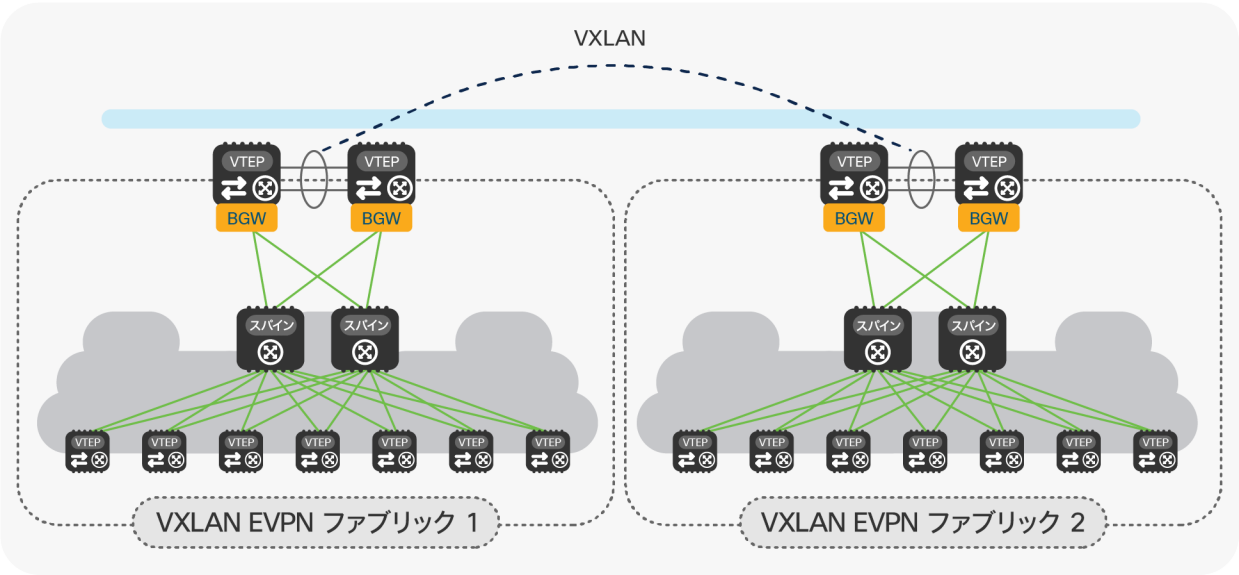

図 20.
vPC BGW ノードを使用したレガシーデータセンターの VXLAN EVPN ファブリックへの移行が終了した状態

この時点で vPC BGW ノードはローカルおよびリモートの VTEP デバイスに接続されているエンドポイント間の接続を拡張できる状態が整い、BGW の全機能を発揮できます。これはローカルサイト内に VTEP ノードが存在しない図 6 のシナリオとは対照的です。

- 最後の任意手順では BGW ノードの vPC 設定を削除し、エニーキャスト BGW に変換します（図 21 を参照）。

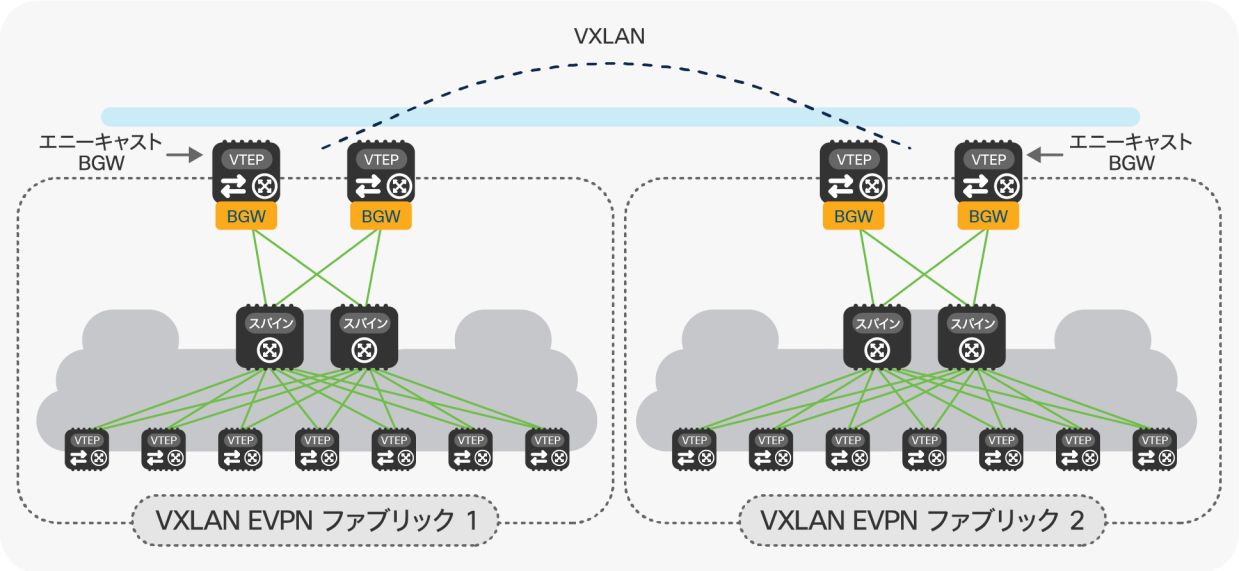

図 21.
vPC BGW からエニーキャスト BGW への変換

- これは VXLAN EVPN ファブリックを相互接続する場合の導入モデルとして推奨されていますが、このモデルを適用できるのは、vPC BGW をデフォルトゲートウェイとして使用するエンドポイントが、元の vPC BGW に接続されていない場合だけです。
- **注：**サイト間のレイヤ 2 およびレイヤ 3 接続が中断されないようにするため、エニーキャストモードへの変換を実行できるのは一度に 1 つの BGW です。

まとめ

データセンターの導入は、高速化とポートの高密度化を実現し、豊富な機能を備えた Cisco Nexus 9000 シリーズ スイッチベースのインフラストラクチャへと急速に移行しつつあります。Cisco Nexus 9000 シリーズ スイッチの Cisco VXLAN EVPN マルチサイト ソリューションは、ファブリックの拡張、コンパートメント化、データセンター相互接続（DCI）などの多くの活用例に対応すべく、まったくのゼロから設計されたソリューションです。

このドキュメントではマルチサイト導入の具体的な選択肢として、vPC ボーダーゲートウェイ（BGW）を活用することで有効に対処できるシナリオを紹介し、中でも個別のデータセンターサイトを相互接続する最新のソリューション（DCI の使用例）を重点的に取り上げました。

VXLAN EVPN ファブリックの相互接続にあたっては、これまでどおりエニーキャスト BGW を利用する手法が推奨されるものの、VXLAN EVPN ファブリックと、従来のテクノロジー（STP、vPC、Cisco FabricPath など）で構築されたレガシー データセンター サイトとの間のレイヤ 2 およびレイヤ 3 接続を拡張する必要がある場合のような特殊なケースでは、vPC BGW の導入によって得られるメリットがあります。

同時に、レガシー データセンター ネットワークを相互接続する従来の DCI テクノロジーを vPC BGW で置き換える

ことができます（データセンターネットワーク内に VXLAN EVPN テクノロジーを導入する前であっても可能）。これができるのは vPC BGW によるネットワーク制御、VTEP マスキング、BUM トラフィック適用関連の機能のおかげです。これらの機能が EVPN マルチサイトアーキテクチャを非常に効率的な DCI テクノロジーにしています。

参考資料

[1] 『VXLAN EVPN マルチサイト設計および導入ホワイトペーパー』

https://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-739942.html#_Toc498025695

[2] vPC ベストプラクティス設定ガイド

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/nexus9000/sw/6-x/interfaces/configuration/guide/b_Cisco_Nexus_9000_Series_NX-OS_Interfaces_Configuration_Guide/b_Cisco_Nexus_9000_Series_NX-OS_Interfaces_Configuration_Guide_chapter_0111.html

[3] 『Cisco FabricPath 環境から VXLAN BGP EVPN への移行』または『クラシック イーサネット環境から VXLAN BGP EVPN への移行』ホワイトペーパー

FabricPath からの移行：https://www.cisco.com/c/en/us/td/docs/dcn/whitepapers/migrating-fabricpath-environment-vxlan-bgp-evpn.html

クラシックイーサネットからの移行：https://www.cisco.com/c/en/us/td/docs/dcn/whitepapers/migrating-classic-ethernet-to-vxlan-bgp-evpn-white-paper.html

付録：vPC BGW を使用した VXLAN EVPN マルチサイトの設計と導入に関する考慮事項

vPC のボーダーゲートウェイ機能の導入により、これまで利用できたエニーキャスト ボーダー ゲートウェイの導入に代わるデータセンターサイトの相互接続モデルが新たに加わりました（図 22）。EVPN マルチサイト環境では、一部のサイトにはエニーキャスト BGW を、別のサイトには vPC BGW を導入するというようにさまざまなバリエーションの BGW を混在させることができます。

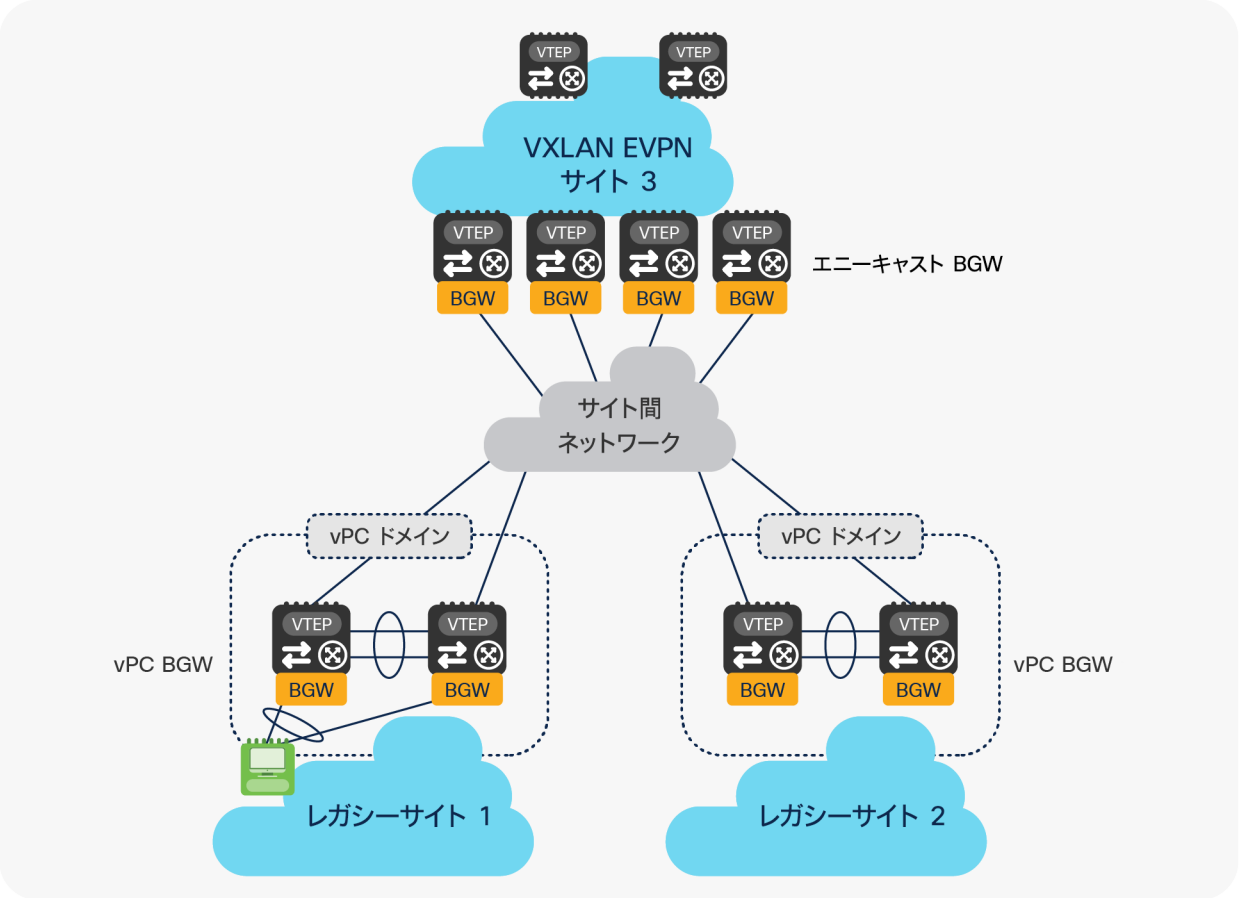

図 22.
エニーキャスト BGW と vPC BGW

vPC BGW 環境には具体的に次のような特徴があります。

- vPC BGW モデルでは、2 台の Cisco Nexus 9000 デバイスを同じ vPC ドメインに取り入れる必要があります。2 つの BGW を vPC ピアリンクで相互接続（およびピアキープアライブリンクを定義）しなければならない点を含め、vPC の一般的なベストプラクティス設定が適用されます。vPC BGW ノードのペアを導入する際に必要なベストプラクティスの vPC 設定については、本文の「vPC BGW によるレガシーデータセンターの VXLAN EVPN ファブリックへの移行」セクションで詳しく説明します。
- レイヤ 2 モードでのローカルエンドポイント（およびサービスノード）の接続は vPC BGW でサポートされています。このローカルエンドポイントが、ローカル接続されたエンティティのファーストホップ ゲートウェイとして機能します。エンドポイントはデュアル接続かシングル接続が可能です。
- サポートされる vPC BGW ペアは各サイトにつき 1 つだけです。
- vPC BGW は BGW 機能を提供し、ローカルエンドポイント接続をサポートするだけでなく、サイトを外部レイヤ 3 ネットワークドメインに相互接続するボーダーリーフノード（つまりノースサウス接続）としても機能します。
- 次の表 1 は、Cisco Nexus 9000 プラットフォームで vPC BGW 機能をサポートするために必要となるハードウェアとソフトウェアの依存関係を示したものです。

表 1. EVPN vPC ボーダーゲートウェイのソフトウェアおよびハードウェアの最小要件

項目	要件
Cisco Nexus ハードウェア	• Cisco Nexus 9300 EX プラットフォーム • Cisco Nexus 9300 FX プラットフォーム • Cisco Nexus 9332C プラットフォーム • Cisco Nexus 9364C プラットフォーム • Cisco Nexus 9500 プラットフォーム（X9700-EX ラインカード装備） • Cisco Nexus 9500 プラットフォーム（X9700-FX ラインカード装備）
Cisco NX-OS ソフトウェア	Cisco NX-OS ソフトウェアリリース 9.2(1) 以降

注： BGW 機能は Cisco Nexus 9348GC-FXP スイッチではサポートされていません。

その他導入に関して考慮すべき事項（たとえば、サイト間における BUM トラフィック転送での入力レプリケーションの使用、各 BGW ノードでのアンダーレイおよびオーバーレイピアリングの確立方法など）は、エニーキャスト BGW 導入モデルと共通です。詳細については、『VXLAN EVPN マルチサイト設計および導入ホワイトペーパー』を参照してください。

vPC BGW の論理インターフェイス

さまざまな論理インターフェイス（ループバック インターフェイスなど）がそれぞれの役割を果たすには、図 23 に示すように vPC BGW デバイスで論理インターフェイスを定義する必要があります。

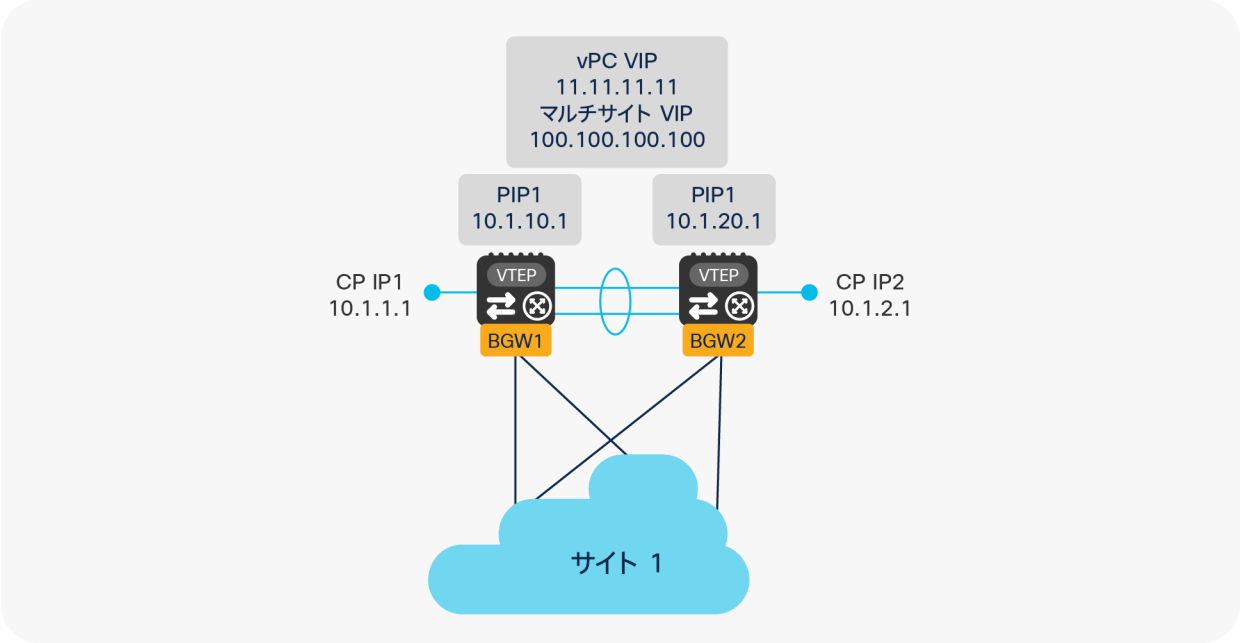

図 23.
vPC BGW 論理インターフェイス

- コントロールプレーンの IP アドレス（CP IP）：各 BGW ノードで定義される一意の IP アドレスです。MP-BGP EVPN オーバーレイがリモート BGW デバイスとのコントロールプレーン隣接関係を確立するために使用されます。この IP アドレスは、VXLAN カプセル化トラフィックの送受信には使用されず、アンダーレイ ルーティング プロトコルのルータ ID（RID）として機能します。

- プライマリ IP アドレス（PIP）：各 BGW ノードで定義される一意の IP アドレスです。レイヤ 3 接続を介して BGW に接続されているデバイスから発信されたトラフィックを送信し、リモートサイトから発信された同じエンティティ宛てのトラフィックを受信するために使用されます。たとえば、BGW ノードがボーダーリーフとして外部レイヤ 3 ドメインとのノースサウス接続を提供する場合などが考えられます。プライマリ IP アドレスを使用する場合は「advertise-pip」を設定して有効化する必要があります。
- vPC 仮想 IP アドレス（vPC VIP）：同じ vPC ドメインの両 BGW ノード部で一般的に定義されるセカンダリ IP アドレスで、次のような 2 つの使用目的があります。

  1. リモートサイトに拡張されたレイヤ 2 ネットワークの BUM トラフィックの発信。
  2. BGW にレイヤ 2 で ローカル接続されたシングルまたはデュアル接続エンドポイント間のトラフィックの発信と受信（図 24）。

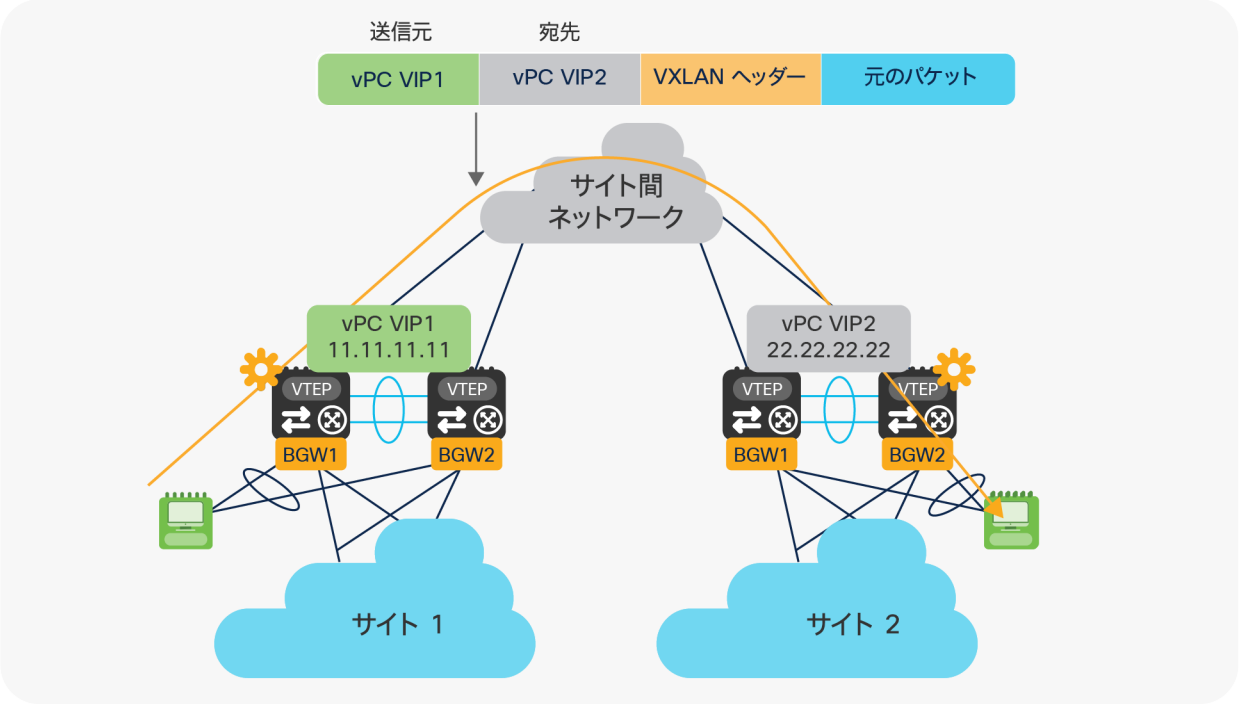

図 24.
vPC BGW ノードでの vPC VIP アドレスを使用したトラフィックの送信と受信

- マルチサイト仮想 IP アドレス（マルチサイト VIP）：同じ vPC ドメインに属する両 BGW ノードで一般的に定義される専用ループバックの IP アドレスです。この IP アドレスは、リモートサイトを宛先とし、なおかつローカルサイトのリーフノードの後方に接続されているエンドポイントを送信元とするトラフィックを送信するために使用されます。また、リモートサイトを送信元とし、なおかつローカルサイトのリーフノードの後方に接続されているエンドポイントを宛先とするトラフィックの受信にも同じ IP アドレスが使用されます（図 25）。

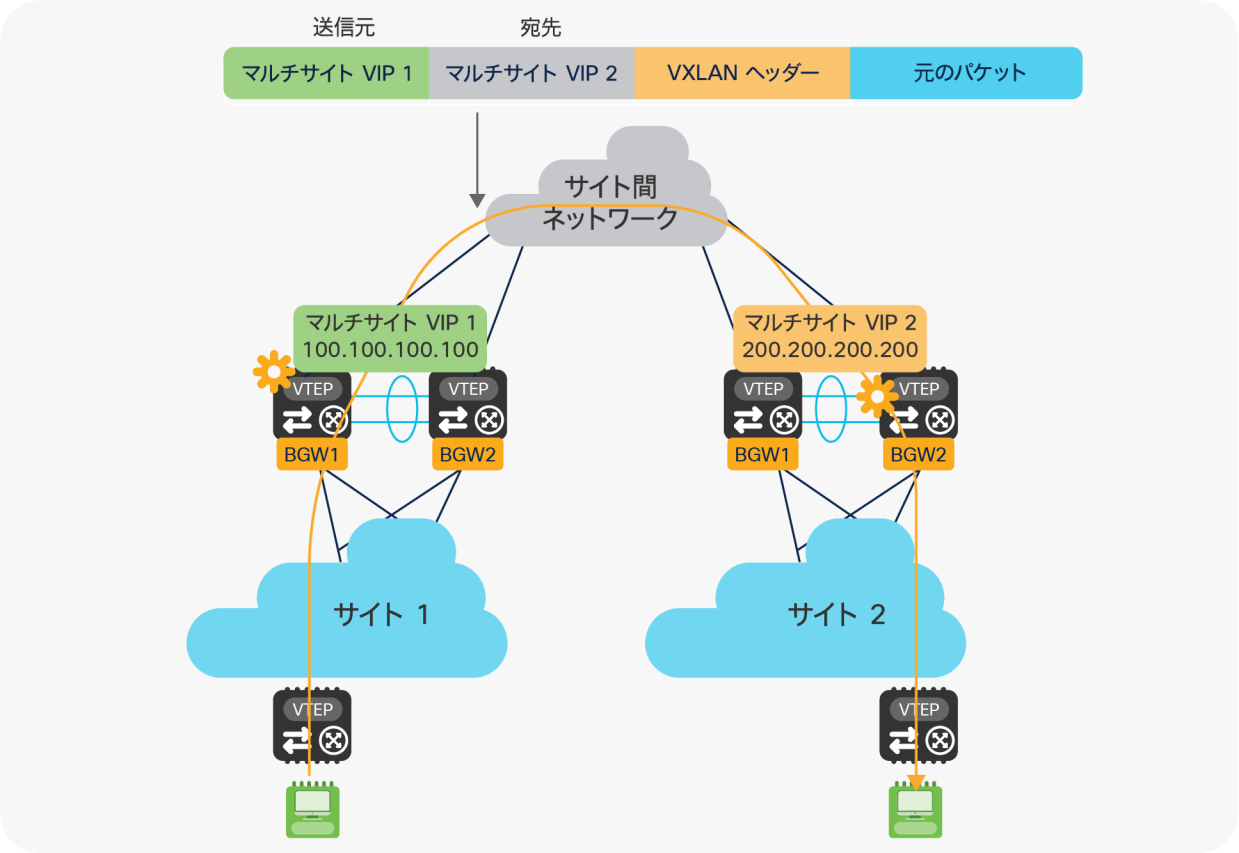

図 25.
vPC BGW ノードでのマルチサイト VIP アドレスを使用したトラフィックの送信と受信

次のサンプルは、vPC BGW ノードで必要なループバックアドレスの設定を示したものです。

```
interface loopback0
 description CP IP or RID
 ip address 10.1.1.1/32 tag 54321
!
interface loopback1
 description PIP1
 ip address 10.1.10.1/32 tag 54321
 ip address 11.11.11.11/32 secondary tag 54321
!
interface loopback100
 description Multi-Site VIP1
 ip address 100.100.100.100/32 tag 54321
!
interface nve1
 host-reachability protocol bgp
 source-interface loopback1
 multisite border-gateway interface loopback100
```

上記のように、PIP、vPC VIP、マルチサイト VIP の割り当てに使用する定義済みのループバック インターフェイスは、論理 NVE インターフェイスの設定で指定する必要があります。このことは、これらの IP アドレスを前述のさまざまなシナリオにおける VXLAN データプレーンのトラフィックに使用する必要があることを示しています。

注： コマンドに「tag 54321」を付加すると、サイト間で使用されるアンダーレイ コントロール プレーン プロトコルへのループバックプレフィックスの再配布が容易になります。その詳細については、本文の「vPC BGW によるレガシーデータセンターの VXLAN EVPN ファブリックへの移行」セクションで説明しています。

EVPN マルチサイト vPC BGW での障害シナリオ

vPC BGW は相互接続されている他方のサイトに対するデータセンターネットワークのインターフェイスを表すため、vPC BGW によって障害シナリオがどのように処理されるかを理解するのは特に重要なことです。

そのために、別のサイト（サイト外部ネットワーク）の vPC BGW 間に到達可能性を提供する転送ネットワークでの障害と、ローカル VTEP ノード（サイト内部ネットワーク）に接続を提供するサイト内部でのネットワーク障害とを区別する必要があります。

インターフェイス トラッキングは各 BGW ノードに実装されたメカニズムで、サイト内部またはサイト外部ネットワークとの接続が失われる可能性を検出することで、そうしたイベントに適切に対処する機能です。さらに必要に応じてトラフィックデータパスから分離されたノードを削除することで、トラフィックが廃棄されてしまうブラックホール化が起こるのを防ぎます。

次のサンプルは、vPC BGW インターフェイスのモニタリング機能を有効にするための設定を示したものです。

```
interface Ethernet1/1
 description L3 Link to Site-External Network
 ip address 10.111.111.1/30
 evpn multisite dci-tracking
!
interface Ethernet1/2
 description L3 Link to Site-Internal Network
 ip address 10.0.1.5/30
 evpn multisite fabric-tracking
```

注： 同じ vPC ドメインの 2 つの BGW ノード間で確立されている vPC ピアリンク接続の状態を追跡する必要はありません。

サイト外部ネットワークからの vPC BGW の分断

図 26 は、vPC BGW ノードでサイト外部ネットワークとのすべての物理接続が失われる障害を示した具体的なシナリオです。

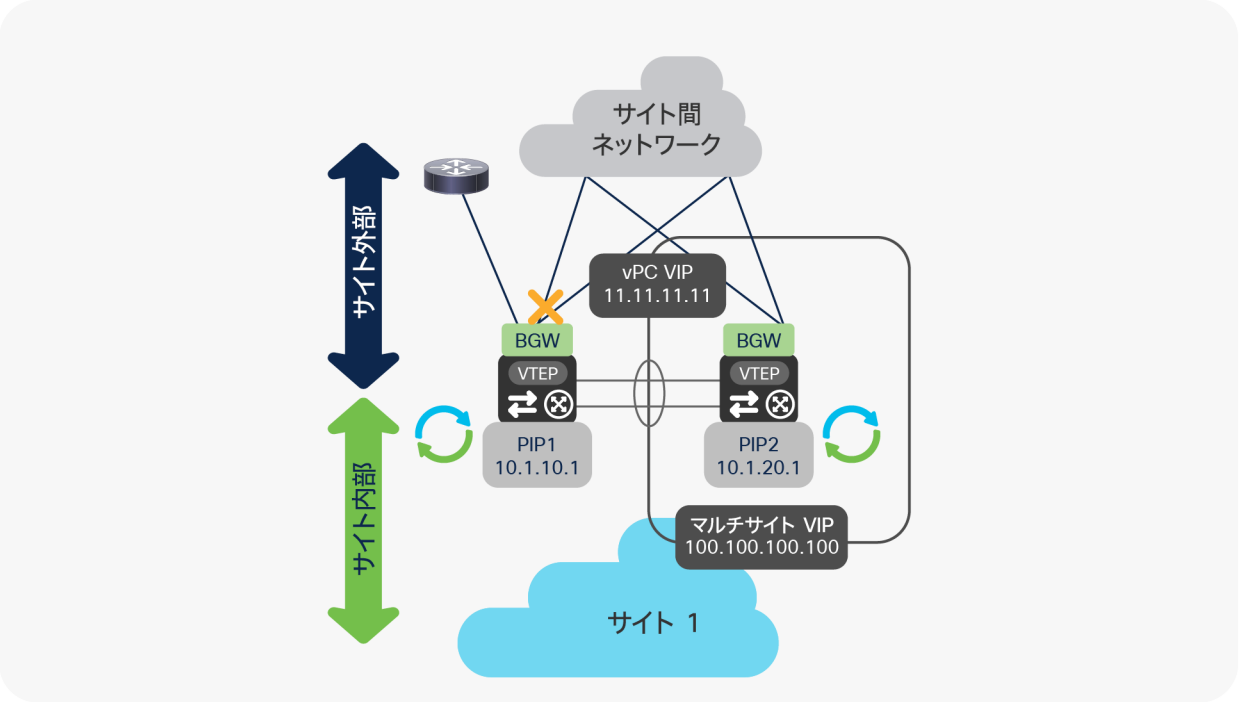

図 26.
サイト外部ネットワークからの vPC BGW の分断

このような状況では、サイト外部ネットワークから分離された vPC BGW ノードで、次のようなイベントが続けて発生します。

- vPC ピアリンクで確立されたレイヤ 3 隣接関係を介して、PIP1 アドレスと vPC VIP アドレスがサイト内部ネットワークとピアの BGW に向けてアドバタイズされ続けます。このアドバタイズの処理は、ローカルサイトに接続されたエンドポイントとリモートサイトのエンドポイントの両方から、外部ネットワークとローカルエンドポイント（分断された BGW ノード経由でのみ到達可能）に接続するために必要です。

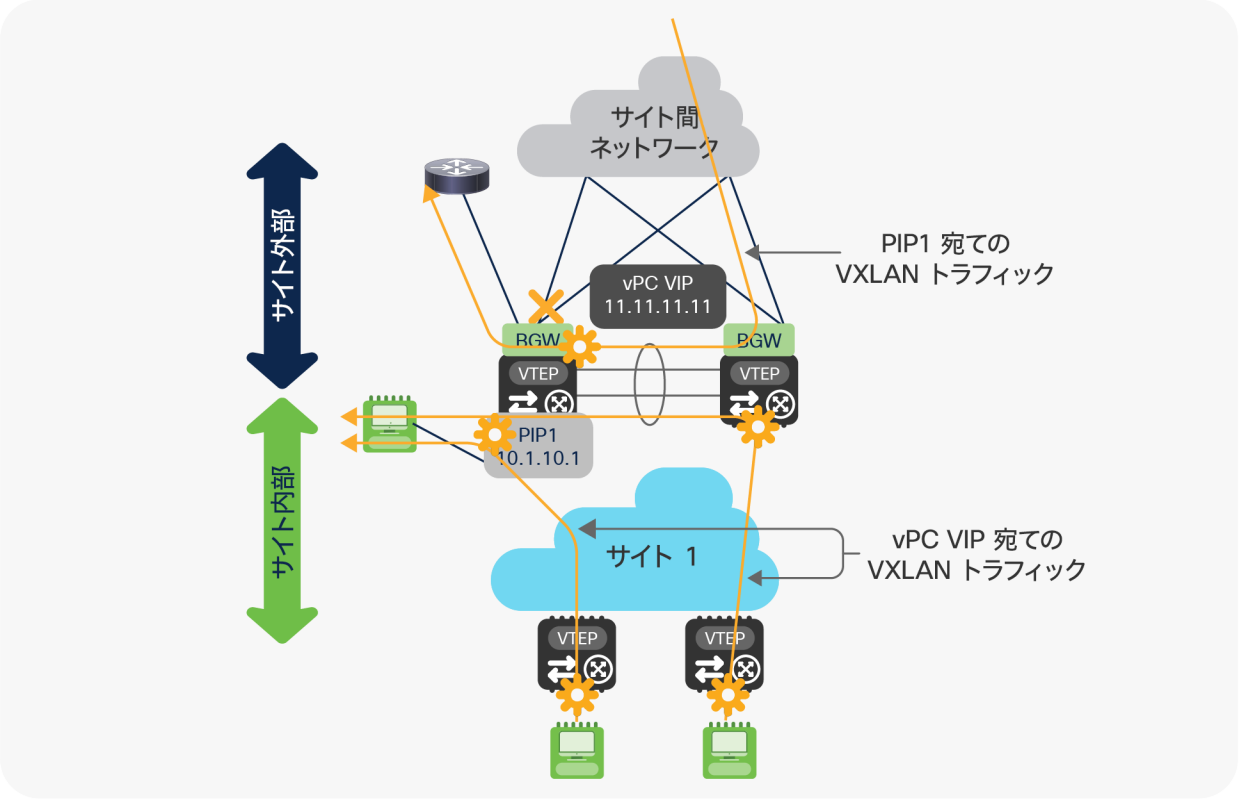

図 27.
分断された BGW ノードでの PIP アドレスと vPC VIP アドレスの使用

図 27 に示すように、vPC BGW ノードがサイト外部ネットワークから分断されると、大量のトラフィックが vPC ピアリンクに流れ始める可能性があります。そのため、この増加したトラフィックに対応できるように、ピアリンクで使用できる帯域幅を適切に設定することが重要です。

- マルチサイト VIP アドレスのサイト内部ネットワークへのアドバタイズが停止します。これによりローカルリーフノードに接続され、なおかつリモートサイトを宛先とする（およびその逆の）エンドポイントから発信されるトラフィックを、サイト外部ネットワークとの接続が健在の vPC BGW ノードに直接誘導できます（この場合 vPC ピアリンク接続を使用する必要はありません）。

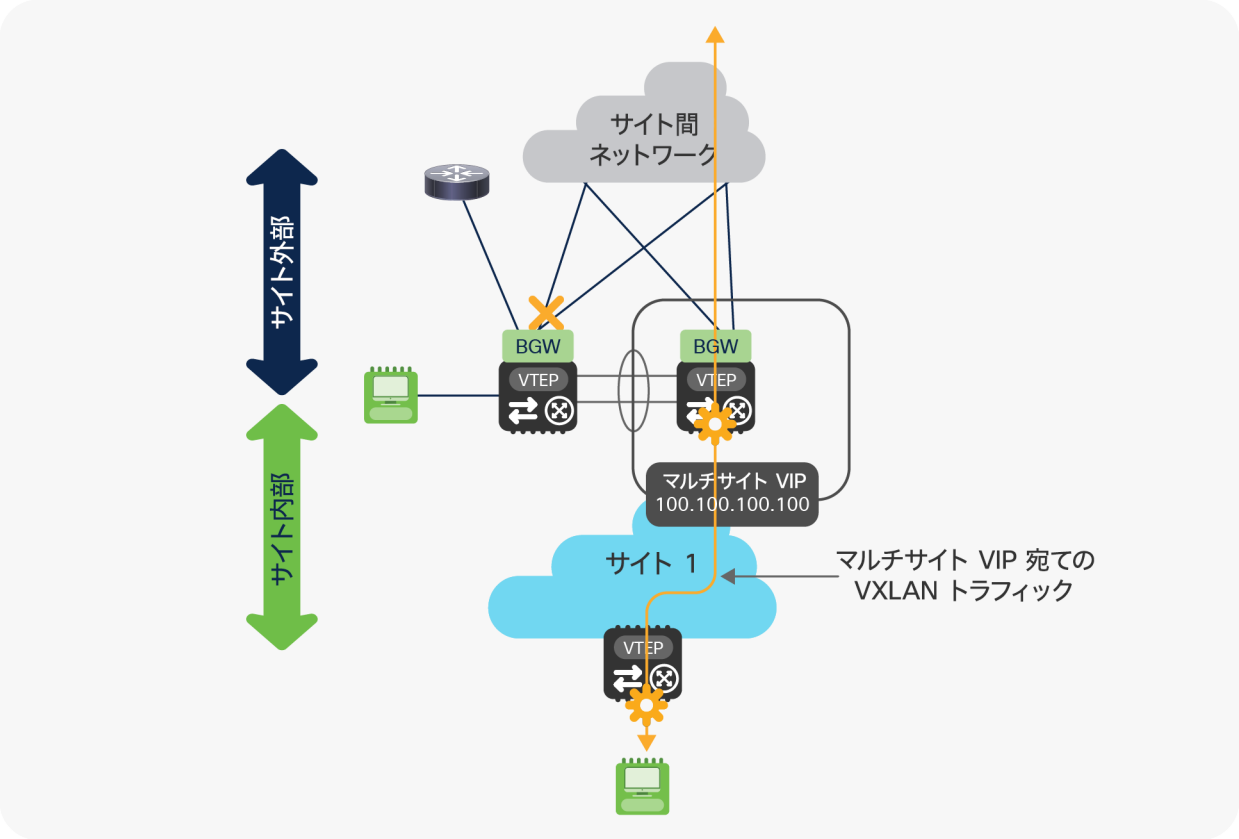

図 28.
サイト外部ネットワークとの接続が健在の vPC BGW 上にあるマルチサイト VIP のみ使用

サイト外部ネットワークから分断されている状態が検出されると、BGW ノードの関連するループバック インターフェイスが管理上の措置として動的にシャットダウンされるため、マルチサイト VIP のアドバタイズが停止する点に注意してください。

- サイト外部ネットワークとの接続が 1 つでも回復すると、BGW ノードはサイト外部ネットワークとのダイレクト アンダーレイ ピアリングの再確立を開始できます。マルチサイト VIP ループバック インターフェイスは、設定時間（デフォルト値は 300 秒）の間、ダウン状態のままになります。

サイト内部ネットワークからの vPC BGW の分断

図 29 は、vPC BGW ノードでサイト内部ネットワークとのすべての物理接続が失われる障害を示した具体的なシナリオです。

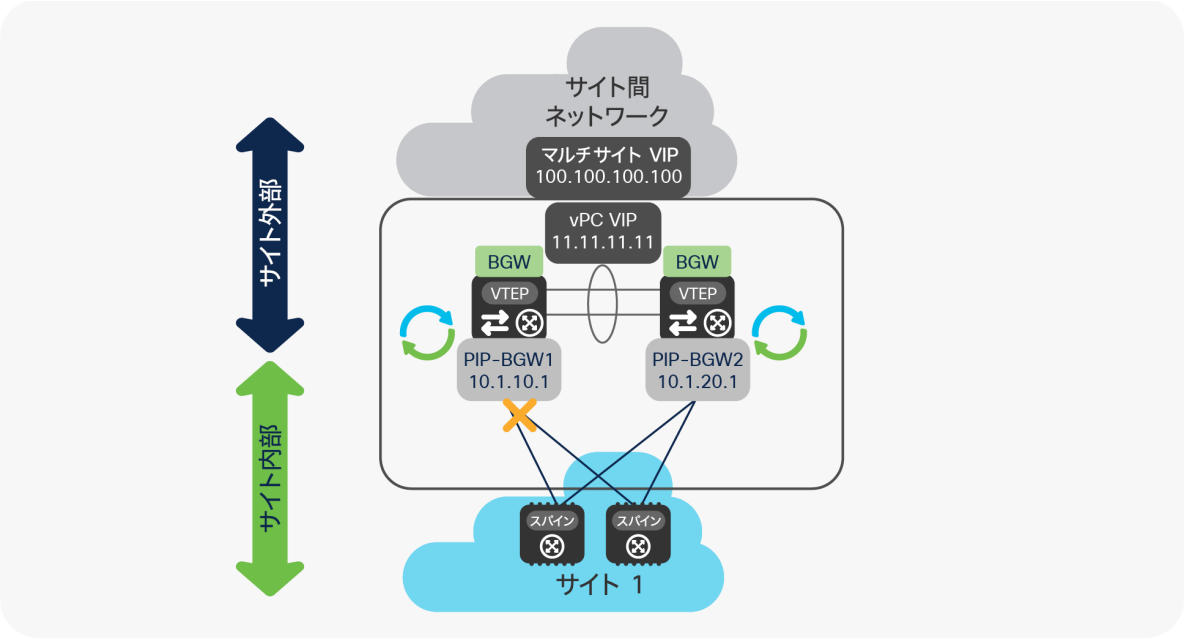

図 29.
サイト内部ネットワークからの vPC BGW の分断

このような状況では、分断された BGW のすべての論理インターフェイス（PIP、vPC VIP、マルチサイト VIP）はアクティブ状態のままになり、そのアドレスはサイト外部ネットワーク（および vPC ピアリンク上に確立されたレイヤ 3 隣接関係を介したピア BGW）に向けて引き続きアドバタイズされます。

この状態は、リモートサイトからの着信トラフィックフローの 5 割を、分断された BGW ノードと直接つながっているエンドポイントやネットワークから発信されるトラフィックフロー全体と合わせて、vPC ピアリンク経由で転送しなければならないことを意味します（図 30）。

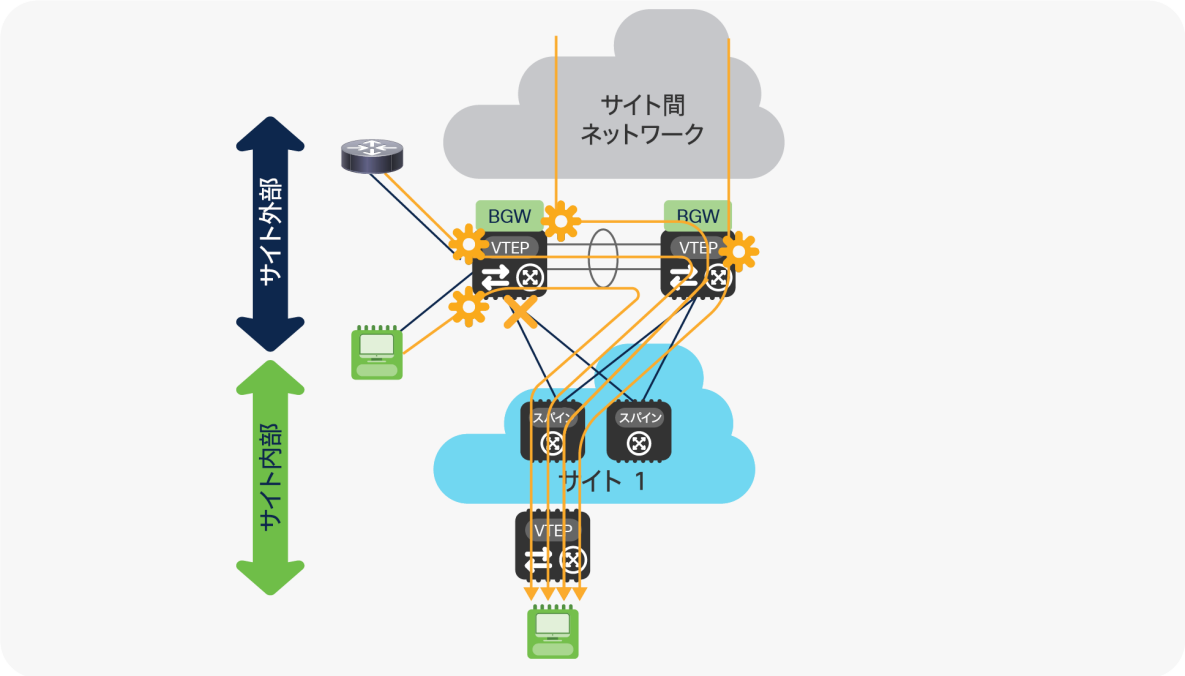

図 30.
分断された BGW ノードでの PIP、vPC VIP、マルチサイト VIP の使用

前述のように、トラフィックフローが増加した場合を考慮して、vPC ピアリンク接続で使用できる帯域幅を適切に設定することが重要です。

サイト内部ネットワークとの最初のリンクが回復すると、BGW ノードはスパインとのアンダーレイ接続を再確立し、ピアリンクを使用せずに、より適切な方法でトラフィックの送受信を開始します。

「ジグザグ」分断のシナリオ

図 31 のような具体的な「ジグザグ」分断シナリオについても、前の 2 つのセクションで説明した動作により対処できます。このシナリオでは複数の vPC BGW ノードで同時に障害が発生し、サイト内部およびサイト外部ネットワークからノードが分断されます。

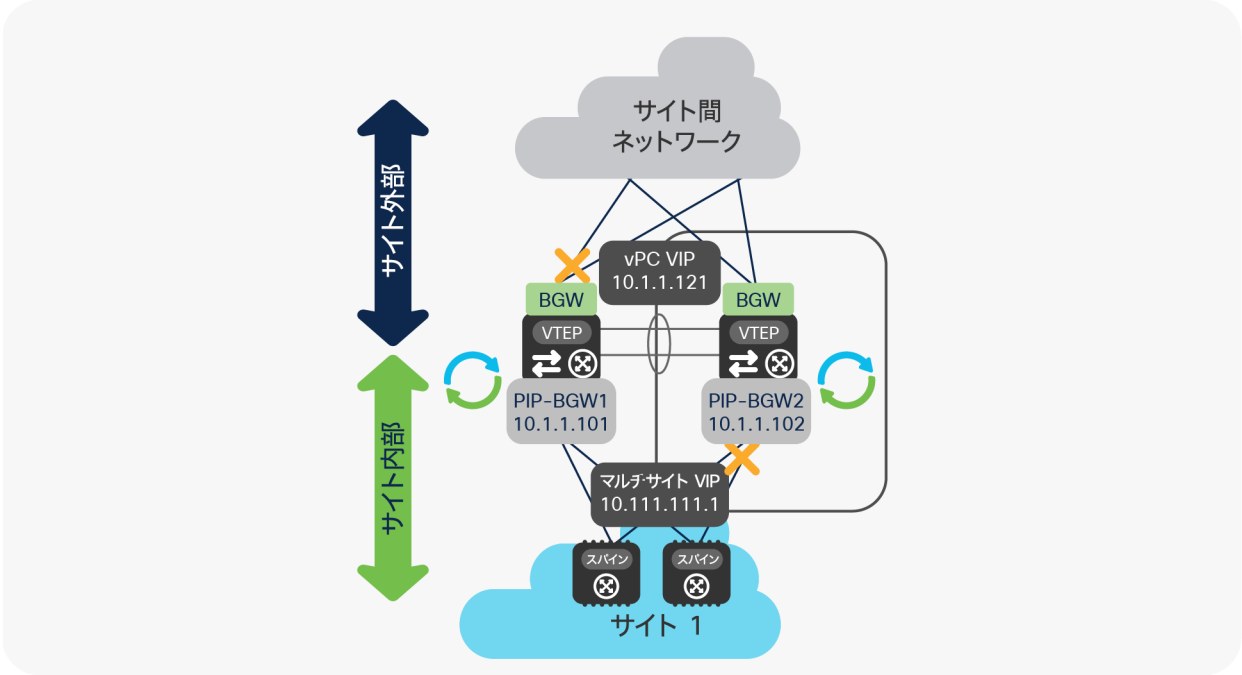

図 31.
「ジグザグ」分断のシナリオ

サイト内部ネットワークから分断された BGW ノードのマルチサイト VIP インターフェイスがアクティブ状態から変化しないことで、リモートサイトから発信された着信トラフィックがピアリンク接続を介してローカルサイトに転送されます（逆も同様）。この状態を表したのが図 32 です。

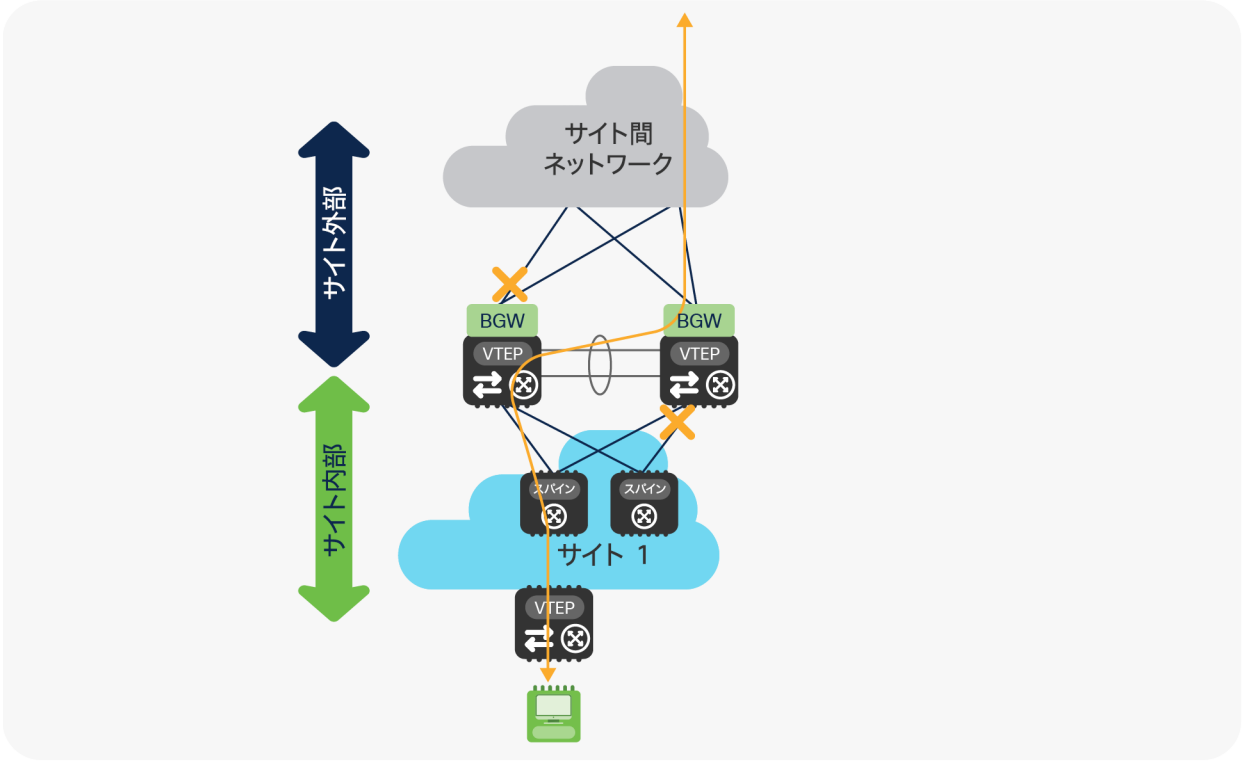

図 32.
「ジグザグ」トラフィックの動き

vPC ピアリンクは、リモートサイトを発信元とし、なおかつ BGW ノード（上の図の左側。サイト外部ネットワークから分断された状態）にローカル接続されたエンドポイントとネットワークを宛先とする着信トラフィックを転送する際にも利用されます。このことから、一般的な考慮事項としてピアリンクを適切に設定する必要があります。

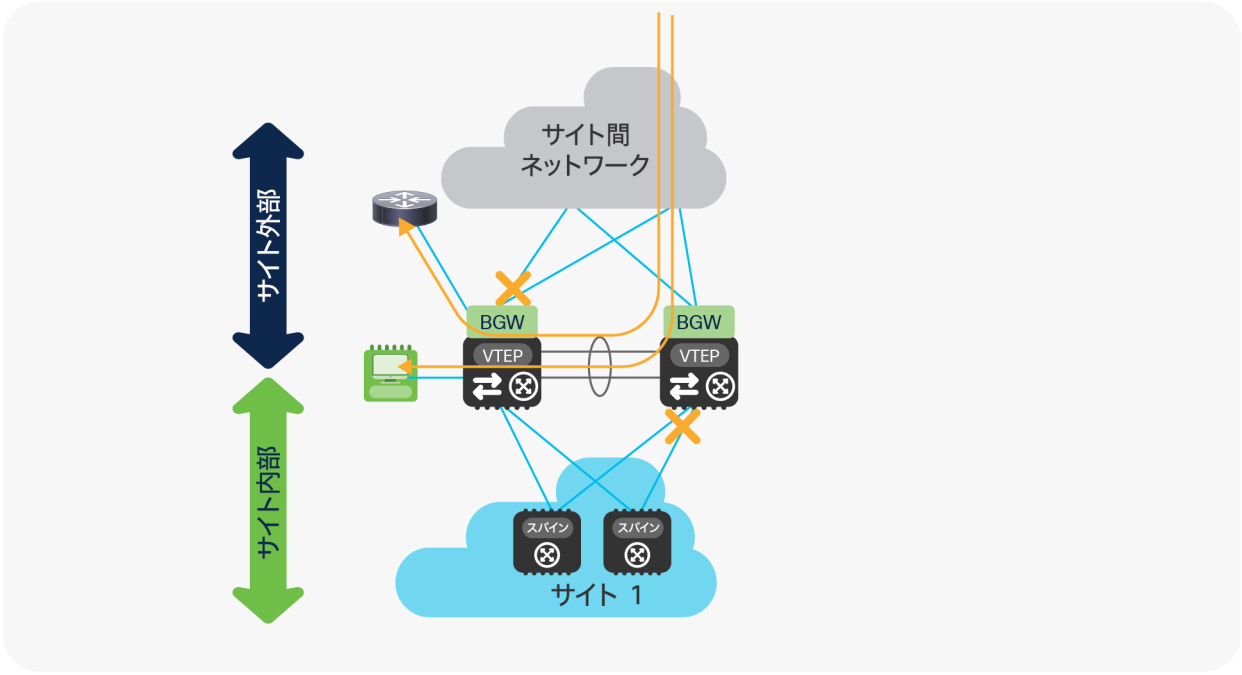

図 33.
BGW ノードにローカル接続されたエンティティ宛ての着信トラフィック

この資料の記載内容は 2021 年 9 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

C11-742114-01JA 21.09