Livre blanc sur la conception et le déploiement multisites VXLAN EVPN

Actualisé: 8 février 2022

Ce que vous apprendrez

Ce document décrit comment réaliser une conception multisite de réseau privé virtuel Ethernet (EVPN) Virtual Private Network (VXLAN) virtual lan (VXLAN) en intégrant les fabrics VXLAN EVPN à l'architecture multisite EVPN pour une extension transparente de couche 2 et de couche 3. Outre les détails techniques, ce document présente des considérations de conception et des exemples de configurations pour illustrer l'approche multisite EVPN. La structure (ou le site) EVPN du protocole BGP (Border Gateway Protocol) VXLAN peut être étendue aux couches 2 et 3 avec diverses technologies. Cependant, le seul objectif de ce document est de savoir comment cette extension peut être réalisée en utilisant l'architecture EVPN Multi-Site, une approche d'interconnectivité intégrée pour les fabrics VXLAN BGP EVPN.

La technologie EVPN Multi-Site est basée sur IETF draft-sharma-multi-site-evpn.

L'architecture VXLAN EVPN Multi-Site est indépendante du réseau de transport entre les sites. Néanmoins, ce document fournit les meilleures pratiques et des recommandations pour un déploiement réussi.

Conditions préalables

Ce document suppose que le lecteur est familiarisé avec la configuration de la structure de centre de données VXLAN BGP EVPN (réseau interne au site). La structure VXLAN BGP EVPN peut être configurée manuellement ou à l'aide de Cisco Data Center Network Manager (DCNM).®

Ce document se concentre entièrement sur les considérations de conception, de déploiement et de configuration pour l'architecture multisite EVPN et les passerelles frontalières (BFW) associées. Il suppose

que les structures de centre de données individuelles (réseaux internes au site) sont déjà configurées et opérationnelles. La solution EVPN Multi-Site vous permet d'interconnecter des structures de centre de données basées sur la technologie VXLAN EVPN. Il vous permet également d'étendre la connectivité des couches 2 et 3 aux réseaux de centres de données construits avec des technologies plus anciennes (héritées) (Spanning Tree Protocol, Virtual Port Channel [vPC], Cisco FabricPath, etc.).

Introduction

Cette section présente un bref aperçu de la technologie sous-jacente à l'architecture multisite VXLAN EVPN. Il présente également plusieurs cas d'utilisation.

Réseaux hiérarchiques

Pendant des décennies, les organisations ont construit des réseaux hiérarchiques, soit en construisant et en interconnectant plusieurs domaines réseau, soit en utilisant simplement des mécanismes d'adressage hiérarchique tels que le protocole Internet (IP). Avec la présence de la couche 2 et de l'espace d'adressage non hiérarchique, les grands domaines pontés ont toujours présenté un défi pour la mise à l'échelle et l'isolation des défaillances. Aujourd'hui, avec l'essor de la mobilité des terminaux, des technologies permettant de créer des extensions de couche 2 plus efficaces et de rétablir les hiérarchies sont nécessaires. En utilisant une interconnectivité dédiée qui peut ramener la hiérarchie perdue, les technologies d'interconnexion de centre de données (DCI) ont été populaires. Cependant, bien que DCI puisse être utilisé pour interconnecter plusieurs centres de données, au sein du centre de données, de grandes structures sont devenues courantes pour faciliter le placement et la mobilité des points de terminaison sans frontières. À la suite de cette tendance, l'explosion de l'état du réseau pour les entrées MAC et ARP s'est présentée. VXLAN était censé relever ce défi, mais il a augmenté le défi, avec des domaines de couche 2 encore plus grands en cours de construction car la limite d'emplacement a été surmontée par la capacité de VXLAN à fournir un réseau de couche 2 sur couche 3.

Pour les tissus, la colonne vertébrale et la feuille, l'arbre gras et les topologies Clos pliées sont devenus essentiellement les topologies standard. Les nouveaux modèles de topologie de réseau construisent des réseaux hiérarchiques bien conçus, mais avec l'ajout de VXLAN en tant que réseau over-the-top, cette hiérarchie était en train d'être aplatie. Alors que la conception du réseau dans la topologie sous-jacente était principalement de couche 3 et qu'une hiérarchie efficace était présente, avec l'introduction du réseau de superposition, cette hiérarchie est devenue cachée. Cet aplatissement présente à la fois des avantages et des inconvénients. L'approche consistant à construire un réseau par-dessus le dessus sans toucher chaque commutateur offre une simplicité, et un tel réseau peut être étendu à plusieurs endroits. Cependant, cette approche présente un risque en l'absence d'isolation des défaillances, en particulier lorsque des réseaux de couche 2 volumineux et étirés sont construits avec cette nouvelle conception de réseau superposée. Tout ce qui est envoyé à travers le point d'entrée dans le réseau de superposition partira au point de sortie respectif. Ces réseaux superposés utilisent l'approche « la plus proche de la source » et « la plus proche de la destination » et construisent dynamiquement des tunnels d'un point à l'autre partout où cela est nécessaire.

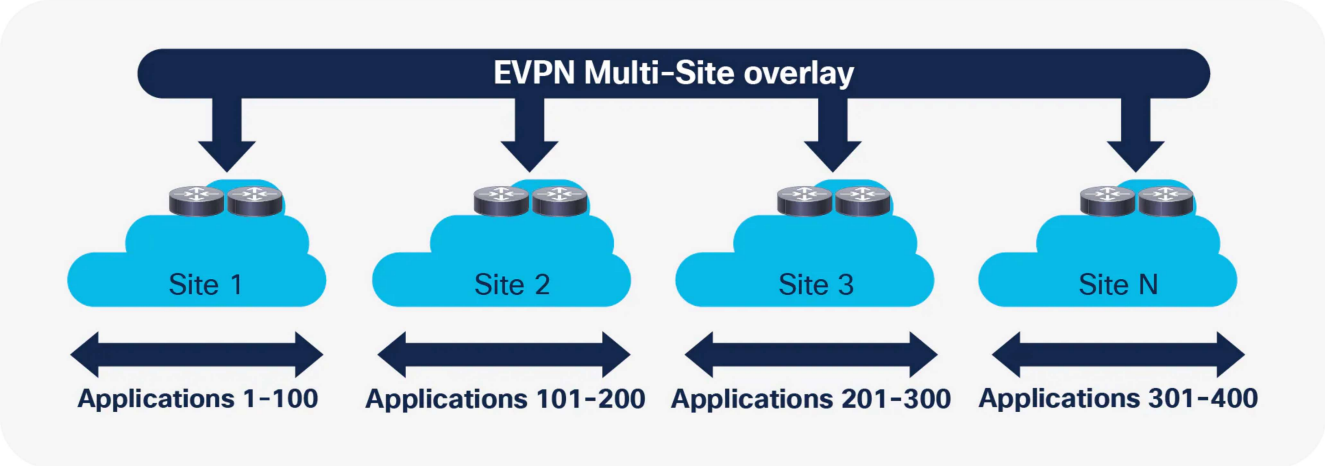
L'architecture multisite EVPN ramène des hiérarchies pour superposer les réseaux. L'architecture EVPN Multi-Site introduit le BGP externe (eBGP) pour les réseaux VXLAN BGP EVPN, alors que jusqu'à présent le BGP intérieur (iBGP) était prédominant. Suite à l'introduction du comportement eBGP next-hop, des systèmes autonomes (AS) aux passerelles frontalières (BGW) ont été introduits, renvoyant les points de

contrôle du réseau au réseau de superposition. Avec cette approche, les hiérarchies sont utilisées efficacement pour compartimenter et interconnecter plusieurs réseaux superposés. Les organisations disposent également d'un point de contrôle pour diriger et appliquer l'extension du réseau à l'intérieur et au-delà d'un seul centre de données.

Cas d'utilisation

L'architecture VXLAN EVPN Multi-Site est une conception pour les réseaux de superposition basés sur VXLAN BGP EVPN. Il permet l'interconnexion de plusieurs fabrics VXLAN BGP EVPN distincts ou de domaines de superposition, et il permet de nouvelles approches de la mise à l'échelle, du compartimentage et de l'ICD du fabric.

Lorsque vous créez une structure de centre de données de grande taille par emplacement, divers défis liés au fonctionnement et au confinement des défaillances existent. En construisant des compartiments de tissus plus petits, vous améliorez les domaines de défaillance et d'opération individuels. Néanmoins, la complexité de l'interconnexion de ces différents compartiments empêche le déploiement généralisé de tels concepts, en particulier lorsque l'extension des couches 2 et 3 est requise (figure 1).



Graphique 1.

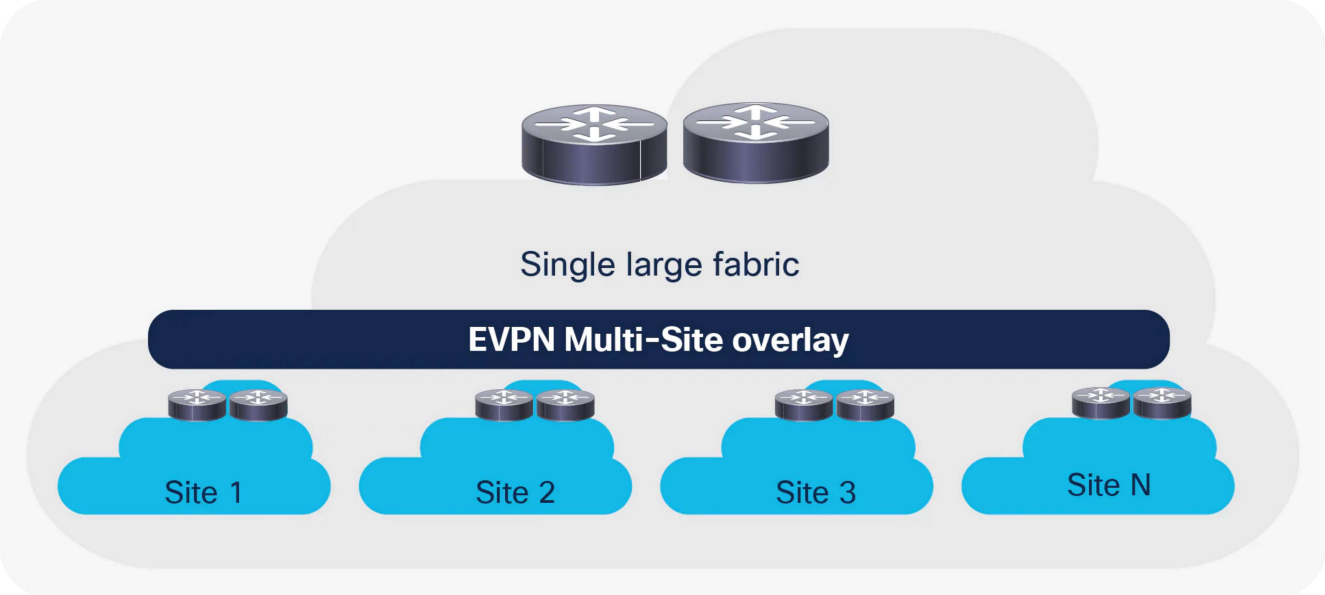
Exemple de compartimentation

L'architecture multisite VXLAN EVPN offre une interconnectivité intégrée qui ne nécessite pas de technologie supplémentaire pour l'extension de couche 2 et de couche 3. Il offre ainsi la possibilité d'une extension sans couture entre les compartiments et les tissus. Il vous permet également de contrôler ce qui peut être étendu. En plus de définir quelle instance VLAN ou VRF (Virtual Routing and Forwarding) est étendue, dans les extensions de couche 2, vous pouvez également contrôler le trafic de diffusion, de monodiffusion inconnue et de multidiffusion (BUM) pour limiter l'effet d'entraînement d'une défaillance dans une structure de centre de données.

Lorsque vous créez des réseaux à l'aide du modèle de mise à l'échelle, un périphérique ou un composant atteint généralement la limite d'échelle avant que le réseau global ne le fasse. L'approche scale-out offre une amélioration pour les structures de centre de données. Néanmoins, une structure de centre de données unique a également des limites d'échelle, et donc l'approche de montée en puissance parallèle pour une seule structure de grand centre de données existe.

En plus de la possibilité d'évoluer au sein d'un seul fabric, avec l'architecture EVPN Multi-Site, vous pouvez évoluer au niveau suivant de la hiérarchie. De même, lorsque vous ajoutez plus de nœuds feuilles

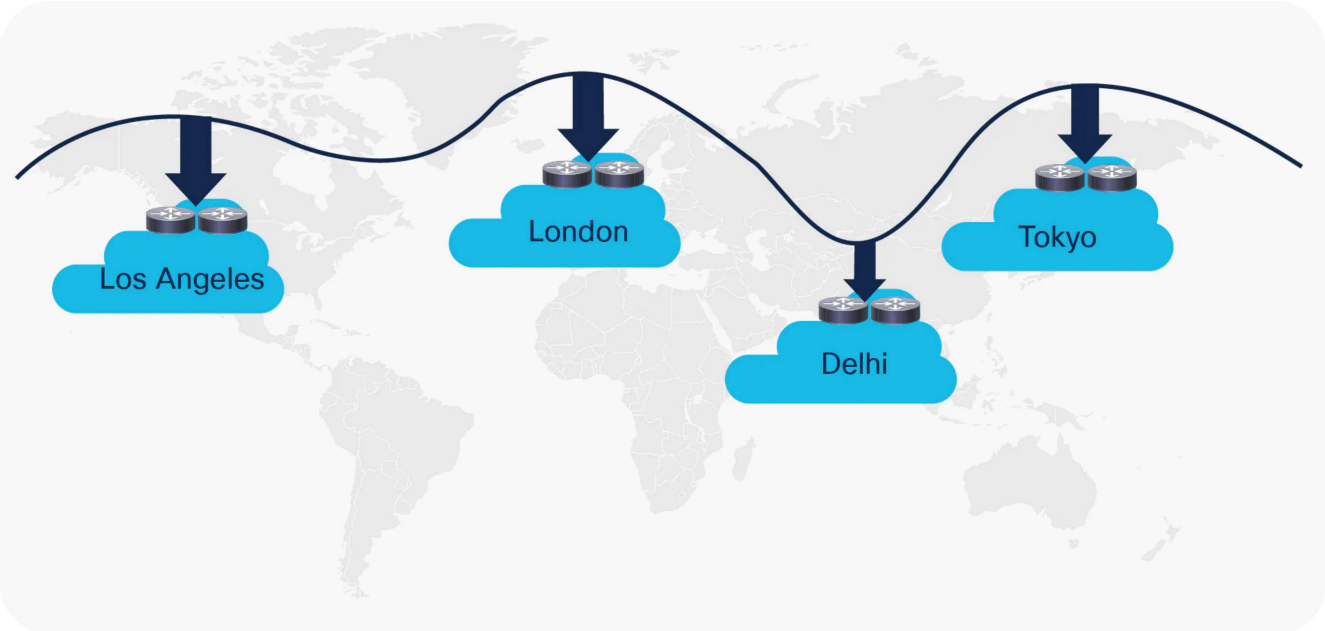
pour la capacité au sein d'une structure de centre de données, dans l'architecture multisite EVPN, vous pouvez ajouter des fabrics (sites) pour mettre à l'échelle horizontalement l'environnement global. Avec cette approche évolutive dans l'architecture multisite EVPN, en plus d'augmenter l'échelle, vous pouvez contenir les contiguïtés de maillage complet du VXLAN entre les points de terminaison de tunnel VXLAN (VTEPs) dans un fabric (Figure 2).



Graphique 2.

Exemple de mise à l'échelle

L'architecture multisite EVPN peut également être utilisée pour les scénarios DCI (Figure 3). Comme pour le compartimentage et la montée en puissance parallèle au sein d'un centre de données, l'architecture multi-sites EVPN a été conçue pour DCI. L'architecture globale permet de positionner et d'interconnecter un ou plusieurs sites par centre de données avec un ou plusieurs sites dans un centre de données distant. Avec l'extension transparente et contrôlée des couches 2 et 3 grâce à l'utilisation de VXLAN BGP EVPN au sein et entre les sites, les capacités de VXLAN BGP EVPN lui-même ont été augmentées. Les nouvelles fonctions liées au contrôle du réseau, au masquage VTEP et à l'application du trafic BUM ne sont que quelques-unes des fonctionnalités qui contribuent à faire de l'architecture multisite EVPN la technologie DCI la plus efficace.



Graphique 3.

Exemple d'interconnexion de centre de données

Exigences

Le tableau 1 résume les exigences relatives à l'architecture multisite EVPN. Le tableau 1 fournit la configuration matérielle et logicielle requise pour les commutateurs Cisco Nexus série 9000 qui fournissent la fonction EVPN Multi-Site BGW.®

Tableau 1. Configuration logicielle et matérielle minimale requise EVPN Multi-Site border gateway

Article	Exigence
Matériel Cisco Nexus	• Plate-forme Cisco Nexus 9300 EX • Plate-forme Cisco Nexus 9300 FX • Plate-forme Cisco Nexus 9300 FX2 • Plate-forme Cisco Nexus 9300 FX3 • Plate-forme Cisco Nexus 9300-GX * • Plate-forme Cisco Nexus 9332C • Plate-forme Cisco Nexus 9364C • Plate-forme Cisco Nexus 9500 avec carte de ligne X9700-EX • Plate-forme Cisco Nexus 9500 avec carte de ligne X9700-FX • Plate-forme Cisco Nexus 9500 avec carte de ligne X9700-GX
Logiciel Cisco NX-OS	Logiciel Cisco NX-OS version 7.0(3)I7(1) ou ultérieure

***La plate-forme est capable d'exécuter la fonction Multi-Site Border Gateway (BGW), veuillez consulter les notes de mise à jour pour le support logiciel.**

Remarque : La configuration matérielle et logicielle requise pour le réflecteur de route BGP interne au site (RR) et le VTEP d'un site VXLAN BGP EVPN reste la même que celles sans le BGW multisite EVPN. Ce document ne couvre pas la configuration matérielle et logicielle requise pour le réseau interne du site VXLAN EVPN. La section « Pour plus d'informations » à la fin de ce document comprend des liens qui permettent d'accéder aux sites Web Cisco spécifiques aux déploiements VXLAN BGP EVPN.

D'autres considérations de conception pour le matériel et les logiciels internes au site et externes au site sont abordées dans les sections suivantes.

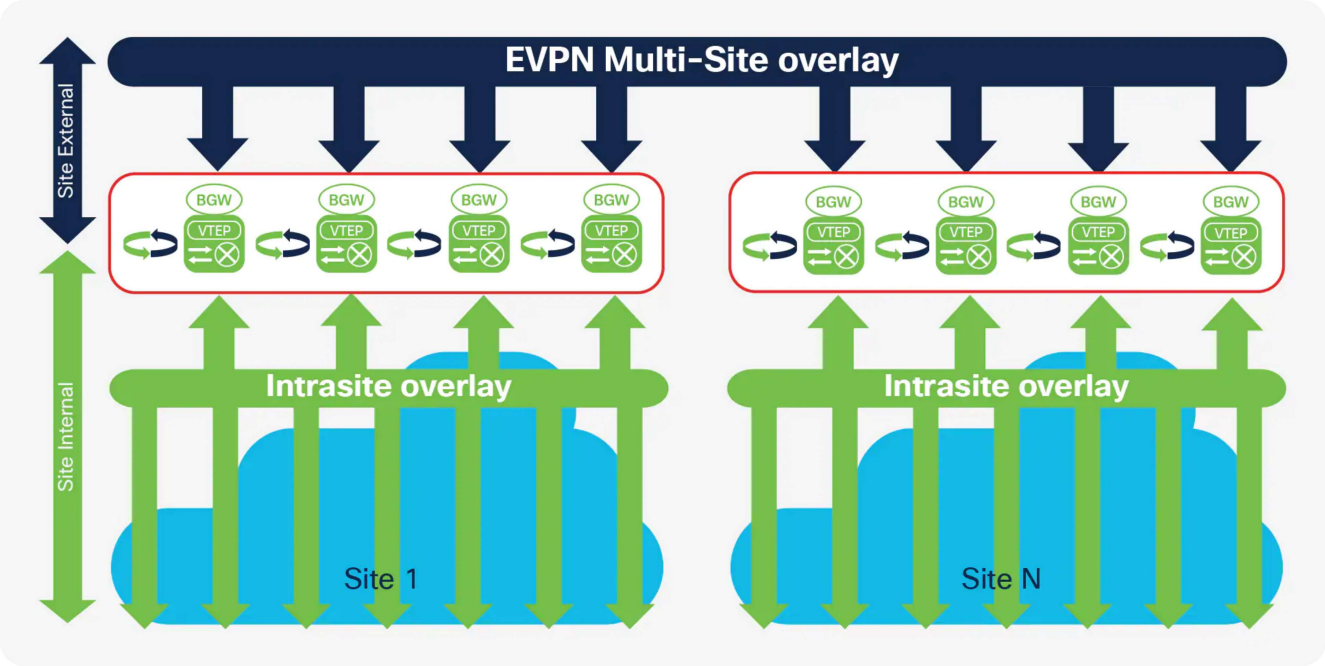
Détails de la technologie

Cette section présente des informations techniques sur les principaux composants de l'architecture multisite EVPN et décrit les scénarios de défaillance.

Passerelle frontalière

Le principal composant fonctionnel de l'architecture multisite EVPN est la passerelle frontalière, ou BGW. Les BFW séparent le fabric-side (fabric interne au site) du réseau qui interconnecte les sites (DCI externe au site) et masquent les VTEP internes au site.

Généralement, un déploiement EVPN Multi-Site se compose de deux sites ou plus, qui sont interconnectés via une superposition VXLAN BGP EVPN Layer 2 et Layer 3 (Figure 4). Dans ce scénario, le BGW est connecté aux VTEP internes au site (généralement via des nœuds de colonne vertébrale) et à un réseau de transport externe au site qui permet au trafic d'atteindre les BFW sur d'autres sites distants. Les BFW sur les sites distants ont derrière eux des VTEP internes au site. Seules les adresses IP sous-jacentes des BFW sont visibles à l'intérieur du réseau de transport entre les BFW. Les VTEP internes au site sont toujours masqués derrière les BGW.



Graphique 4.

Déploiement multisite EVPN

Du point de vue de BGW, le rôle des VTEP internes au site est de partager les fonctions VXLAN et BGP-EVPN communes. Pour interagir avec un BGW, un nœud interne au site doit prendre en charge les fonctions suivantes :

- VXLAN avec multidiffusion indépendante du protocole (PIM), multidiffusion toute source (ASM) ou réplication d'entrée (BGP EVPN Route Type 3) dans la sous-couche
- BGP EVPN Route Type 2 et Route Type 5 pour le plan de contrôle de superposition
- Réflecteur de route capable d'échanger BGP EVPN Route Type 4
- Périphériques compatibles avec les opérations, l'administration et la maintenance VXLAN (OAM) pour la prise en charge OAM de bout en bout

Du point de vue du réseau externe du site, aucune exigence spécifique n'est exigée, à l'exception de l'accessibilité du transport IP entre les BFW et de l'adaptation d'une taille de paquet mTU (Maximum Transmission Unit) accrue. Les BGV utilisent toujours la réplication d'entrée (IR) pour le trafic BUM de couche 2 entre les BGV de différents sites, mais ils peuvent utiliser PIM ASM ou la réplication d'entrée au sein d'un site donné. Cette fonctionnalité offre une flexibilité pour les déploiements existants et une indépendance de transport pour le réseau externe du site.

Remarque : L'architecture multisite EVPN utilise l'encapsulation VXLAN pour le plan de données, ce qui nécessite 50 ou 54 octets de surcharge en plus du MTU Ethernet standard (1550 ou 1554).

Le BGW effectue localement la procédure de séparation de site interne-externe. Par conséquent, le BGW ne nécessite pas d'appareil voisin pour exécuter cette fonction. Tout comme un VTEP traditionnel peut se connecter d'un réseau interne au site à un BGW, un VTEP traditionnel peut également se connecter à un BGW à partir d'un réseau externe au site. Autrement dit, un BGW sur le site source ne nécessite pas de BGW voisin sur le site de destination ; un VTEP traditionnel suffira. Cette flexibilité intégrée au BGW permet des déploiements au-delà des appariements multi-sites EVPN traditionnels. L'un de ces cas de déploiement est décrit dans la section « Frontière partagée » de ce document, et l'autre est décrit dans la section « Intégration de sites hérités ».

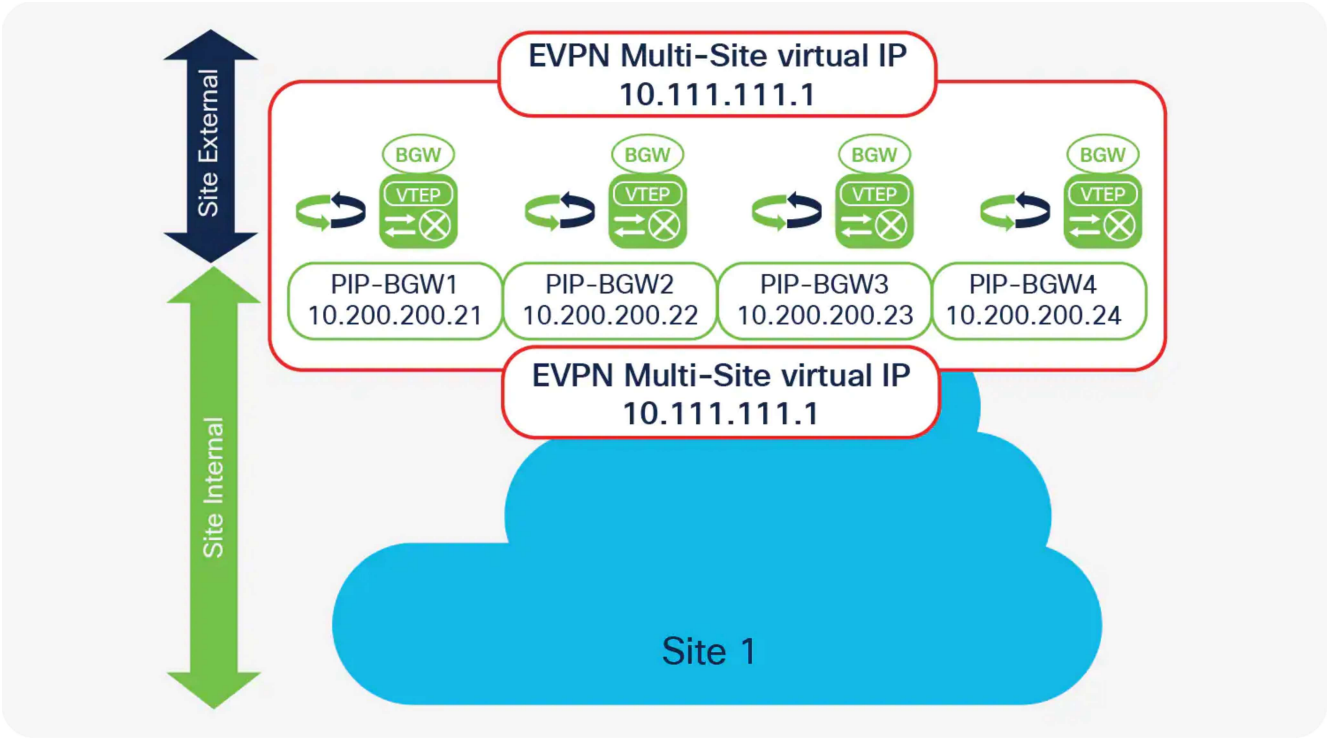
Notez que même si un VTEP traditionnel fonctionnerait pour se connecter à un BGW à partir d'un réseau externe au site, ces VTEP connectés en externe n'exécuteraient aucune fonction BGW étendue telle que le masquage VTEP interne au site.

Placement de la passerelle frontalière

Avec l'architecture EVPN Multi-Site, deux emplacements de placement peuvent être envisagés pour le BGW. Un ensemble dédié de BFW peut être placé au niveau de la couche foliaire, avec les BFW connectés à la colonne vertébrale comme n'importe quel autre VTEP dans le tissu (VTEP interne au site). Alternativement, les BGW peuvent être co-localisés sur la colonne vertébrale du tissu. Si le BGW est sur la colonne vertébrale, de nombreuses fonctions sont surchargées ensemble: par exemple, réflecteur d'itinéraire, point de rendez-vous (RP), trafic est-ouest et fonctions de connectivité externe. Dans ce cas, vous devez prendre en compte des facteurs supplémentaires liés à l'échelle, à la configuration et aux scénarios de défaillance.

Passerelle de frontière Anycast

L'anycast BGW (A-BGW) exécute la fonction BGW comme décrit dans la section précédente. L'A-BGW permet la mise à l'échelle horizontale des BFW dans un modèle scale-out et sans partage du destin des dépendances inter-périphériques. Depuis Cisco NX-OS 7.0(3)I7(1), l'A-BGW est disponible sur les plates-formes à l'échelle du cloud Cisco Nexus série 9000 (plates-formes Cisco Nexus 9000 Series EX et FX), avec jusqu'à quatre BFW anycast disponibles par site (Figure 5).



Graphique 5.

Passerelle frontalière Anycast

Le nom « A-BGW » fait référence au partage d'une adresse IP virtuelle (VIP) commune ou d'une adresse IP anycast entre les BGW dans un site commun. Ce document utilise l'adresse IP virtuelle pour faire référence également à l'adresse IP anycast multisite EVPN.

L'adresse IP virtuelle sur le BGW est utilisée pour toutes les communications du plan de données quittant le site et entre les sites lorsque l'extension EVPN Multi-Site est utilisée pour atteindre un site distant. L'adresse IP virtuelle unique est utilisée à la fois dans le site pour atteindre un point de sortie et entre les sites, les BFW utilisant toujours l'adresse IP virtuelle pour communiquer entre eux. L'adresse IP virtuelle est représentée par une interface de bouclage dédiée associée à l'interface NVE (Network Virtualization Endpoint) (**loopback d'interface frontière-passerelle multisite100**).

Avec cette approche, et avec l'existence d'un réseau ECMP (Equal-Cost Multipath), tous les BGV sont toujours également accessibles et actifs pour le transfert du trafic de données. Le réseau de transport sous-jacent à l'intérieur ou entre les sites est responsable du hachage du trafic VXLAN parmi les chemins à coût égal disponibles. Cette approche évite la polarisation, compte tenu de l'entropie du VXLAN, et augmente la résilience. En cas d'échec d'un ou de plusieurs BGV, les BFW restants annoncent toujours l'adresse IP virtuelle et sont donc immédiatement disponibles pour prendre en charge tout le trafic de données. L'utilisation d'adresses IP anycast ou d'adresses IP virtuelles fournit une résilience basée sur le réseau, au lieu d'une résilience qui repose sur des hellos de périphérique ou des protocoles d'état similaires.

En plus de l'adresse IP virtuelle ou de l'adresse IP anycast, chaque BGW a sa propre personnalité individuelle représentée par l'adresse IP VTEP principale (PIP) (bouclage source-interface1). L'adresse PIP est responsable dans le BGW de la gestion du trafic BUM. Chaque BGW utilise son adresse PIP pour effectuer la réplication BUM, soit dans la sous-couche de multidiffusion, soit lors de la publicité de BGP EVPN Route Type 3 (multidiffusion incluse), utilisée pour la réplication d'entrée. Par conséquent, chaque BGW joue un rôle actif dans l'expédition BUM. Comme l'adresse IP virtuelle, l'adresse PIP est annoncée

au réseau interne du site ainsi qu'au réseau externe du site. L'adresse PIP est utilisée pour gérer le trafic BUM entre les BFW sur différents sites, car l'architecture EVPN Multi-Site utilise toujours la réplication d'entrée pour ce processus.

L'adresse PIP est également utilisée dans deux autres scénarios étroitement liés.

Si le BGW fournit une connectivité externe avec VRF-lite à côté du déploiement EVPN Multi-Site, les préfixes de routage qui sont appris à partir des périphériques externes de couche 3 sont annoncés à l'intérieur de la structure VXLAN avec l'adresse PIP comme adresse du tronçon suivant. Du point de vue du BGW, ces préfixes IP appris en externe sont considérés comme provenant localement d'un BGW, en utilisant la famille d'adresses BGP EVPN. Ce processus crée un BGP EVPN Route Type 5 individuel (itinéraire de préfixe IP) à partir de chaque BGW qui a appris un préfixe IP pertinent en externe. Dans le meilleur des cas, votre réseau interne au site dispose d'une route ECMP pour atteindre les réseaux multi-sites non-EVPN.

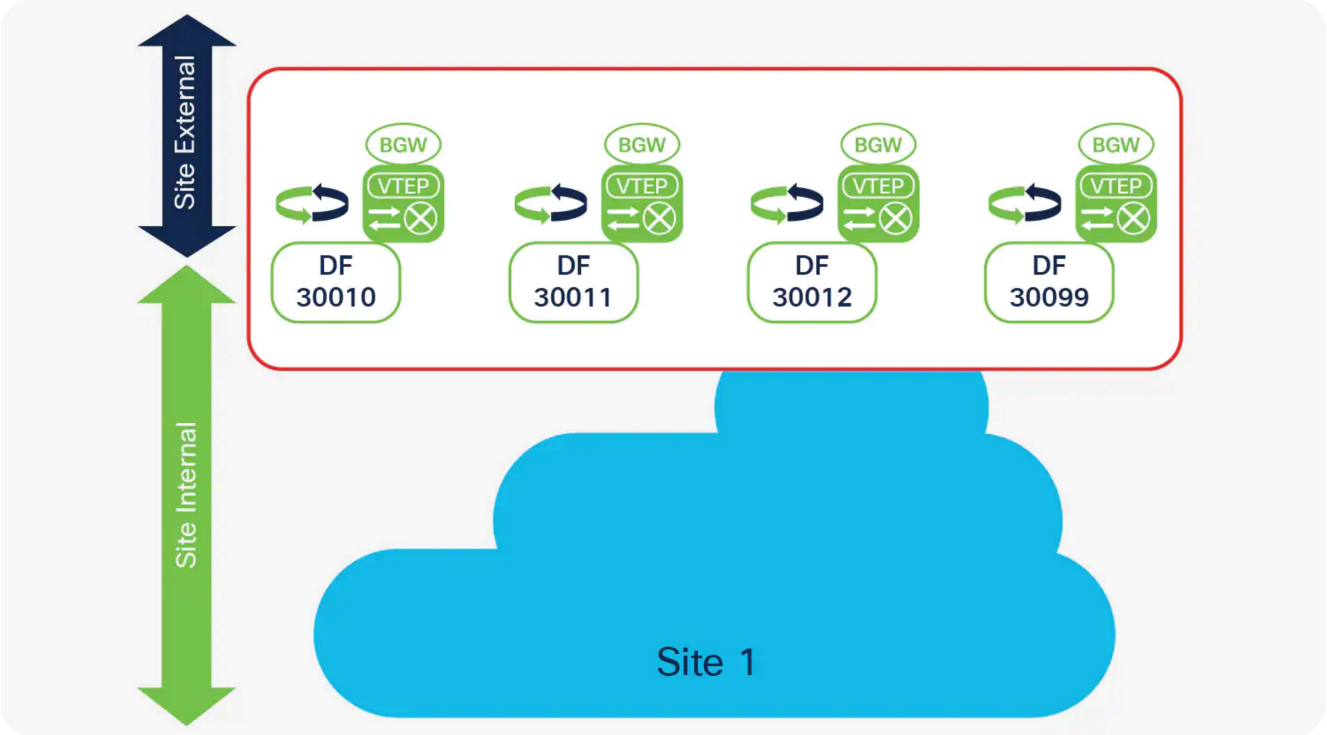
Remarque : Les préfixes IP appris externes peuvent être redistribués à BGP EVPN à partir de n'importe quel monodiffusion BGP IPv4/IPv6, OSPF (Open Shortest Path First) ou tout autre protocole de routage statique ou dynamique qui permet la redistribution vers BGP EVPN.

Un scénario étroitement lié est le cas dans lequel le BGW annonce un préfixe IP avec sa propre adresse PIP via une connectivité locale. Un point de terminaison peut être directement connecté à un BGW, mais son adresse IP ne peut être apprise que par routage sur une interface physique ou une sous-interface. La connexion entièrement active des services réseau de couche 4 à 7 (L4-L7) (par exemple, les pare-feu et les équilibreurs de charge) peut être réalisée via le routage ECMP avec un protocole de routage statique ou dynamique.

Remarque : L'utilisation de VLAN et d'interfaces virtuelles de commutation (SVU) locales à un BGW ou sur plusieurs BFW n'est actuellement pas prise en charge. Cette restriction s'applique également aux canaux de port de couche 2 avec ou sans multihébergement. Pour les services réseau L4-L7 qui nécessitent ce modèle de connectivité, utilisez un VTEP interne au site (un VTEP traditionnel).

Transitaire désigné

Chaque A-BGW participe activement à l'acheminement du trafic BUM. Plus précisément, la fonction designated-Forwarder (DF) pour le trafic BUM est distribuée sur une base d'identificateur de réseau VXLAN (VNI) par couche 2. Pour synchroniser les redirecteurs désignés, les mises à jour BGP EVPN Route Type 4 (Ethernet segment route) sont échangées entre les BFW au sein d'un même site (Figure 6).



Graphique 6.

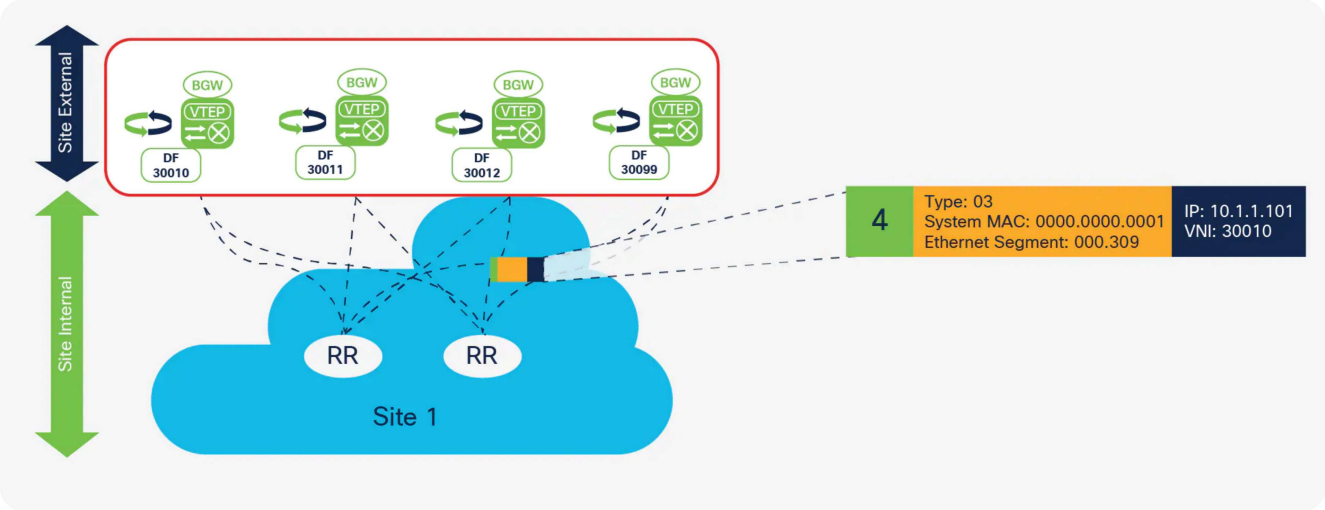
Transitaire désigné

Pour participer à l'élection du transitaire désigné, la configuration du même ID de site est requise. Cet ID est défini dans le cadre de la configuration BGW (**evpn multisite border-gateway <site-id>**). En plus de l'ID de site, l'utilisation du même VNI de couche 2 est nécessaire pour élire le transitaire désigné parmi les BGW éligibles.

L'affectation du transitaire désigné est effectuée sur une base VNI par couche 2, à l'aide d'un processus de tourniquet pour répartir les affectations de manière égale. Une liste ordinale d'adresses PIP est utilisée et, en fonction de tous les ordres de configuration VNI de couche 2 ou de la liste ordinale, le rôle de redirecteur désigné est distribué de manière circulaire.

Remarque : Chaque BGW aura un rôle de transitaire désigné actif si le nombre de VNI de couche 2 dépasse le nombre de BGW.

Pour échanger les messages d'élection du transitaire désigné entre les BFW, l'appairage BGP EVPN est requis car les messages d'élection sont constitués de publicités BGP EVPN Route Type 4. Plus naturellement, le BGW s'accorderait avec un réflecteur de route interne au site (fabric), qui contient également toutes les informations de point de terminaison provenant des VTEP internes au site. Avec le réflecteur d'itinéraire déjà présent dans le tissu, et avec tous les VTEP, y compris le BGW, qui l'appairent, l'échange de messages électoraux de transitaire désigné est réalisé (Figure 7).



Graphique 7.

Élection du transitaire désigné à l'aide de réflecteurs d'itinéraire

Dans les cas où il n'existe pas de réflecteur de route ou dans lesquels le réflecteur de route n'est pas capable de relayer BGP EVPN Route Type 4, une session iBGP peut être considérée comme une alternative. L'appairage iBGP doit être compatible avec la famille d'adresses EVPN et avoir un maillage complet établi entre les interfaces de bouclage des BFW.

Remarque: L'échange BGP EVPN Route Type 4 ne doit se produire que par le biais d'appairages internes au site. Si l'échange d'élection du transitaire désigné se produit via les réseaux interne au site (fabric) et externe au site (DCI), un temps de convergence prolongé peut être expérimenté dans certains scénarios de défaillance. Par défaut, cet appairage est appliqué via le mécanisme de prévention de la boucle de chemin d'accès système autonome BGP, car les systèmes autonomes source et de destination pour les BFW locaux du site sont les mêmes. Dans les cas où des fonctions telles **que as-override** et **allowas-in sont utilisées**, vous devez accorder une attention particulière à l'appairage de superposition externe du site.

Scénarios de défaillance

Le BGW est le dispositif de liaison entre les VTEP internes au site et tout ce qui est externe au site. En raison de l'importance du BGW, vous devez tenir compte non seulement de l'échelle et de la résilience, mais également du comportement lors d'une situation de défaillance. Pour l'architecture multisite EVPN, vous devez envisager deux scénarios de défaillance principaux : une défaillance dans le fabric (défaillance interne au site) et une défaillance dans la zone externe du site. Avec la résilience recommandée pour la conception globale de la connectivité, l'architecture multi-sites EVPN est équipée pour résister aux défaillances qui nécessitaient auparavant un temps de convergence important ou un recalcul du chemin de données.

Isolation du tissu

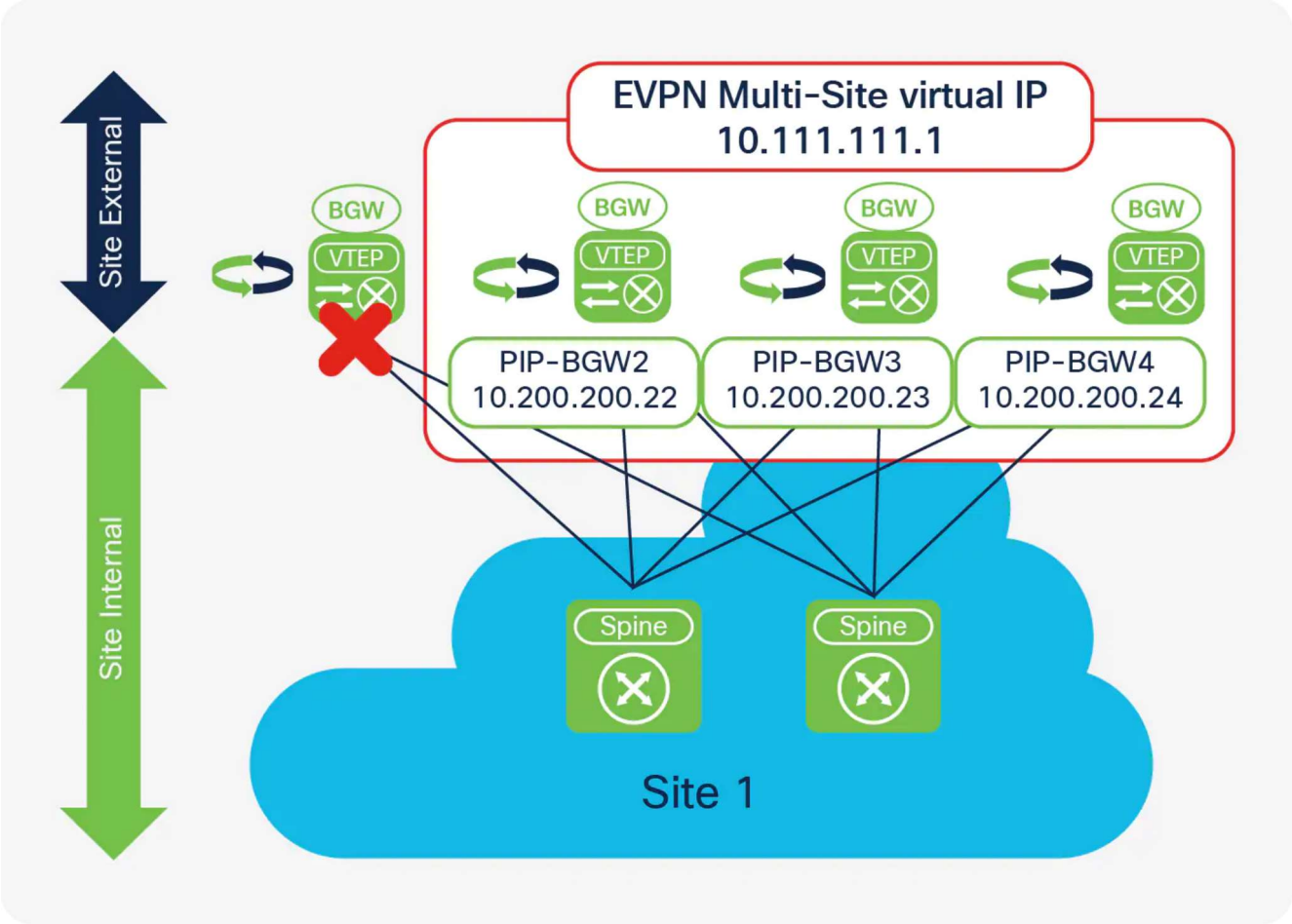


Figure 8.

Fabric Isolation

Failure detection in the site-internal interfaces is one of the main mechanisms offered by EVPN Multi-Site architecture to reduce traffic outages. The site-internal or fabric interfaces commonly are connected to the spine layer, to which more VTEPs are connected. Assuming a fabric with two spine switches and four BGWs, a full mesh of links is established between the neighboring spine and BGW interfaces. On the BGW itself, the site-internal interfaces are specially configured to understand their locations in the network (**evpn multisite fabric-tracking**).

The EVPN Multi-Site fabric-tracking function detects whether one or all of the site-internal interfaces are available. As long as one of these interfaces is operational and available, the BGW can extend Layer 2 and Layer 3 traffic to remote sites. If all fabric-tracking interfaces are reported to be down, the following steps are performed:

- The isolated BGW stops advertising the virtual IP address to the site-external underlay network.
- The isolated BGW withdraws all of its advertised BGP EVPN routes (Route Type 2, Route Type 3, Route Type 4, and Route Type 5).
- The remaining BGWs withdraw all BGP EVPN Route Type 4 (Ethernet segment) routes received from the now isolated BGW because reachability is missing.

Note: You do not need to stop advertising from the site-internal underlay because all site-internal interfaces are considered to be down.

As a result of these actions, the BGW will be isolated from a VTEP perspective in both the site-internal and site-external networks (Figure 8). Therefore, all traffic originating from remote sites and destined for the virtual IP address is rerouted to the remaining BGWs that still host the virtual IP address and have it active. With the disappearance of the BGW traffic to the site-internal network, the advertisements of this PIP address and the capability to participate in designated-forwarder election is removed. As a consequence, the designated-forwarder role for the VNIs previously “owned” by the isolated BGW is now renegotiated between the remaining BGWs.

On recovery from a failure of all site-internal interfaces, first the underlay routing adjacencies are established and then the site-internal BGP sessions to the route reflector are reestablished. To allow the underlay and overlay control planes to converge before data traffic is forwarded by the BGW, you can configure a restore delay for the virtual IP address to delay its advertisement to the underlay network control plane. The EVPN Multi-Site delay-restore setting is a subconfiguration of the BGW site ID configuration (**delay-restore time 300**).

DCI isolation

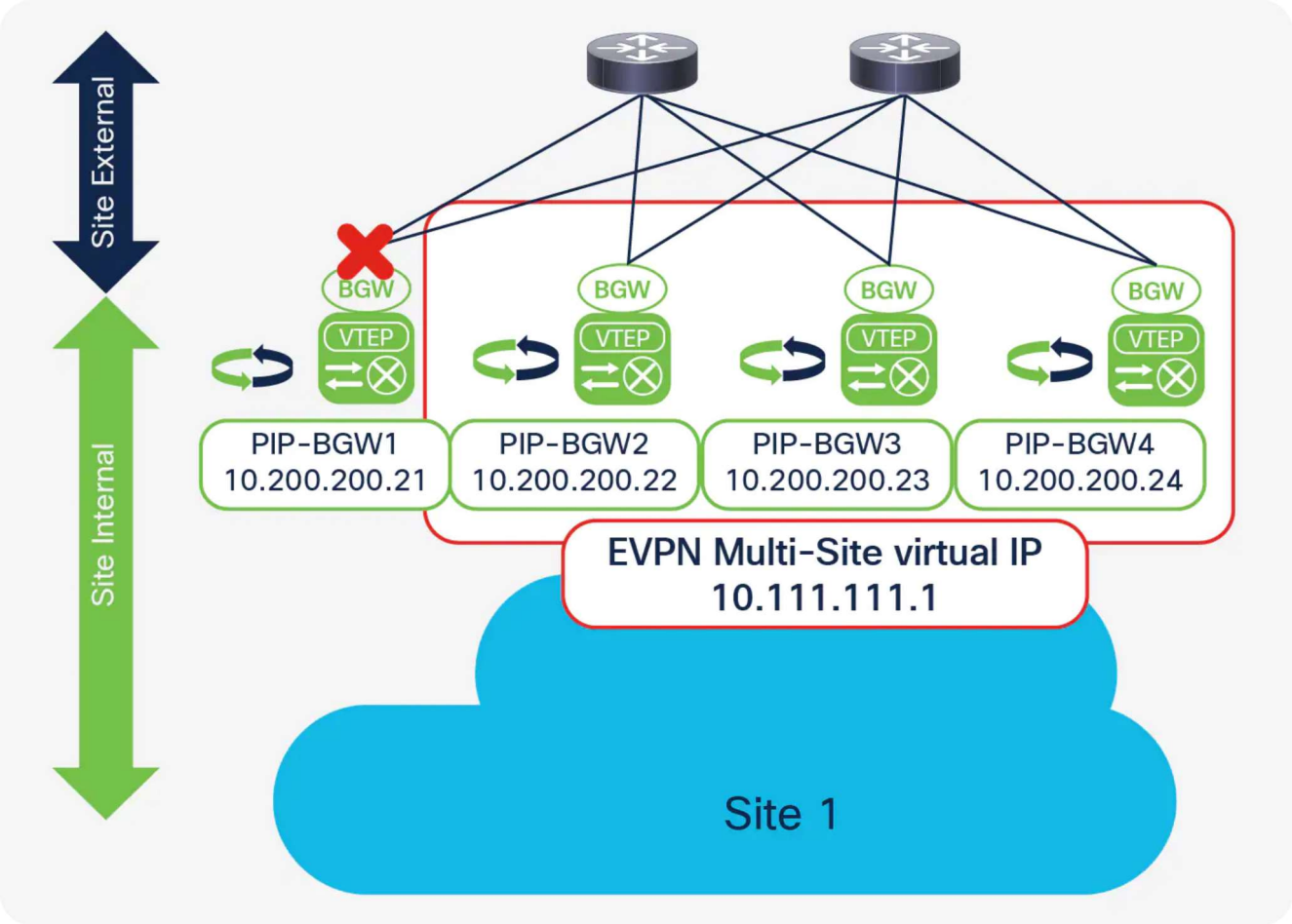


Figure 9.

DCI Isolation

Similar to the site-internal interfaces, the site-external interfaces in EVPN Multi-Site architecture use interface failure detection. The site-external or DCI interfaces commonly are connected to the network between sites, at which more BGWs are present. The site-external interfaces offer a configuration similar to that for the site-internal interfaces to understand their locations and the need for tracking (**evpn multisite dci-tracking**).

The DCI-tracking function in EVPN Multi-Site architecture detects whether one or all of the site-external interfaces are up and operational. If one of the many interfaces remains up, the site-external interfaces are considered working, and the BGW can extend Layer 2 and Layer 3 services to remote sites.

In the rare case in which all DCI-tracking interfaces are down, the BGW performs the following actions:

- It stops advertising the virtual IP address to the site-internal underlay network.
- It withdraws all BGP EVPN Route Type 4 (Ethernet segment) route advertisement.
- It converts the BGW to a traditional VTEP (the PIP address stays up).

Note: You do not need to stop advertising from the site-external underlay because all site-external interfaces are considered to be down.

As a result of these actions, the BGW will continue to operate only as a site-internal VTEP. Therefore, all traffic to the virtual IP address is rerouted to the remaining BGWs that still host the virtual IP address and have it active. The advertisements to participate in designated-forwarder election are removed from the DCI-isolated BGW (Figure 9).

On recovery from a failure of all site-external interfaces, first the underlay routing adjacencies are established, and then the site-external BGP sessions are reestablished. To allow the underlay and overlay control planes to converge before data traffic is forwarded by the BGW, you can configure a restore delay for the virtual IP address. The EVPN Multi-Site delay-restore setting is a subconfiguration of the BGW site ID configuration (**delay-restore time 300**) and applies to both the site-internal and site-external networks.

Design considerations

EVPN Multi-Site architecture has many different deployment scenarios that apply to different use cases. The topology that works best depends on the use case.

Topologies

This document considers the following major topologies:

- DCI
  - BGW to cloud
  - BGW back to back
- Multistage Clos (three tiers)
  - BGW between spine and superspine
  - BGW on spine

Although all of these designs look similar, you need to consider different factors when deploying them. The following sections describe the four topologies and the deployment details.

BGW-to-cloud model

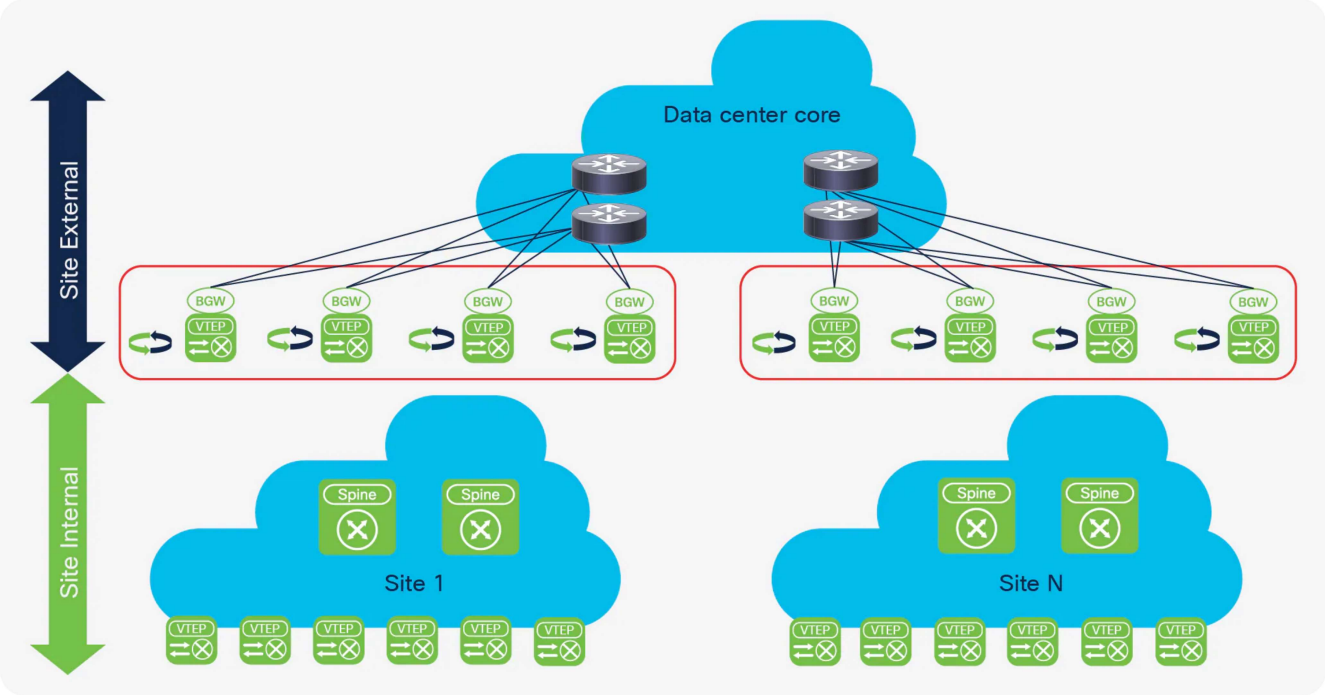


Figure 10.

BGW-to-Cloud Model

A common choice is to deploy the BGWs at the border of the fabric with the border leaf and DCI node functions. The BGW-to-cloud model (Figure 10) has a redundant Layer 3 cloud between the different sites. In this deployment model, the Layer 3 cloud provides to each site redundant connectivity points to which the BGWs can connect. Assuming four BGWs and two data center core devices, full-mesh connectivity can be established among them all, using the basic principle of building triangles, not squares. Similar connectivity can be achieved by the other sites, so that every BGW has redundant connectivity to the Layer 3 cloud, which also reduces the convergence time in a link-failure scenario.

The only specific requirements for the Layer 3 cloud are that it provide IP connectivity between the virtual IP and PIP addresses of the BGWs and accommodate the MTU for the VXLAN-encapsulated traffic across the cloud. The Layer 3 cloud can be any routed service, such as a flat Layer 3 routed network, a Multiprotocol Label Switching (MPLS) Layer 3 VPN (L3VPN), or other provider services. Whenever a VPN-like service is provided in the Layer 3 cloud, note that the physical interfaces on the BGW site must remain in the default VRF instance. Multiprotocol-BGP (MP-BGP) peering with VPN address families is supported only as part of the default VRF instance.

If a deployment consists of many sites and many BGWs, the need for full-mesh eBGP peerings between any BGWs for the overlay control plane may create additional complexity. The introduction of a Route Server (RS) can simplify the design and reduce the burden of having so many BGP peerings. A BGP route server is basically an eBGP route reflector, which in BGP terminology doesn't exist. A BGP route server performs the same route reflection function as an iBGP route reflector. Neither type of reflector needs to be in the data path to perform this function. Such a route server can be placed in the Layer 3 cloud or in a separate location reachable from every BGW. The route server will act as a star point for all the control-plane peerings for all the BGWs and will help ensure reflection of BGP updates. For resiliency, a pair of route servers is recommended.

BGW back-to-back model

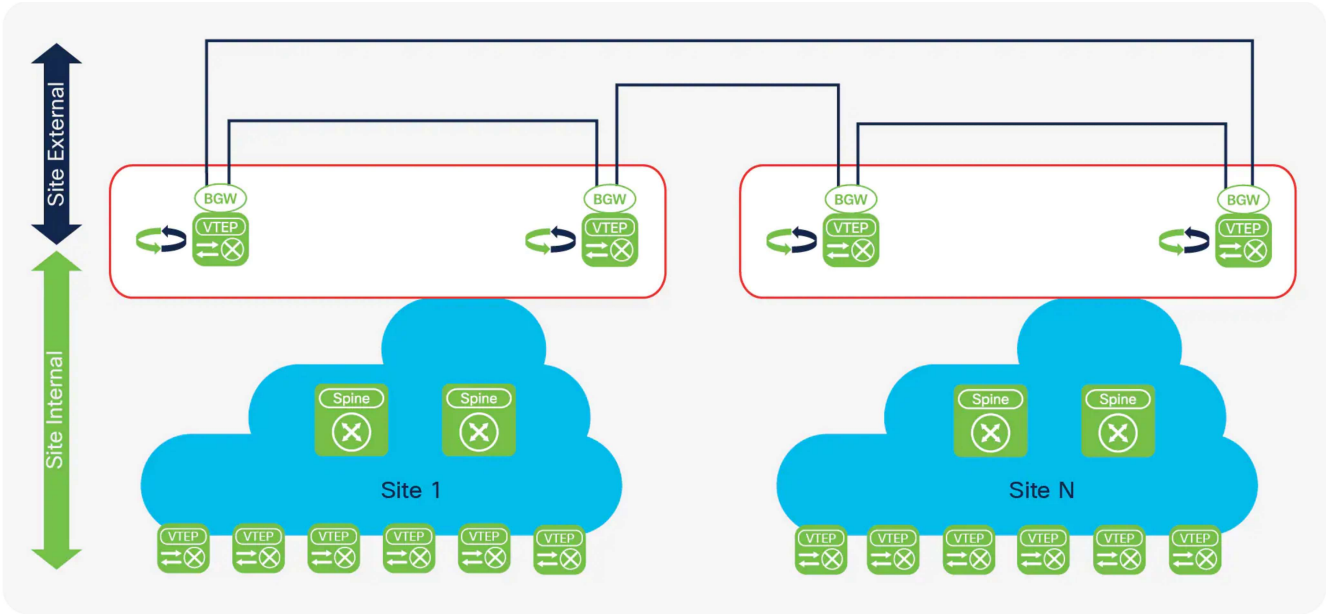


Figure 11.

BGW back-to-back model

The back-to-back connectivity model (Figure 11) provides an alternative to the topology in which the BGWs are connected to a Layer 3 cloud. For the back-to-back topology, you need to consider how the BGWs are interconnected within the site and between sites. In addition to physical-connectivity issues, you need to consider scenarios such as link failure, designated-forwarder reelection, and BUM-traffic forwarding (especially in a failure scenario).

Assuming two BGWs per site, the back-to-back connectivity model builds a square between the two BGWs at the local site and the two BGWs at the remote site. A permutation of this topology is a square with an additional cross between the BGWs, which is slightly more resilient and does not require designated-forwarder reelection if a single link fails. The Layer 3 underlay between all BGWs is achieved with a point-to-point subnet and the advertisement of the virtual IP and PIP addresses of the BGWs into this routing domain.

Note: The minimum back-to-back topology, the square, will not provide ECMP for fast convergence and traffic depolarization. In the extended back-to-back topology, with the square plus the full mesh between the BGWs, ECMP is available.

The only specific requirements for back-to-back connectivity are that it provide IP connectivity between all virtual IP and PIP addresses for the BGWs and accommodate the MTU for the VXLAN-encapsulated traffic across the links.

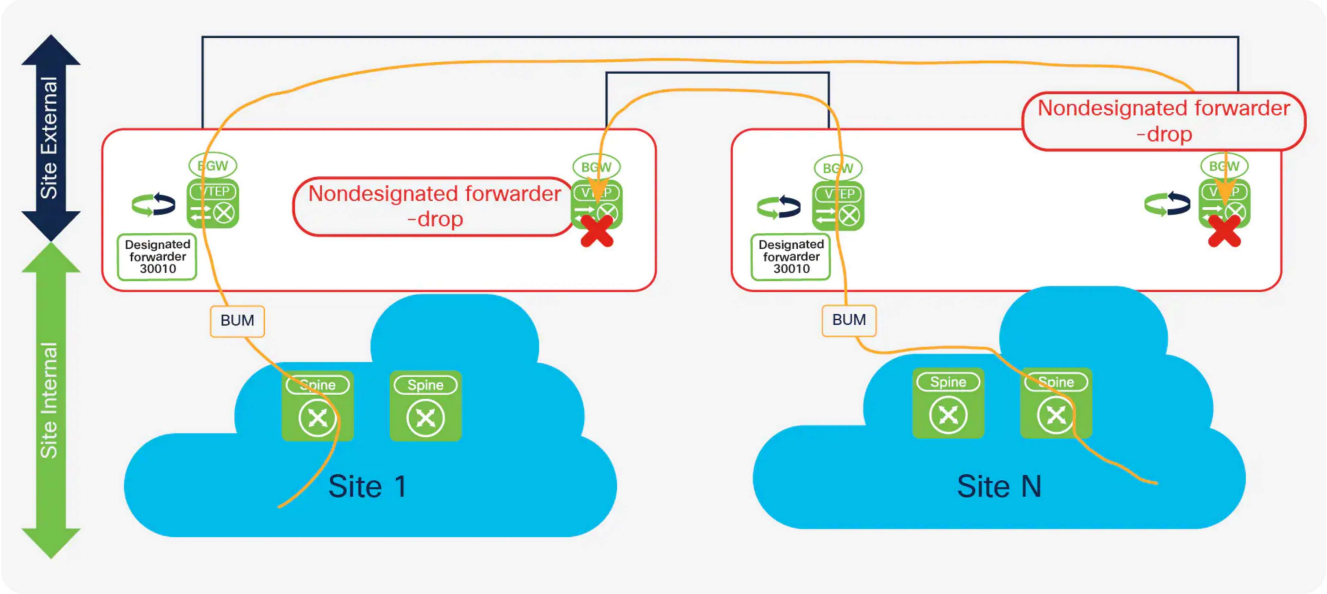


Figure 12.

BGW back-to-back model (BUM traffic not acceptable)

The minimum back-to-back topology is a square. The connection between the BGWs in the same site allows proper BUM-traffic handling during normal operations and failure scenarios, without requiring designated-forwarder reelection. In a square topology, in which the designated forwarder at the local site is connected to the nondesignated-forwarder spine at the remote site, BUM traffic cannot be forwarded to the remote site without the link between the BGW at the same site (Figure 12). The compensation link between the site-local BGWs allows BUM traffic to be forwarded flawlessly. The link between the BGWs can be considered a backup path to the remote site and can be configured with DCI tracking enabled (Figure 13).

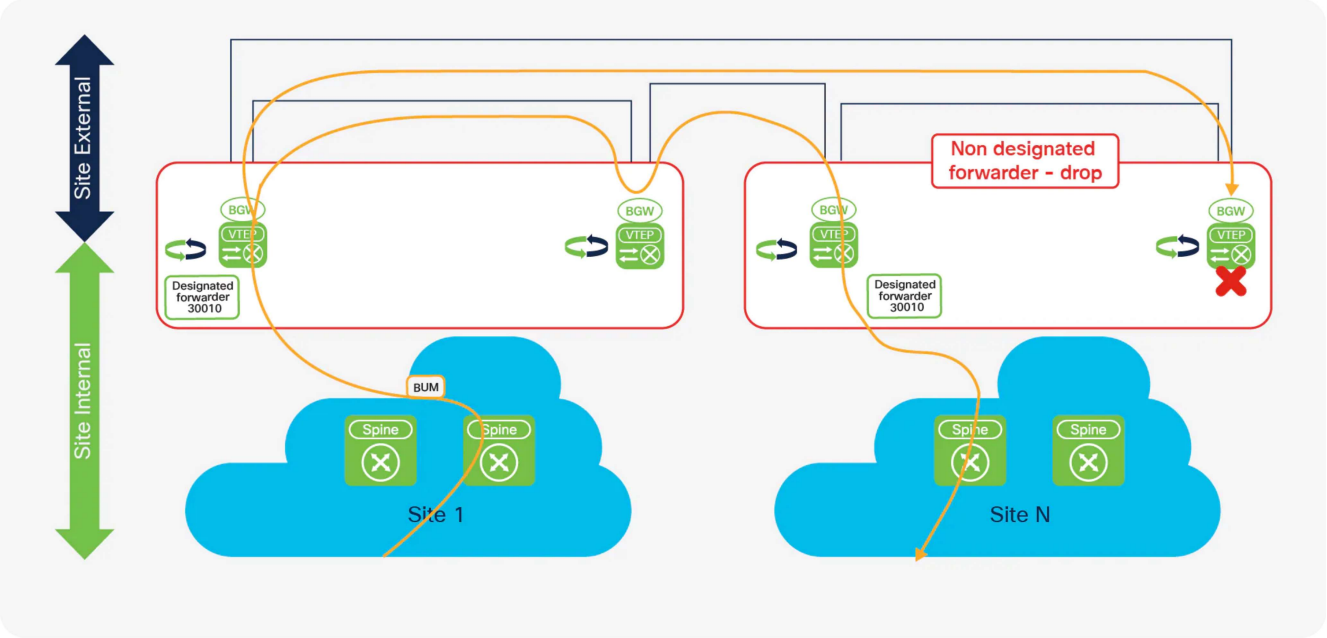


Figure 13.

BGW back-to-back model (BUM traffic acceptable)

Note: BUM replication between sites will always include in the replication list all BGWs with the respective destination Layer 2 VNIs.

Model with BGW between spine and superspine

The model in which the BGWs are placed between the spine and superspine (Figure 14) is similar to the BGW-to-cloud scenario. With a spine-and-leaf folded Clos model creating the site-internal network, the BGWs are placed on top of the spine. The superspine layer is part of the site-external network. With all the BGWs of the various sites connected to the superspine, you achieve a topology with the same network layers as in the BGW-to-cloud model. The main difference is in the geographical radius of such a topology. Whereas the BGW-to-cloud approach considers the Layer 3 cloud to be extended across a long distance, the superspine likely exists within a physical data center. With the superspine model, all BGWs of all sites connect to all superspines. This approach creates a high-speed backbone within a data center, also known as the data center core.

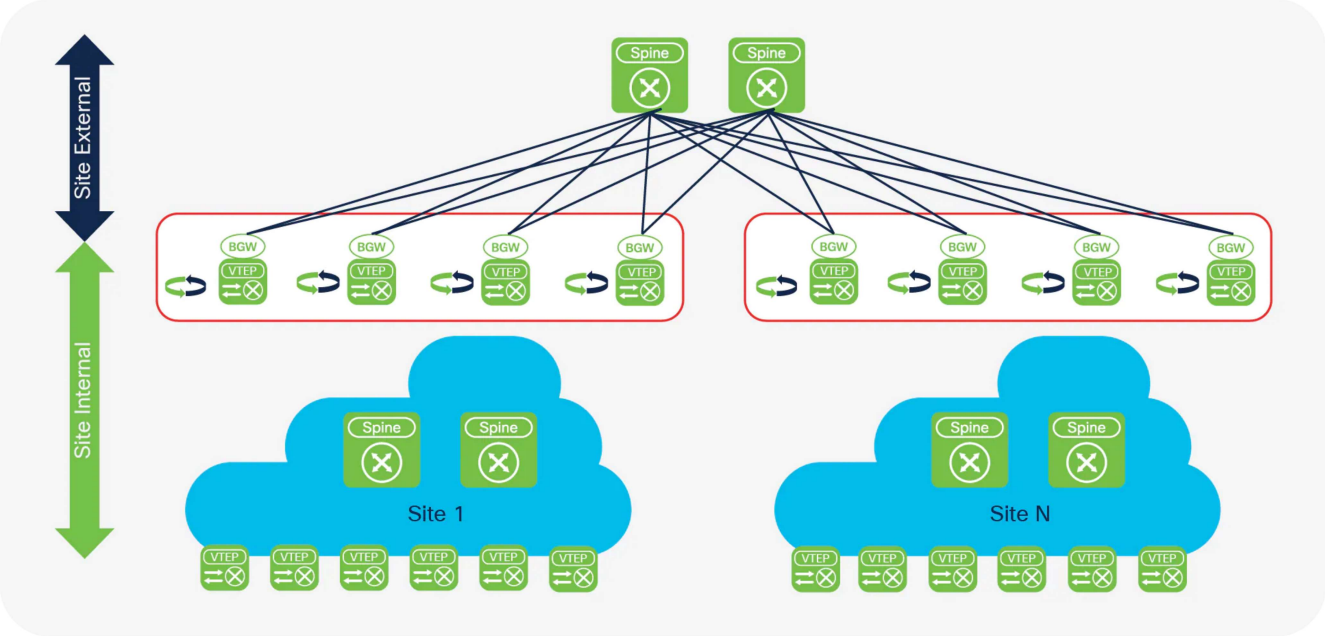


Figure 14.

Model with BGWs between spine and superspine

The deployment of the BGWs between the spine and superspine presents a deployment use case different from the DCI use case. With the BGWs between the spine and superspine, data center fabrics are scaled by interconnecting them in a hierarchical fashion. The achievement here is not simply extension of connectivity across fabrics. This approach also uses the masking that EVPN Multi-Site architecture provides to reduce the amount of peering between all VTEPs and thus to increase scale. With EVPN Multi-Site architecture and the BGWs, you can compartmentalize functional building blocks within the data center. The easy interconnection of these compartments is achieved through the integrated Layer 2 and Layer 3 extension provided by EVPN Multi-Site architecture. With selective control-plane advertisement and the enforcement of BUM traffic at the BGWs, you can achieve more control over extension between fabrics.

As with the BGW-to-cloud approach, the use of a BGP route server can be beneficial when you deploy BGWs between the spine and superspine. With many sites and many BGWs per site, the number of peering can easily grow dramatically. The route-server approach allows you to rein in the control-plane exchanges between all the BGWs across sites with a simplified peering model.

BGW-on-spine model

The previous topologies used dedicated BGW nodes. In the BGW-on-spine model (Figure 15), the BGW is co-located with the spine of the site-internal network (fabric). When the BGW and spine are combined, the exit points of the fabric and the spine are on the same set of network nodes. You thus need to consider, for example, how leaf-to-leaf communication occurs and how BGW-to-BGW communication occurs.

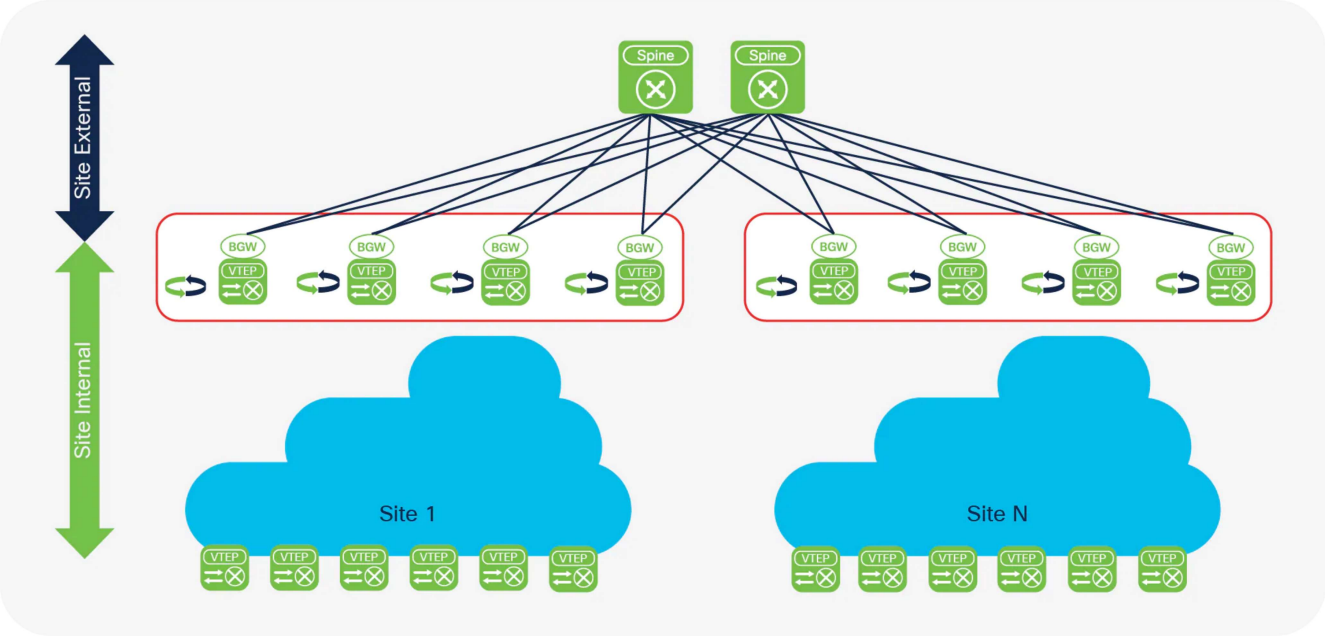


Figure 15.

BGW-on-spine model

For the site-internal VTEP or leaf-to-leaf communication, the traffic pattern is through the BGW and spine combination. Also, the services that a leaf requires are reachable through one hop at the BGW and spine.

BGW-to-BGW communication is less natural. For example, consider the designated-forwarder election exchange. The BGW and spine don't have any direct connection or BGP peering between them, so the control-plane exchange to synchronize the BGWs must be achieved through additional iBGP peering (full mesh). In this design, the only path available for the designated-forwarder exchange between the BGWs is through the site-internal VTEPs (leaf nodes). Although this approach doesn't create any problems from a traffic volume or a resiliency perspective, the use of a control-plane exchange between the BGW traversing the leaf node is not natural.

Underlay and overlay

The main functional component of the EVPN Multi-Site architecture consists of the BGW devices. Their deployment affects the way that the overlay network performs its Layer 2 and Layer 3 services. Given that stability is of paramount importance for the overlay, proper design of the underlay network is critical.

For EVPN Multi-Site architecture, numerous best practices and recommendations have been established to successfully deploy the overall solution. This document focuses mainly on two main models for the underlay. It also discusses the overlay.

- The I-E-I model focuses on an Interior Gateway Protocol (IGP) and iBGP (IGP-iBGP)-based site-internal network (fabric) with eBGP-eBGP at the external site (DCI).
- The E-E-E model uses eBGP-eBGP within the site (fabric) as well as between sites (DCI).

Note: Although Cisco supports both models, the I-E-I deployment scenario is recommended. For additional information about the E-E-E deployment model and why I-E-I is the recommended approach, see the "For more information" section at the end of this document.

The two models can be mixed in the sense that one site can run on "E" (eBGP-eBGP) and the other, remote site can run on "I" (IGP-iBGP). From an intersite underlay, eBGP can be replaced with any routing protocol, as long as a clean separation exists between the site-internal and site-external routing domains. As described later in this section, the "E" (eBGP) portion for the overlay is mandatory.

In addition to choosing the underlay routing protocols, you must separate the site-internal and site-external routing domains. In the case of I-E-I, the underlays will not likely be redistributed between the "I" (IGP) and the "E" (eBGP) domains. Furthermore, you must actively separate the site-internal underlay from the site-external underlay in the E-E-E case, because by default BGP automatically exchanges information between the underlay domains. In cases in which the site-internal and site-external underlays are joined, unanticipated forwarding and failure cases may occur.

The following sections present the main design principles for successfully deploying the EVPN Multi-Site architecture. The two primary topologies discussed here are the BGW-to-cloud model and the model with the BGW between the spine and superspine.

Site-internal underlay (fabric)

The site-internal underlay can be deployed in various forms. Most commonly, an IGP is used to provide reachability between the intrasite VTEP (leaf), the spine, and the BGWs. Alternative approaches for underlay unicast reachability use BGP; eBGP with dual- and multiple-autonomous systems are known designs.

For BUM replication, either multicast (PIM ASM) or ingress replication can be used. EVPN Multi-Site architecture allows both modes to be configured. It also allows different BUM replication modes to be used at different sites. Thus, the local site-internal network can be configured with ingress replication while the remote site-internal network can be configured with a multicast-based underlay.

Note: BGP EVPN allows BUM replication based on either ingress replication or multicast (PIM ASM). The use of EVPN doesn't preclude the use of a network-based BUM replication mechanism such as multicast.

BGW: Site-internal OSPF underlay

Figure 16 shows the BGW with a site-internal topology.

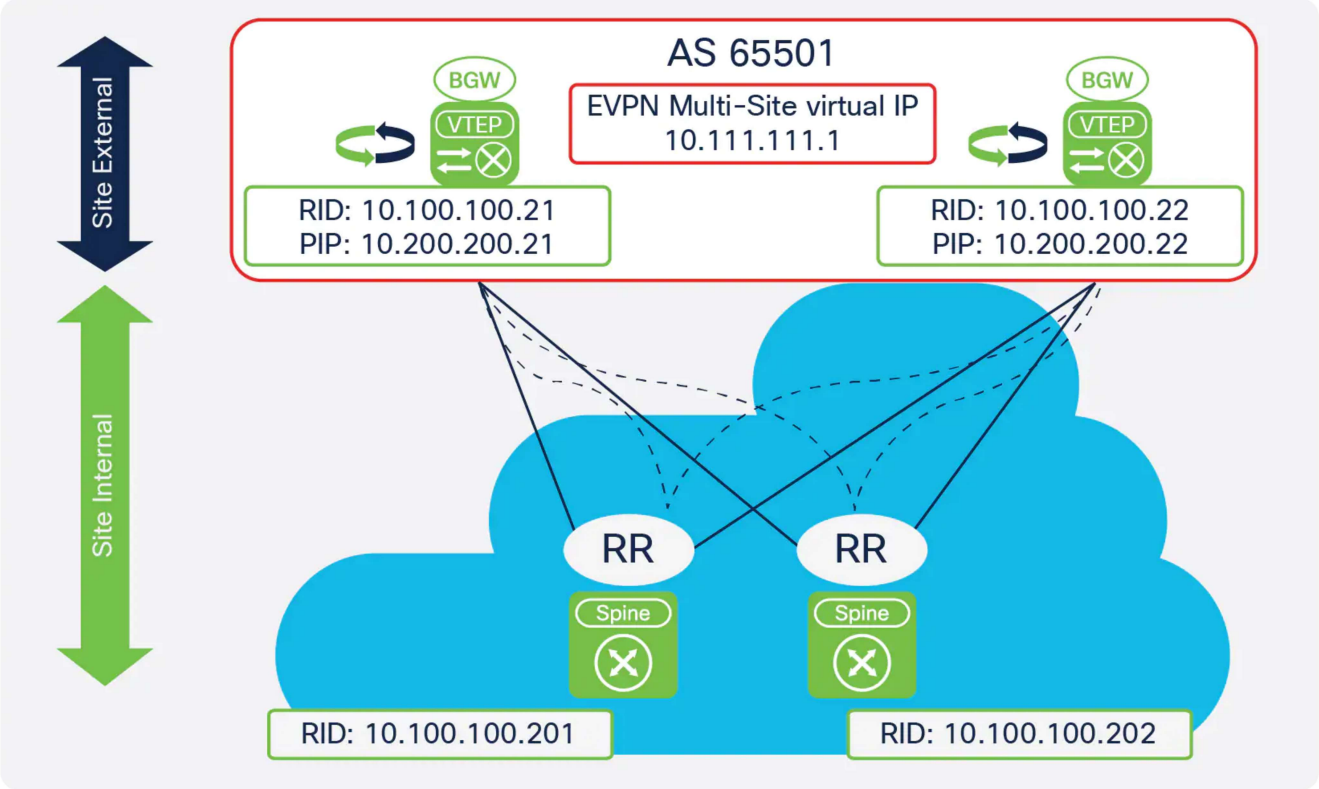


Figure 16.

BGW with site-internal topology

The configuration for a BGW with a site-internal OSPF underlay is shown here.

`version 7.0(3)I7(1)`	This version is the minimum software release required for EVPN Multi-Site architecture.

`feature ospf`	Enable **feature ospf** for underlay IPv4 unicast routing.

`feature pim`	Enable **feature pim** for multicast-based BUM replication. **Note:** This setting is not required if ingress replication is used for the intrasite underlay.

`router ospf UNDERLAY` `  router-id 10.100.100.21`	Define the OSPF process tag and OSPF router ID. **Note:** The OSPF router ID matches the loopback0 IP address.

`interface loopback0`
`  description RID AND BGP PEERING`
`  ip address 10.100.100.21/32 tag 54321`
`  ip router ospf UNDERLAY area 0.0.0.0`
`  ip pim sparse-mode`
	Define the loopback0 interface for the routing protocol router ID and overlay control-plane peering (that is, BGP peering).
	The IP address is extended with a tag to allow easy selection for redistribution.
	The OSPF process tag is used for site-internal underlay routing.
	Note: The **ip pim sparse-mode** setting is needed only for intrasite multicast-based BUM replication.

Note: The loopback interface used for the router ID and BGP peering must be advertised to the site-internal underlay as well as to the site-external underlay. If deemed beneficial, separate loopback interfaces can be used for site-internal and site-external purposes as well as for the various routing protocols (router ID, peering, etc.).

`interface loopback1`
`  description NVE INTERFACE (PIP VTEP)`
`  ip address 10.200.200.21/32 tag 54321`
`  ip router ospf UNDERLAY area 0.0.0.0`
`  ip pim sparse-mode`
	Define the loopback1 interface as the NVE source interface (PIP VTEP).
	The IP address is extended with a tag to allow easy selection for redistribution.
	The OSPF process tag is used for site-internal underlay routing.
	Note: The **ip pim sparse-mode** setting is needed only for intrasite multicast-based BUM replication.

Note: The loopback interface used for the individual VTEP (PIP) must be advertised to the site-internal underlay as well as to the site-external underlay.

`Interface loopback100`
`  description MULTI-SITE INTERFACE (VIP VTEP)`
`  ip address 10.111.111.1/32 tag 54321`
`  ip router ospf UNDERLAY area 0.0.0.0`
	Define the loopback100 interface as the EVPN Multi-Site source interface (anycast and virtual IP VTEP).
	The IP address is extended with a tag to allow easy selection for redistribution.
	The OSPF process tag is used for site-internal underlay routing.

Note: The loopback interface used for the EVPN Multi-Site anycast VTEP (virtual IP address) must be advertised to the site-internal underlay as well as to the site-external underlay.

`Interface Ethernet1/53`	Define site-internal underlay interfaces facing the spine. Adjust the MTU value for the interface to accommodate your environment (minimum value is 1500 bytes plus VXLAN encapsulation). You can use point-to-point IP addressing or IP unnumbered addressing (IP unnumbered support started in 7.0(3)I7(2)) for site-internal underlay routing (point-to-point IP addressing with /30 is shown here). Specify the OSPF network type (point to point) and OSPF process tag for site-internal underlay routing. **Note:** The **ip pim sparse-mode** setting is needed only for site-internal multicast-based BUM replication. Specify EVPN Multi-Site interface tracking for the site-internal underlay (**evpn multisite fabric-tracking**). This command is mandatory to enable the Multi-Site virtual IP address on the BGW. At least one of the physical interfaces that are configured with fabric tracking must be up to enable the Multi-Site BGW function (keeping the virtual IP VTEP address active).
`description SITE-INTERNAL INTERFACE`	
`no switchport`	
`mtu 9216`	
`medium p2p`	
`ip address 10.1.1.34/30`	
`ip ospf network point-to-point`	
`ip router ospf UNDERLAY area 0.0.0.0`	
`ip pim sparse-mode`	
`evpn multisite fabric-tracking`	
`interface Ethernet1/54`	
`description SITE-INTERNAL INTERFACE`	
`no switchport`	
`mtu 9216`	
`medium p2p`	

```
  ip address
10.1.2.34/30
  ip ospf
network
point-to-
point
  ip router
ospf
UNDERLAY
area 0.0.0.0
  ip pim
sparse-mode
  evpn
multisite
fabric-
tracking
```

Site-internal overlay

The site-internal overlay for VXLAN BGP EVPN always behaves like an iBGP deployment, whereas the underlay can use eBGP. This is the case regardless of whether a single-autonomous-system, dual-autonomous-system, or multiple-autonomous-system design is used. For a single-autonomous-system deployment, the overlay control-plane configuration is straightforward. For a dual- or multiple-autonomous-system design, additional BGP configurations are needed. This document focuses on EVPN Multi-Site architecture, so the site-internal overlay configuration for dual- and multiple-autonomous-system designs is omitted. For configuration guidance for dual- and multiple-autonomous-system designs, see the "For more information" section at the end of this document.

Note: If BGP EVPN control-plane communication between BGWs traverses a site-internal BGP route reflector, the route reflector must support BGP EVPN Route Type 4. If the route reflector doesn't support BGP EVPN Route Type 4, direct BGW-to-BGW full-mesh iBGP peering must be configured. BGP EVPN Route Type 4 is used for EVPN Multi-Site designated-forwarder election.

BGW: Site-internal iBGP overlay

The configuration for a BGW with a site-internal iBGP overlay is shown here.

`version 7.0(3)I7(1)`	This version is the minimum software release required for EVPN Multi-Site architecture.

`feature bgp`	Enable **feature bgp** for underlay IPv4 unicast routing.

`feature nv overlay`	Enable **feature nv overlay** for VXLAN VTEP capability.
`nv overlay evpn`	Extend the capability of VXLAN with EVPN (**nv overlay evpn**).

evpn multisite border-gateway <site-id> **delay-restore time 300**	Define the node as an EVPN Multi-Site BGW with the appropriate site ID. **Note:** All BGWs at the same site must have the same site ID (site ID 1 is shown here). As a subconfiguration of the BGW definition, a time-delayed restore operation for BGW virtual IP address advertisement can be set.

`interface nve1`
`  host-reachability protocol bgp`
`  source-interface loopback1`
** multisite border-gateway interface loopback100**
	Define the NVE interface (VTEP) and extend it with EVPN (**host-reachability protocol bgp**).
	Define the loopback1 interface as the NVE source interface (PIP VTEP).
	Define the loopback100 interface as the EVPN Multi-Site source interface (anycast and virtual IP VTEP).

`router bgp 65501`
`  neighbor 10.100.100.201`
`    remote-as 65501`
`    update-source loopback0`
`    address-family l2vpn evpn`
`      send-community`
`      send-community`
	Define the BGP routing instance with a site-specific autonomous system.
	Note: The BGP router ID matches the loopback0 IP address.
	Define the neighbor configuration with the EVPN address family (L2VPN EVPN) for the site-internal overlay control plane facing the route reflector.
	Configure the iBGP neighbor by specifying the source interface loopback0. This setting allows underlay ECMP reachability from BGW loopback0 to route-reflector loopback0.

```
extended


  neighbor
10.100.100.202
    remote-as
65501
    update-
source
loopback0
    address-
family l2vpn
evpn
      send-
community
      send-
community
extended
```

Site-external underlay (DCI)

The site-external underlay is the network that interconnects multiple VXLAN BGP EVPN fabrics. It is a transport network that allows reachability between all the EVPN Multi-Site BGWs and external VTEPs. Some deployment scenarios use an additional spine tier (superspine), and other deployments have a routed Layer 3 cloud.

The site-external underlay network can be deployed with various routing protocols, but eBGP is typically used to provide reachability between the BGWs of multiple sites, given its interdomain nature. Alternative approaches for underlay reachability include the use of IGP, but this document focuses solely on eBGP.

For BUM replication between sites, EVPN Multi-Site architecture exclusively uses ingress replication to simplify the requirements of the site-external underlay network.

Note: Ingress replication to handle BUM replication between sites (site-external network) doesn't limit the use of the available BUM replication mode to a given site (site-internal network). EVPN Multi-Site architecture allows the use of multicast (PIM ASM) for BUM replication within one site, while other sites can use ingress replication or multicast.

BGW: Site-external eBGP underlay

Figure 17 shows the BGW with a site-external topology.

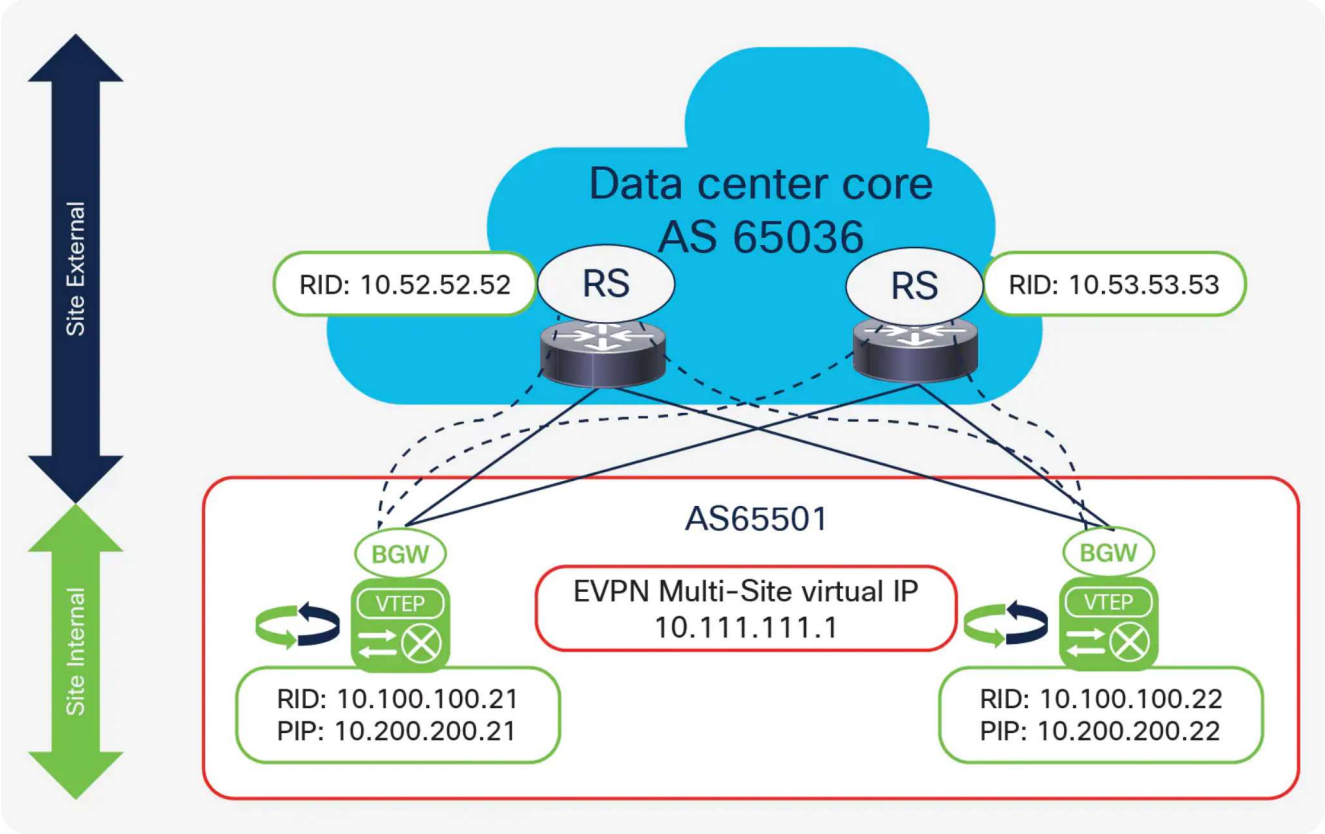


Figure 17.

BGW with site-external topology

The configuration for a BGW with a site-external eBGP underlay is shown here.

`version 7.0(3)I7(1)`	This version is the minimum software release required for EVPN Multi-Site architecture.

`feature bgp`	Enable **feature bgp** for underlay IPv4 unicast routing.

`interface Ethernet1/1`
`  no switchport`
`  mtu 9216`
`  ip address 10.52.21.1/30 tag 54321`
`  evpn multisite dci-tracking`
	Define site-external underlay interfaces facing the external Layer 3 core.
	Adjust the MTU value of the interface to accommodate your environment (the minimum value is1500 bytes plus VXLAN encapsulation).
	Point-to-point IP addressing is used for site-external underlay routing (point-to-point IP addressing with /30 is shown here). The IP address is extended with a tag to allow easy selection for redistribution.
	Note: The **ip pim sparse-mode** setting is not needed because site-external BUM replication always uses ingress replication.
	EVPN Multi-Site interface tracking is used for the site-external underlay (**evpn multisite dci-tracking**). This command is mandatory to enable the Multi-Site virtual IP address on the BGW. At least one of the physical

`interface Ethernet1/2`
`  no switchport`
`  mtu 9216`
`  ip address 10.53.21.1/30 tag 54321`
`  evpn multisite dci-tracking`
	interfaces that are configured with DCI tracking must be up to enable the Multi-Site BGW function.

`route-map RMAP-REDIST-DIRECT permit 10` `  match tag 54321`	The route map is used to select all IP addresses that are attached to an interface and that carry the tag extension.

`router bgp 65501`
`  router-id 10.100.100.21`
`  log-neighbor-changes`
`  address-family ipv4 unicast`
 	
`redistribute direct` **`route-map RMAP-REDIST-DIRECT`**
`    maximum-paths 4`
	Define the BGP routing instance with a site-specific autonomous system.
	Note: The BGP router ID matches the loopback0 IP address.
	Activate the IPv4 unicast global address family (VRF default) to redistribute the required loopback and physical interface IP addresses within BGP.
	Enable BGP multipathing (**maximum-paths**).
	Note: The redistribution from the locally defined interfaces (direct) to BGP is performed through route-map classification. Only IP addresses in the VRF default instance that are extended with the matching tag of the route map are redistributed.

`neighbor 10.52.21.2`
`remote-as 65036`
`update-source Ethernet1/1`
`address-family ipv4 unicast`
	The neighbor configuration for the IPv4 unicast global address family (VRF default) facilitates site-external underlay routing.
	Configure the eBGP neighbor by selecting the source interface for this eBGP peering.

`neighbor 10.53.21.2`
`remote-as 65036`
`update-source Ethernet1/2`
`address-family ipv4 unicast`	

Site-external overlay

The site-external overlay for VXLAN BGP EVPN must use eBGP, because the eBGP next-hop behavior is used for VXLAN tunnel termination and reorigination.

In the case of EVPN Multi-Site architecture, a site-internal MAC address or IP prefix advertisement originates from the local BGWs with their anycast VTEPs as the next hop. Similarly, the BGWs of the local site receive a MAC address or IP prefix advertised from remote BGWs with their anycast VTEPs as the next hop. This behavior follows eBGP's well-known and proven process of changing the next hop at the autonomous system boundary. EVPN Multi-Site architecture uses eBGP not only for VXLAN tunnel termination and reorigination, but also for its loop prevention mechanism offered through the as-path attribute. With this approach, on the control plane, prefixes originating at one site will never be imported back into the same site, thus preventing routing loops. On the data plane, designated-forwarder election and split-horizon rules complement the control-plane loop-prevention functions.

Note: BGP EVPN control-plane communication between BGWs at different sites can be achieved using either a full mesh or a route server (eBGP route reflector).

BGW: Site-external eBGP overlay

The configuration for a BGW with a site-external eBGP overlay is shown here.

`version 7.0(3)I7(1)`	This version is the minimum software release required for EVPN Multi-Site architecture.

`feature bgp`	Enable **feature bgp** for underlay IPv4 unicast routing.

`feature nv overlay`	Enable **feature nv overlay** for the VXLAN VTEP capability.
`nv overlay evpn`	Extend VXLAN with EVPN (**nv overlay evpn**).

evpn multisite border-gateway <site-id> **delay-restore time 300**	Define the node as an EVPN Multi-Site BGW with the appropriate site ID. **Note:** All BGWs at the same site must have the same site IDs (site ID 1 is shown here). As a subconfiguration of the BGW definition, a time-delayed restore operation for BGW virtual IP address advertisement can be set.

`interface nve1`
`host-reachability protocol bgp`
`source-interface loopback1`
multisite border-gateway interface loopback100
	Define the NVE interface (VTEP) and extend it with EVPN (**host-reachability protocol bgp**).
	Define the loopback1 interface as the NVE source interface (PIP VTEP).
	Define the loopback100 interface as the EVPN Multi-Site source interface (anycast and virtual IP VTEP).

Note: Feature enablement and VXLAN, BGP EVPN, and EVPN Multi-Site global configuration have already been described in the “BGW: Site-internal iBGP overlay”.

`router bgp 65501`	Define the BGP routing instance with a site-specific autonomous system. **Note:** The BGP router ID matches the loopback0 IP address.
`router-id 10.100.100.21`	Configure the neighbor with the EVPN address family (L2VPN EVPN) for the site-external overlay control plane facing the route server or remote BGW (peering to a pair of route servers is shown here).
`log-neighbor-changes`	Configure the eBGP neighbor by specifying the source interface loopback0. This setting allows underlay ECMP reachability from BGW loopback0 to route-server loopback0.

```
  neighbor
10.52.52.52
    remote-as
65036
    update-
source
loopback0
    ebgp-
multihop 5
    peer-type
fabric-external
    address-
family l2vpn
evpn
      send-
community
      send-
community
extended
      rewrite-
evpn-rt-asn


  neighbor
10.53.53.53
    remote-as
65036
    update-
source
loopback0
    ebgp-
multihop 5
    peer-type
fabric-external
    address-
family l2vpn
evpn
      send-
community
```

Note: Site-external EVPN peering is always considered to use eBGP with the next hop the remote site BGWs.

With the route server or remote BGW potentially multiple routing hops away, you must increase the BGP session Time-To-Live (TTL) setting to an appropriate value (**ebgp-multihop**).

In defining the site-external BGP peering session (**peer-type fabric external**), rewrite and reorigination are enabled. (This function is explained in detail in the upcoming section “Site-external route server”.

The autonomous system portion of the automated route target (ASN:VNI) will be rewritten upon receipt from the site-external network (**rewrite-evpn-rt-asn**) without modification of any configurations on the site-internal VTEPs. The route-target rewrite will help ensure that the ASN portion of the automated route target matches the destination autonomous system.

```
        send-
community
extended
        rewrite-
evpn-rt-asn
```

Route server (eBGP route reflector)

EVPN Multi-Site architecture requires every BGW from a local site to peer with every BGW at remote sites. This full-mesh requirement is not mandatory for a proper exchange of information in a steady-state environment, but given the various failure scenarios that are possible, a full mesh is the recommended configuration (Figure 18). When you deploy two sites with two BGWs in each topology, the number of BGP peerings remains manageable. However, when you scale out the EVPN Multi-Site environment and add more sites and BGWs to each site, the number of full-mesh BGP peerings becomes difficult to manage and creates a burden on the control plane.

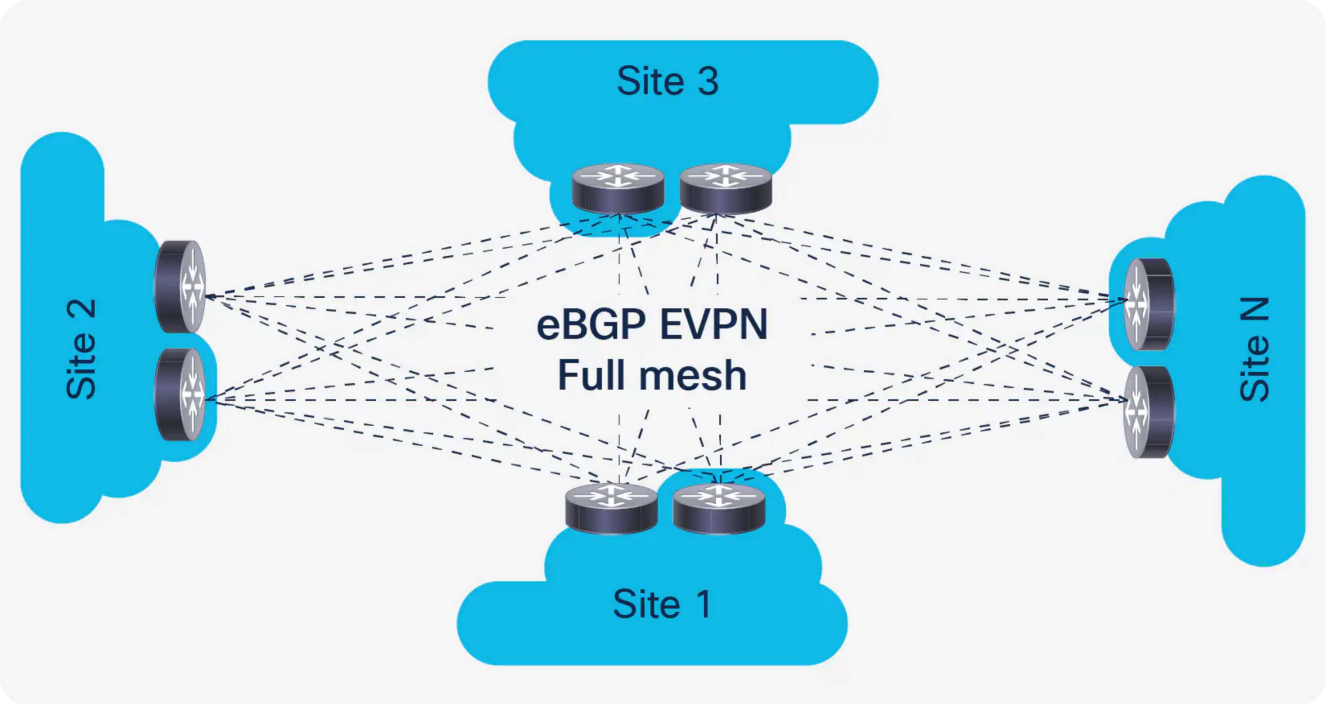


Figure 18.

EVPN Multi-Site without route server

A more elegant approach to a scale-out EVPN Multi-Site environment is to use a star point to broker the site-external overlay control plane (Figure 19). Such nodes are well known in iBGP environments as route reflectors. They are present to reflect routes that are being sent from their clients that don't require a full mesh anymore. This approach allows the environment to scale well from control-plane peering, and it also eases the management burden of configuration and operation. BGP route reflectors are limited to providing their services to iBGP-based peering. In the case of eBGP networks, the route-reflector function is absent or nonexistent. However, for eBGP networks, a function similar to the route-reflector function is offered by the route server, as described in IETF RFC 7947: Internet Exchange BGP Route Server.

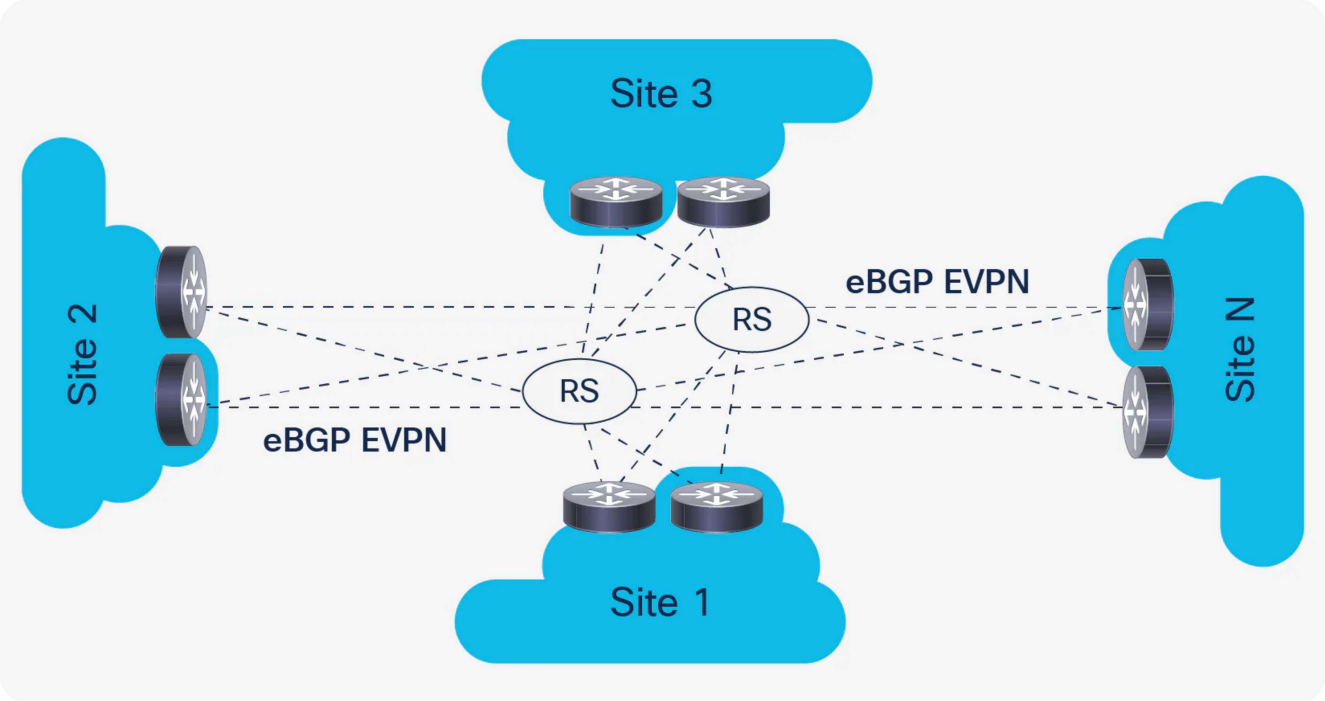


Figure 19.

EVPN Multi-Site with route server

Like a route reflector, a route server performs a pure control-plane function and doesn't need to be in the data path between any of the BGWs. To help ensure that the route-server deployment provides resiliency for the EVPN Multi-Site control-plane exchange in any failure scenario, connectivity or device redundancy is required. Various platforms support the configuration of a route server in either a hardware-only or software-only design. Cisco NX-OS offers the route-server capability in the Cisco Nexus Family switches, which can be connected on a stick or within the data path as a node for the site-external underlay. The route server must be able to support the EVPN address family, reflect VPN routes, and manipulate the next-hop behavior (**next-hop unchanged**). In addition, the route server should support route-target rewrite to simplify the deployment.

Site-external route server

The configuration for a site-external route server is shown here.

`feature bgp`	Enable **feature bgp** for underlay IPv4 unicast routing.

`route-map UNCHANGED permit 10` `  set ip next-hop unchanged`	The route map enforces the policy to leave the overlay next hop unchanged when the route server is used. **Note:** The route server is not a VTEP or BGW and hence should not have the next hop pointing to itself.

`router bgp 65036`
 	
`address-family l2vpn evpn`
 	
`retain route-target all`
	Define the BGP routing instance with a site-independent autonomous system.
	You must ensure that all the received EVPN advertisements are reflected even if all the tenant VRF instances are not created on the route server. The route targets must be preserved while that function is performed (**retain route-target all**).

`  template peer OVERLAY-PEERING`
 	
`update-source loopback0`
`    ebgp-multihop 5`
 	
`address-family l2vpn evpn`
 	
`send-community both`
 	
`route-map UNCHANGED out`
	The per-neighbor configuration for the overlay control-plane function in a route server can be simplified. The configuration of the BGP reachability function across multiple hops (**ebgp-multihop**) and preservation of the next hop between the BGWs are common settings. These configuration knobs, including the source interface, can be combined in a BGP peer template.
	Note: BGP peer templates are part of the BGP instance configuration.

`  neighbor`	Configure the neighbor in the IPv4 unicast global address family (VRF

`10.100.100.21 remote-as 65501`
`  inherit peer OVERLAY-PEERING`
`  address-family l2vpn evpn`
`  rewrite-evpn-rt-asn`
	default) to peer with the site-external loopback interface (loopback0) of the BGW.
	Configure the eBGP neighbor by using BGP peer templates and activating the EVPN address family (address family L2VPN EVPN).
	The autonomous system portion of the automated route target (ASN:VNI) will be rewritten upon receipt from the site-external network (**rewrite-evpn-rt-asn**) without modification of any configuration on the site-internal VTEPs. If a route server stands in between the BGWs of the individual sites, an additional rewrite to the destination autonomous system is performed. The route-target rewrite helps ensure that the ASN portion of the automated route target matches the destination autonomous system.

`neighbor 10.100.100.22 remote-as 65501`
`  inherit peer OVERLAY-PEERING`
`  address-family l2vpn evpn`
`  rewrite-evpn-rt-asn`	

`neighbor 10.101.101.41 remote-as 65520`
`  inherit peer OVERLAY-PEERING`
`  address-family l2vpn evpn`
`  rewrite-evpn-rt-asn`	

`neighbor 10.101.101.42`	

```
remote-as
65520
    inherit
peer OVERLAY-
PEERING
    address-
family l2vpn
evpn
       rewrite-
evpn-rt-asn
```

Note: The use of a route server is optional, but it simplifies the EVPN Multi-Site deployment.

Route-target rewrite

Previous configuration sections mentioned the capability to rewrite the automated route-target macros.

In VXLAN EVPN, Cisco NX-OS uses an automated route-target derivation in which a prefix is followed by a 2-byte Autonomous System Number (ASN). The suffix of the route target is populated with the VNI, which has a total size of 4 bytes. The prefix portion with the ASN is derived from the BGP instance that is locally configured on the respective node, and the VNI is derived from either the Layer 2 or Layer 3 configuration and its use depends on whether a MAC or IP address import must be performed. Table 2 shows an example.

Table 2. Sample route-target prefix and suffix

Prefix	Suffix
2-byte ASN	**4-byte VNI**
65501	50000

When the MP-BGP and VPN address families are used, the route target defines what is imported into a given VRF instance. The route target is defined based on the export configuration of the VRF instance in which the prefix was learned. The route target is attached to the BGP advertisement as an extended community to the prefix itself. At the remote site, the import configuration of the VRF instance defines the route-target extended community that is matched and the information that is imported.

In EVPN Multi-Site architecture, each site is defined as an individual BGP autonomous system. Thus, with the use of automated route targets, the configurations of the VRF instance and the route-target extended community potentially diverge. For instance, if the local site uses ASN 65501 and the remote site uses ASN 65520, the route targets will be misaligned, and no prefixes learned from the control plane will be imported.

To allow the site-internal configuration to use the automated route target and require no change to any VTEP, the rewriting of the autonomous system portion on the route target must be possible, because the export route target at the local site must match the import route target at the remote site. In EVPN Multi-Site architecture, the route target can be rewritten during ingress at the remote site.

The autonomous system portion of the route target will be rewritten with the ASN specified in the BGP peering configuration. This action allows, for example, route-target 65501:50000 at the local site to be rewritten as 65520:50000 upon receipt of the BGP advertisements at the BGW of the remote site. If a route server is between the BGWs, additional route-target rewrite must be performed on the route server. In this case, for example, route-target 65501:50000 at the local site can be rewritten as 65036:50000 on the route server and then as 65520:50000 at the remote site. This example assumes a symmetric VNI deployment (the same VNI across sites).

This approach enables successful export and import route-target matching by using automated route-target derivation with route-target rewrite. Neither the existing VTEP configuration or the static route-target configuration needs to be changed.

The route-target rewrite function is performed on the EVPN Multi-Site BGW facing the site-external overlay peering.

Note: As of Cisco NX-OS 7.0(3)I7(1), automated route-target derivation and route-target rewrite are limited to a 2-byte ASN. This limitation as a result of the route-target format (ASN:VNI) used, which allows space for a 2-byte prefix (ASN) with a 4-byte suffix (VNI). In cases in which a 4-byte ASN is required, you can use common route targets across sites.

Peer-type fabric-external function

Whereas the route-target rewrite function is an optional configuration to simplify the deployment, the definition of site-external overlay peering on the EVPN Multi-Site BGW is mandatory.

EVPN Multi-Site architecture adds the function that enables intermediate nodes, the BGWs, to terminate and reoriginate VXLAN encapsulation at Layer 2 and Layer 3. In BGP EVPN–based overlay networks, the control plane defines what the data plane and VXLAN use to build adjacencies, for example. The EVPN Multi-Site architecture is based on IETF draft-sharma-multi-site-evpn.

IETF RFC-7432 and draft-ietf-bess-evpn-overlay, draft-ietf-bess-evpn-prefix-advertisement, and draft-ietf-bess-evpn-inter-subnet-forwarding specify that BGP EVPN Route Type 2 and Route Type 5 carry the Router MAC (RMAC) address of the next hop's VTEP (Table 3). EVPN Multi-Site architecture masks the original advertising VTEP (usually a local leaf node) behind the BGW, and hence the RMAC must match the BGW in between rather than the advertising VTEP. The introduction of the peer-type fabric-external function helps ensure that the advertised VTEP IP information is properly rewritten (virtual IP address) and that the RMAC address present in EVPN Route Type 2 and Route Type 5 matches the virtual MAC address of the BGW. With the implementation of this function, every IETF RFC and draft conforming VTEP can peer with a BGW either site internal or site external without specifically needing to have EVPN Multi-Site BGW capabilities.

Note: Cisco NX-OS follows the following implementation as defined by IETF RFC-7342, draft-ietf-bess-evpn-overlay, draft-ietf-bess-evpn-prefix-advertisement, and draft-ietf-bess-evpn-inter-subnet-forwarding.

Table 3. IETF specifications for EVPN Multi-Site architecture

RFC or draft		
RFC-7432	VLAN-based service interface BGP EVPN routes	Section 6.1 Section 7
draft-ietf-bess-evpn-overlay	Encapsulation options	Section 5
draft-ietf-bess-evpn-prefix-advertisement	Interface-less IP-VRF-to-IP-VRF advertisement	Section 4.4.1
draft-ietf-bess-evpn-inter-subnet-forwarding	Symmetric intersubnet forwarding	Section 5

To successfully peer with an EVPN Multi-Site BGW, RFC and draft conformity must be achieved, and a common BUM replication mode must be used. Supported site-internal BUM replication modes are multicast (PIM ASM) and ingress replication. The supported site-external BUM replication mode is ingress replication.

Per-tenant configuration

Previous sections discussed EVPN Multi-Site design scenarios and underlay and overlay configurations. This section explores the configurations needed for the VNIs, for either Layer 2 or Layer 3 extension. This section also discusses how to limit the extension, from either the control plane (selective advertisement) or data plane (BUM enforcement).

This section begins by exploring the name-space mapping for VNIs and the use of VNIs across multiple sites with EVPN Multi-Site architecture.

Symmetric VNI

EVPN Multi-Site architecture allows the extension of Layer 2 and Layer 3 segments beyond a single site. Using EVPN Multi-Site architecture, you can extend Layer 2 VNIs to enable seamless endpoint mobility and address other use cases that require communication bridged beyond a single site. Use cases involving Layer 3 extension beyond a single site primarily require multitenant awareness or VPN services. With the multitenant capability in BGP EVPN and specifically in EVPN Multi-Site architecture, multiple VRF instances or tenants can be extended beyond a single site using a single control plane (BGP EVPN) and a single data plane (VXLAN).

All the use cases for EVPN Multi-Site architecture have the name space provided by VXLAN—the VXLAN network identifier, or VNI—as a central feature. This 24-bit name space, with about 16 million potential identifiers, is an integral part of VXLAN and is used by VXLAN BGP EVPN and EVPN Multi-Site architecture.

As of Cisco NX-OS 7.0(3)I7(1) for the Cisco Nexus 9000 Series EX- and FX-platform switches, all deployed sites must follow a consistent assignment of VNIs for either Layer 2 or Layer 3 extension. Therefore, a VLAN or VRF instance at the local site must be mapped to the same VNI that is used at the remote site. This consistent mapping is called symmetric VNI assignment. Subsequent releases will expand this capability to enable asymmetric VNI assignment, in which different VNIs can be stitched together at the BGW level.

Selective advertisement

With the use of Layer 2 and Layer 3 extension to facilitate endpoint mobility, the boundaries of hierarchical addressing are nonexistent. Thus, an individual endpoint's MAC address and host IP address must be seen within a site or across sites whenever bridging communication is required. The host IP address is not especially important for the bridging itself, but it is needed to provide optimal routing between endpoints. To help ensure that endpoints in different IP subnets can communicate without hairpinning through a remote site, knowledge of the /32 and /128 host routes is crucial.

EVPN Multi-Site architecture not only facilitates these Layer 2 and Layer 3 extension use cases, but it also provides ways to optimize such environments, building hierarchical networks even when Layer 2 extension is needed. EVPN Multi-Site selective advertisement limits the control-plane advertisements on the BGW depending on the presence of per-tenant configurations. If a VRF instance is configured on the BGW to allow a multitenant-aware Layer 3 extension, the data plane is configured, and control-plane advertisement in BGP EVPN is enabled. With this approach, only after the VRF instance is configured and associated with the VTEP is the relevant IP host and IP subnet prefix information advertised to the site-external network. The same approach is followed for Layer 2 extension and MAC address advertisement, with advertisements sent to the site-external network only after the Layer 2 segment has been configured and associated with the VTEP.

These advertisement control functions are provided simply to keep the site-external network manageable and to prevent saturation of the control-plane tables with unnecessary entries. In addition, if VRF route-target imports are configured unintentionally, the selective advertisement approach helps preserve hardware table space on the BGW and even on the VTEPs beyond it.

Selective advertisement is implicitly enabled. Control-plane advertisements are limited based on the local VRF and VNI configurations on the BGWs.

Layer 3 extension

The configuration to enable Layer 3 extension through an EVPN Multi-Site BGW closely follows the configuration for a normal VTEP. However, for an EVPN Multi-Site BGW, no endpoint-facing Layer 2 or Layer 3 configuration is defined. All the per-tenant configuration settings for Layer 3 are provided solely to allow VXLAN traffic termination and reencapsulation for transit through the BGW. The configuration used for the BGW transit functions also facilitates the selective advertisement control explained in the previous section.

Note: All BGWs at a given site must have the same configurations for Layer 3 extensions.

`vlan 2003` `  vn-segment 50001`	Define the Layer 3 VNI and attach it to a BGW local VLAN. **Note:** The VLAN ID has no significance for any endpoint-facing function. It is a resource allocation setting only.

```
vrf
context
BLUE
  vni
50001
  rd
auto

address-
family
ipv4
unicast

route-
target
both
auto

route-
target
both
auto
evpn

address-
family
ipv6
unicast

route-
target
both
auto

route-
target
both
auto
evpn
```

Define a VRF context (IP VRF) with the appropriate instance name.

The Layer 3 VNI chosen refers to the **vn-segment** ID chosen in the previous step.

The route distinguisher for the IP VRF instance can be derived automatically by using the router ID followed by the internal VRF ID (RID:VRF-ID). Similarly, the route target can be derived automatically by using the BGP autonomous system followed by the VNI defined as part of the VRF instance (ASN:VNI). The route targets must be enabled for the IPv4/IPv6 address family and specifically for EVPN.

Note: The use of the automated route distinguisher and route target is optional, but it is a best practice.

```
interface loopback 51
  vrf member BLUE
  ip address 10.55.55.1/32
```

Note: In cases where only Layer 3 extension is configured on the BGW an additional loopback interface is required. The loopback interface must be present in the same VRF instance on all BGW and with an individual IP address per BGW. Ensure the loopback interfaces IP address is redistributed into BGP EVPN, specially towards Site-External.

```
interface Vlan2003
  mtu 9192
  vrf member BLUE
  no ip redirects
  ip forward
  ipv6 forward
  no ipv6 redirects
```

Define a Layer 3 interface to enable the previously defined VNI to become a fully functional Layer 3 VNI.

Verify that the MTU accommodates your needs and that the forwarding matches the IPv4/IPv6 requirements.

Note: The SVI identifier must match the identifier that was chosen earlier. The VRF member name must match the VRF context name in the next step.

```
interface nve1
  member vni 50001 associate-vrf
```

Associate the Layer 3 VNI with the NVE interface (VTEP) and associate it with the VRF type.

Note: In addition to configuring the Layer 3 extension, you may need to add the VRF information in the configuration of the BGP instance. This step is mandatory if external connectivity for locally connected devices is required.

Layer 2 extension

As with Layer 3 extension, the configuration to enable Layer 2 extension through an EVPN Multi-Site BGW is similar to the configuration used for a normal VTEP. However, for an EVPN Multi-Site BGW, no

endpoint-facing Layer 2 or Layer 3 configuration is defined (that is, no distributed IP anycast gateway). All the Layer 2 configuration settings are provided solely to help ensure VXLAN traffic termination and reencapsulation for transit through the BGW only. The configuration for Layer 2 extension also promotes selective advertisement beyond the BGW.

Note: All BGWs for a given site must have the same configuration for Layer 2 extensions.

`vlan 10` `  vn-segment 30010`	Define the Layer 2 VNI and attach it to a BGW local VLAN. **Note:** The VLAN ID has no significance for any endpoint-facing function. It is a resource allocation setting only.

`interface nve1`
`  member vni 30010`
`  multisite ingress-replication`
`  [ingress-replication protocol bgp]`
`  [mcast-group 239.1.1.0]`
	Associate the Layer 2 VNI with the NVE interface (VTEP) and configure the relevant site-internal and site-external BUM replication modes (dual mode).
	Note: Site-external BUM replication always uses ingress replication. Site-internal BUM replication can use multicast (PIM ASM) or ingress replication.
	Note: Configure only one site-internal BUM replication mode: either multicast (PIM ASM) or ingress replication.

`evpn`
`  vni 30010 l2`
`  rd auto`
`  route-target import auto`
`  route-target`
	Define a VRF context (MAC VRF instance) with the appropriate Layer 2 VNI and the forwarding mode (L2).
	The Layer 2 VNI chosen refers to the **vn-segment** ID chosen in the previous step.
	The route distinguisher of the MAC VRF instance can be derived automatically by using the router ID followed by the internal VRF ID (RID:VRF-ID). Similarly, the route target can be derived automatically by using the BGP autonomous system followed by the VNI defined as part of the VRF instance (ASN:VNI). The route targets must be enabled for the IPv4/IPv6 address family and specifically for EVPN.
	Note: The use of an automated route distinguisher and route target is optional, but it is a best practice.

```
export
auto
```

Note: As of Cisco NX-OS 7.0(3)I7(1) for the Cisco Nexus 9000 Series EX- and FX-platform switches, local endpoint connectivity is not supported on an EVPN Multi-Site BGW.

BUM traffic enforcement

Layer 2 extension is a common use case. It is also a scenario in which failure replication is largely exposed. To provide a safer approach for Layer 2 extension, EVPN Multi-Site architecture allows you to control Layer 2 BUM traffic leaving the local site. EVPN Multi-Site architecture uses separate flood domains for site-internal and site-external traffic. This approach allows you to filter traffic between the flood domains. It also introduces split-horizon rules to help ensure that traffic entering the BGW from one flood domain does not return to the same flood domain. If BUM traffic reaches the BGW from the site-internal network, forwarding is allowed only to the site-external network, and if BUM traffic reaches the BGW from the site-external network, forwarding is allowed only to the site-internal network.

EVPN Multi-Site architecture allows selective rate limiting for BUM traffic classes that are known to saturate network infrastructure during broadcast storms, loops, and other traffic-generating failure scenarios. The BGW provides the capability to enforce these traffic classes individually through a rate limiter. Only traffic leaving the local site following termination and reorigination within the BGW will be enforced. The BUM enforcement takes place before the traffic is reoriginated on the BGW for transmission to a remote site.

As of Cisco NX-OS 7.0(3)I7(1) for the Cisco Nexus 9000 Series EX- and FX-platform switches, the classification and rate limiting are applied globally to each BGW. The configured rate-limiting level represents the amount of BUM traffic allowed from each interface that faces the site-external network.

`evpn storm-control broadcast level 0-100`	Define storm control for EVPN Multi-Site Layer 2 extension. The percentage can be adjusted from 0% (block all classified traffic) to 100% (allow all classified traffic).
`evpn storm-control multicast level 0-100`	**Note:** The classification and use of storm control for EVPN Multi-Site architecture is comparable to that for storm control on a physical Layer 2 interface.
`evpn storm-control unicast level 0-100`	

External connectivity

In an EVPN Multi-Site environment, the requirement for external connectivity is as relevant as the requirement for extension between sites. External connectivity includes the connection of the data center

to the rest of the network: to the Internet, the WAN, or the campus. All options provided for external connectivity are multitenant aware and focus on Layer 3 transport to the external network domains.

This document discusses two models for providing external connectivity to EVPN Multi-Site architecture:

- With the placement of the BGWs at the border between the site-internal and site-external domains, a set of nodes is already available at each site that can provide encapsulation and decapsulation for transit traffic. In addition to the EVPN Multi-Site functions, the BGW allows coexistence of VRF-aware connectivity with VRF-lite.
- In addition to per-BGW or per-site external connectivity, connectivity can be provided through a shared border. In this case, a dedicated set of border nodes are placed at the site-external portion of multiple sites. All of these sites connect through VXLAN BGP EVPN to this shared border set, which then provides external connectivity. The shared-border approach also allows MPLS L3VPN, LISP, or VRF-lite hand-off to multiple sites.

VXLAN BGP EVPN provides optimal egress route optimization using the distributed IP anycast gateway function at every VTEP. This optimization is achieved by equipping every VTEP with a first-hop gateway and the information needed to take the best path to a given destination.

With stretched IP subnets across multiple sites, the explicit location of a subnet becomes unclear, and more granular information must be provided in the routing tables. Both the external connectivity models mentioned here allow ingress route optimization by VXLAN BGP EVPN through host-route advertisement (/32 and /128).

In the shared-border model, additional ingress route optimization can be applied depending on the platform used. This topic is discussed in greater detail in the “Shared border” section.

VRF-lite coexistence

The VRF-lite coexistence model (Figure 20) uses the traditional approach to providing external connectivity to a VXLAN BGP EVPN fabric. In particular, this model uses the approach of interautonomous system option A, in which the site-internal network uses MP-BPG with VPN address families. Interautonomous system option A requires the presence of a route distinguisher and route target, although in VRF-lite these would not normally be necessary. For the purposes here, this document uses the terms “VRF-lite” and “interautonomous system option A” interchangeably. For external connectivity, the use of physical Layer 3 interfaces is preferred, with each interface in a separate VRF instance. To use multiple VRF instances on a single physical Layer 3 interface, the use of subinterfaces is recommended.

Note: The EVPN Multi-Site BGW with VRF-lite coexistence is supported starting NX-OS 7.0(3)I7(3)

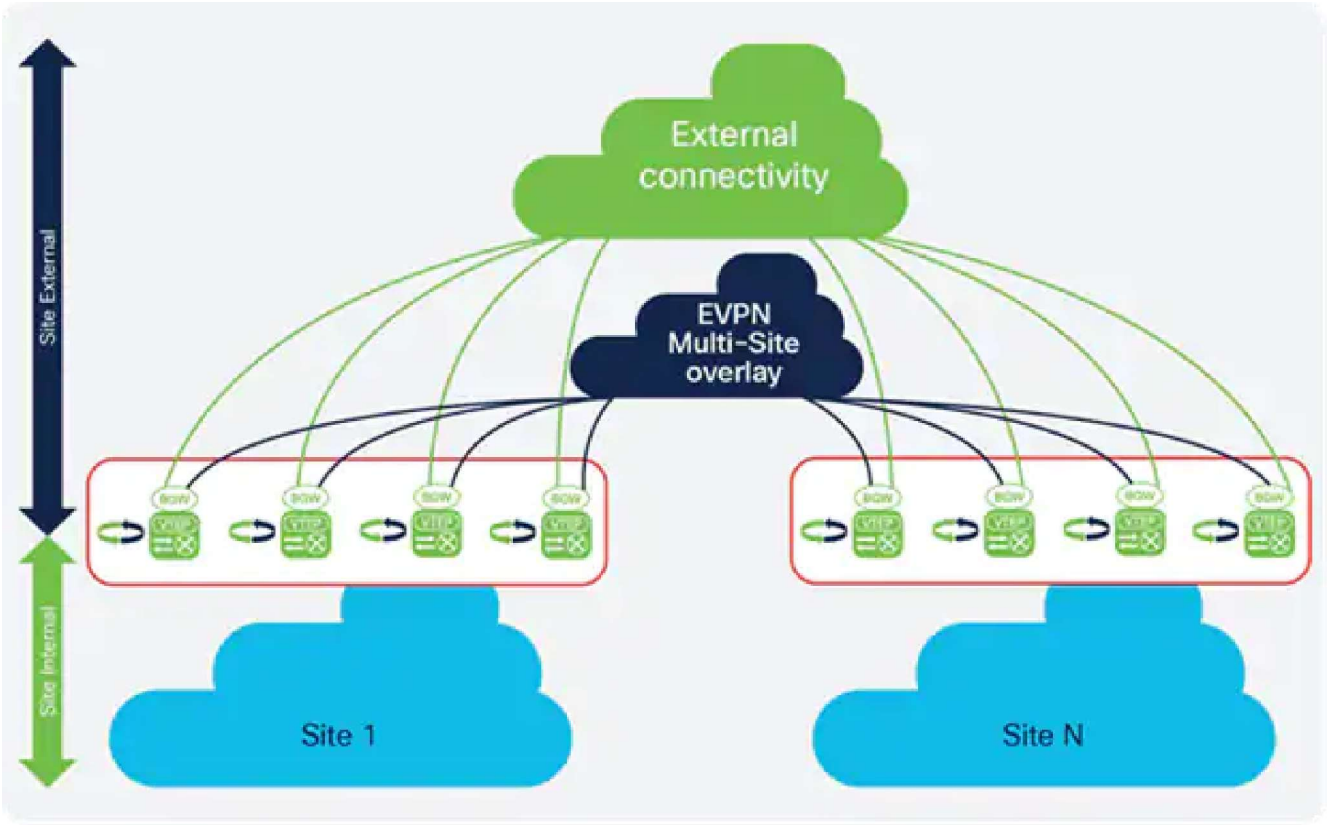


Figure 20.

VRF-lite coexistence

Note: The EVPN Multi-Site BGW does not support the coexistence of external connectivity with IEEE 802.1q tagged Layer 2 interfaces (trunk) and SVIs (interface VLAN), either with or without vPC. More generally, SVIs cannot currently be defined on the BGW.

Because BGP is already in use for EVPN and EVPN Multi-Site architecture, it is the recommended option for exchanging routing information with external routers (VRF-lite external connectivity with the use of a subinterface). Dynamic routing protocols and static routing can also be used, but as a best practice the eBGP approach for VRF-lite coexistence on the BGW is preferred. The physical Layer 3 interface for external connectivity must be dedicated and can't be shared with the site-external connectivity for EVPN Multi-Site architecture.

`vrf context BLUE`
`  vni 50001`
`  rd auto`
 	
`address-family ipv4 unicast`
	Verify that the VRF context (IP VRF instance) with the appropriate instance name has been prepared. The correct Layer 3 VNIs, address families, and route targets must be defined to allow the site-internal VTEPs to have external connectivity.
	Note: For the external connectivity, interautonomous system option A and route distinguishers and route targets are required for the site-internal VXLAN BGP EVPN control plane.

```
route-
target
both
auto

route-
target
both
auto
evpn

address-
family
ipv6
unicast

route-
target
both
auto

route-
target
both
auto
evpn
```

Note: Selective advertisement is defined by the configuration of the per-tenant information on the BGW. In cases in which external connectivity (VRF-lite) and EVPN Multi-Site architecture both are active on the same BGW, the advertisements are always enabled. If this behavior is not desired, you should consider using a dedicated border for external connectivity and EVPN Multi-Site architecture.

```
interface
Ethernet1/3.4

encapsulation
dot1q 4
  vrf member
BLUE
  ip address
10.55.21.1/30
```

Define a Layer 3 subinterface associated with the previously defined VRF, with a point-to-point subnet and IEEE 802.1q tag (VLAN id). This interface connects to the external router.

Note: The VLAN ID and point-to-point subnet must match the neighboring interface. The subinterface ID doesn't need to match the VLAN ID, but consistency is recommended to simplify troubleshooting.

`router bgp 65501` `  vrf BLUE`	Define the VRF instance in the BGP instance.

`    address-family ipv4 unicast` `      advertise l2vpn evpn`	Extend the VRF instance in the BGP instance with the IPv4/IPv6 unicast address family and enable it for EVPN. **Note:** The IPv6 unicast address family is not shown, but it follows same configuration process.

`    neighbor 10.55.21.2`
`      remote-as 65099`
`      update-source Ethernet1/3.4`
`      address-family ipv4 unicast`
	Create the eBGP peering with the neighbor autonomous system and the relevant source interface. Enable the IPv4 unicast address family for this peering.
	Note: The IPv6 unicast address family is not shown, but it follows the same configuration process.

In addition to using route peering to the external router through eBGP, you may sometimes want to advertise the default route to the fabric. Two methods are used to advertise the default route to the fabric:

- The default route is learned through eBGP from the external router on a per-VRF basis. This default route is automatically passed through the BGW and advertised to the site-internal VTEPs through BGP EVPN.
- The default route is learned through a static or dynamic routing protocol (not eBGP). This approach requires the BGW to locally originate the default route and inject it into the BGP EVPN control plane facing the site-internal VTEPs.

Figure 21 shows both approaches.

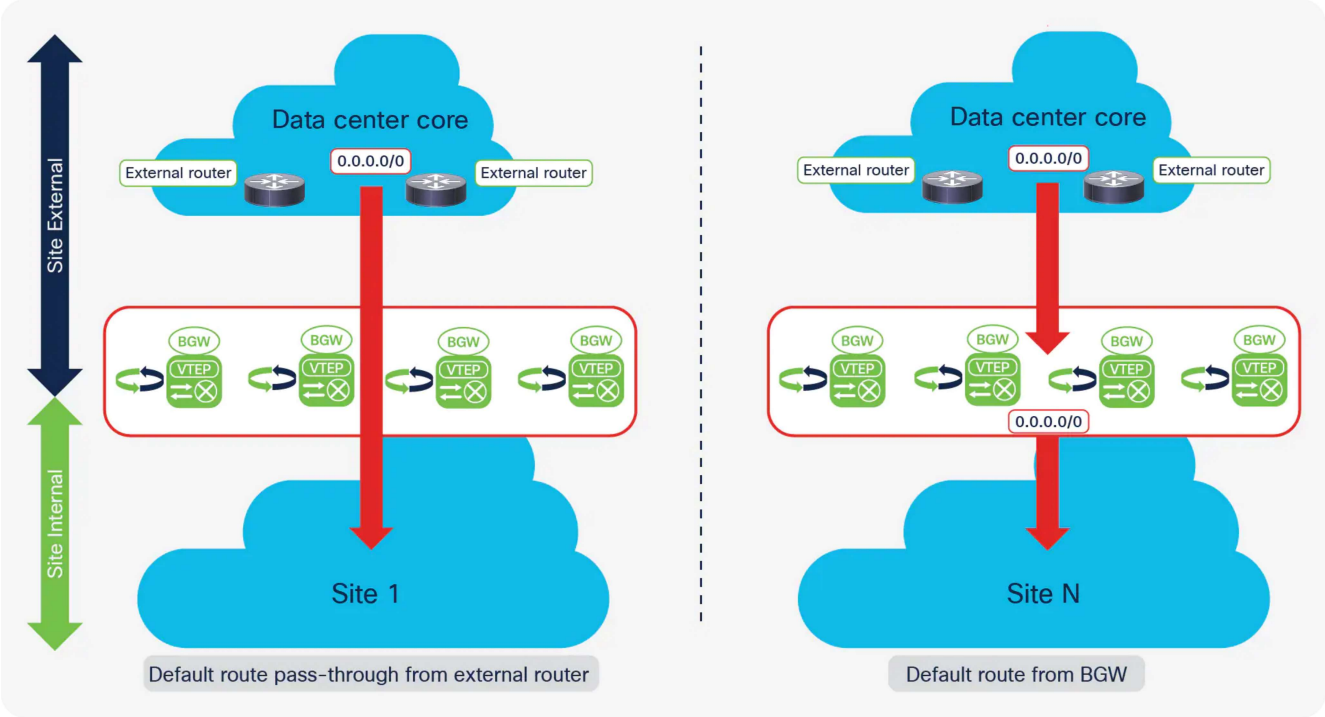


Figure 21.

Default route: External router versus BGW

The first method requires some route filtering to prevent the fabric from becoming a transit network, but no additional configuration is required to receive and advertise the default route to the site-internal VTEPs.

The following configuration example focuses on the second method, using a static route to the external router. The route-filtering configuration example covers both methods.

`vrf context BLUE` `  ip route 0.0.0.0/0 10.55.21.2`	Define a static default route to the next-hop IP address of the external router in the appropriate VRF instance. **Note:** The default route can also be received through a dynamic routing protocol.

`ip prefix-list DEFAULT-ROUTE seq 5 permit 0.0.0.0/0 le 1`	Define a prefix list that matches the default route.

`route-map EXTCON-RMAP-FILTER deny 10` `  match ip address prefix-list DEFAULT-ROUTE`	Define a route map that matches the prefix list, and prevent that match from being advertised to the external connectivity. **Note:** The default route should be advertised only to the site-internal VTEPs.

`route-map EXTCON-RMAP-FILTER permit 1000`	Extend the route map to allow everything that does not match the previous definitions.

`router bgp 65501`
`  vrf BLUE`
 	
`address-family ipv4 unicast`
 	
`network 0.0.0.0/0`
	Define a network statement to advertise the default route to BGP. Because this route is originated locally or learned remotely, it will become an EVPN Route Type 5 route for the site-internal VTEPs.

`    neighbor 10.55.21.2`
`      remote-as 65099`
`      update-source Ethernet1/3.4`
`      address-family ipv4 unicast`
`        route-map EXTCON-RMAP-FILTER out`
	Attach the route filter to the external connectivity peering facing the external router.
	Note: Without the route filter, the VXLAN BGP EVPN fabric can accidentally become a transit network for traffic external to the fabric.

If a single EVPN Multi-Site instance loses external connectivity, but other sites still have external connectivity, EVPN Multi-Site Layer 2 and Layer 3 extension will be used to reach external connectivity for remote sites. If this approach is deemed not beneficial, you can filter external connectivity routes between EVPN Multi-Site fabrics.

In addition to preventing the VXLAN BGP EVPN fabric from becoming a transit network, you can introduce use another optimization through route filtering. The advertisement of host routes (/32 and /128) is performed by default in VXLAN BGP EVPN. This default behavior can be altered by suppressing the host

routes with route summarization at the border facing the external domain or through route filtering (Figure 22).

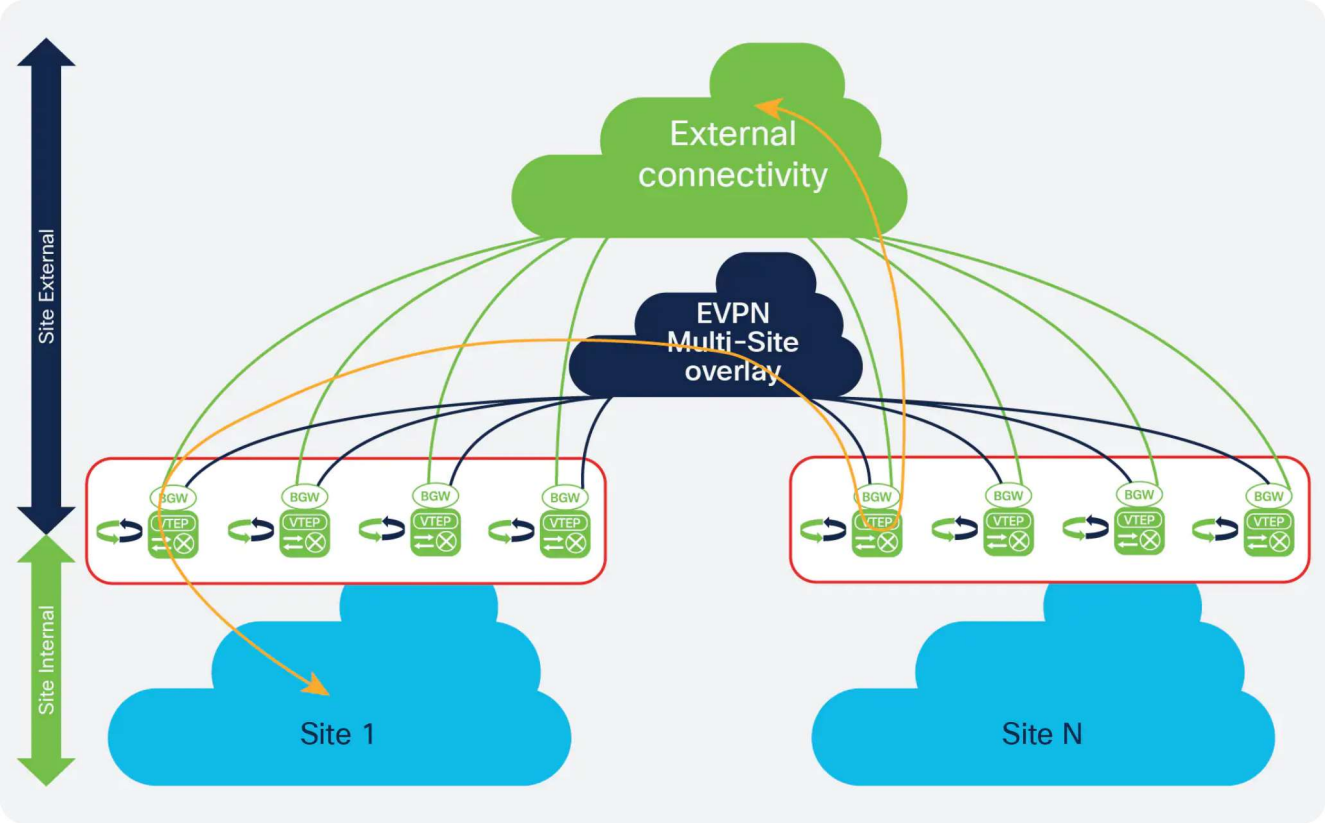


Figure 22.

External connectivity through EVPN Multi-Site

Using the same constructs of the prefix list and route map, you can suppress host routes as shown in the following configuration.

`ip prefix-list HOST-ROUTE seq 5 permit 0.0.0.0/0 eq 32`	Define a prefix list that matches all the host routes. **Note:** IPv6 host-route filtering can be achieved in a similar way.

`route-map EXTCON-RMAP-FILTER deny 20` `  match ip address prefix-list HOST-ROUTE`	Define a route map that matches the prefix list, and prevent that match from being advertised to the external connectivity. **Note:** This route map is an extension of the one previously created for the default route filtering.

`route-map EXTCON-RMAP-`	Extend the route map to allow everything that did not

`FILTER permit 1000`	match the previous definitions.

As a result of the external connectivity configuration, you can route to an external domain, preventing the VXLAN BGP EVPN fabric from becoming a transit network and suppressing host-route advertisements. By disabling host-route advertisements, however, you are not using optimal ingress route optimization. You need to consider this fact when stretching an IP subnet across multiple VXLAN EVPN sites that are extended with EVPN Multi-Site architecture, because ingress routing will then choose any BGW that advertises external connectivity.

Note: The suppression of host routes is not supported between VXLAN BGP EVPN sites that are connected with EVPN Multi-Site architecture. This is specifically the case for the EVPN Multi-Site Layer 2 extension.

Shared border

The shared border acts as a common external connectivity point for multiple VXLAN BGP EVPN fabrics that are interconnected with EVPN Multi-Site architecture. Unlike the BGW, the shared border is completely independent of any VXLAN EVPN Multi-Site software or hardware requirements, it is sloley a border node topologically outside of a single or multiple Sites. The shared border operates like a traditional VTEP, but unlike the site-internal VTEPs discussed previously, the shared border is a site-external VTEP. In the case of external connectivity, the shared border operates solely in Layer 3 mode, and hence no BUM replication between the BGW and shared border nodes is necessary. What you must configure on the shared border is the VXLAN BGP EVPN VTEP and its presence in a different autonomous system than the one that includes the BGWs.

The shared border can enable external connectivity with various Layer 3 technologies, depending on hardware and software capabilities. Some examples are Cisco Nexus 9000 Series Switches (VRF-lite), Cisco Nexus 7000 Series Switches (VRF-lite, MPLS L3VPN, and LISP), Cisco ASR 9000 Series Aggregation Services Routers (VRF-lite and MPLS L3VPN), and Cisco ASR 1000 Series routers (VRF-lite and MPLS L3VPN). This document focuses on the required configuration of the BGW that connects to the shared border. Configuration knobs required on the shared border are discussed, but not the various Layer 3 hand-off technologies for external connectivity.

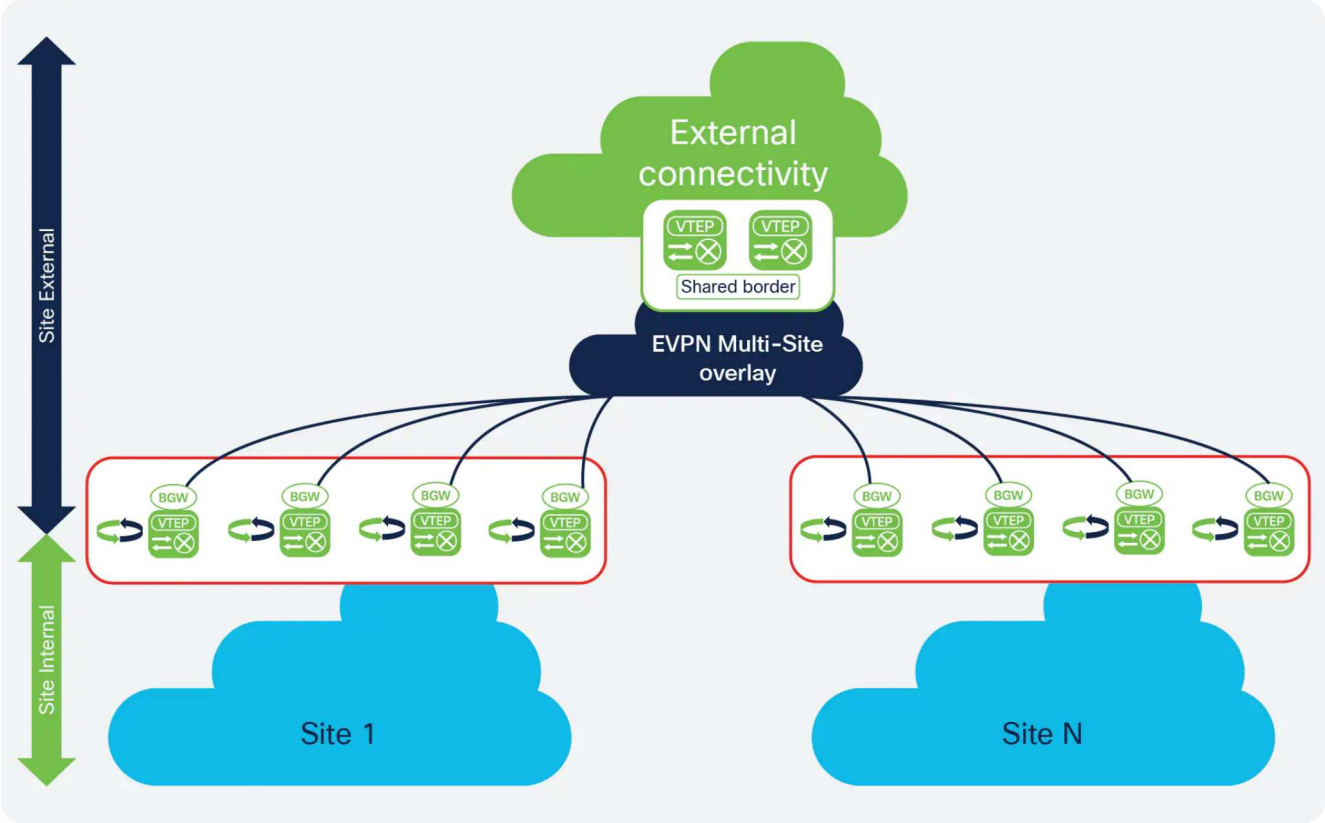


Figure 23.

EVPN Multi-Site shared border

For an EVPN Multi-Site BGW to connect with a shared border, it requires a configuration similar to that for connecting the gateway to the BGW of a remote site (Figure 23). Unlike the EVPN Multi-Site site-external underlay configuration, the configuration of the interface facing the shared border nodes doesn't require interface tracking. It is specifically not necessary to influence the availability of the EVPN Multi-Site virtual IP address, because if the shared border becomes absent, no external routes can be advertised to the site-internal network.

The configuration presented here shows the site-external underlay and overlay configuration on a BGW. The underlay between the BGW and the shared border must be reachable, specifically between the loopback interfaces that provide the VTEP and the overlay peering function. The VXLAN BGP EVPN connectivity between the BGW and the shared border requires a physical Layer 3 interface, as previously discussed for EVPN Multi-Site architecture. For the BGW-to-cloud, BGW-between-spine-and-superspine, and BGW-on-spine deployment models, the existing EVPN Multi-Site site-external underlay interfaces can be used to reach the shared border. When choosing between shared and dedicated external connectivity interfaces, note that you also need to consider your needs for bandwidth and additional resiliency.

BGW to shared border: Site-external eBGP underlay

The configuration for a BGW to a shared border with a site-external eBGP underlay is shown here.

`interface Ethernet1/3`	Define site-external underlay interfaces facing the external Layer 3 core with the shared border present.
`no switchport`	Adjust the MTU setting for the interface to a value that accommodates your environment (the minimum value is 1500 bytes plus VXLAN encapsulation).

`mtu 9216` `ip address 10.55.41.1/30 tag 54321`	Point-to-point IP addressing is used for site-external underlay routing (point-to-point IP addressing with /30 is shown here). The IP address is extended with a tag to allow easy selection for redistribution. **Note:** No EVPN Multi-Site interface tracking (**evpn multisite dci-tracking**) is required for the site-external underlay facing the shared border.

`router bgp 65520`
`router-id 10.101.101.41`
`address-family ipv4 unicast`
`redistribute direct route-map RMAP-REDIST-DIRECT`
`maximum-paths 4`
	Define the BGP routing instance with a site-specific autonomous system.
	Note: The BGP router ID matches the loopback0 IP address.
	Activate the IPv4 unicast global address family (VRF default) to redistribute the required loopback and, if needed, the IP addresses of the physical interfaces within BGP.
	Enable BGP multipathing (**maximum-paths**).
	Note: The redistribution from the locally defined interfaces (direct) into BGP is performed through route-map classification. Only IP addresses in VRF default that are extended with the matching tag of the route map are redistributed.

`neighbor 10.55.41.2`
`remote-as 65099`
`update-source Ethernet1/3`
`address-family ipv4 unicast`
	Configure the neighbor for the IPv4 unicast global address family (VRF default) to facilitate site-external underlay routing.
	eBGP neighbor configuration is performed by specifically selecting the source interface for this eBGP peering.

BGW to shared border: Site-external eBGP overlay

The configuration for a BGW to a shared border with a site-external eBGP overlay is shown here.

`router bgp 65520`	Define the BGP routing instance with a site-specific autonomous system. **Note:** The BGP router ID matches the loopback0 IP address.

`router-id 10.100.100.41`	Configure the neighbor with the EVPN address family (L2VPN EVPN) for the site-external overlay control plane facing the shared border.
`log-neighbor-changes`	eBGP neighbor configuration is performed by specifying the source interface to loopback0. This setting allows underlay ECMP reachability from BGW loopback0 to shared-border loopback0.
	Note: Site-external EVPN peering is always considered to use eBGP with the next hop the shared border.
`neighbor 10.55.55.55`	With the shared border potentially multiple routing hops away, you must increase the BGP session TTL setting to an appropriate value (**ebgp-multihop**).
`remote-as 65099`	When you define the site-external BGP peering session (**peer-type fabric external**), rewrite and reorigination are enabled.
`update-source loopback0`	The autonomous system portion of the automated route target (ASN:VNI) can be rewritten for the site-external network (**rewrite-evpn-rt-asn**) without the need to modify any configuration settings on the shared border. The route-target rewrite helps ensure that the ASN portion of the automated route target matches the destination autonomous system.
`ebgp-multihop 5`	
`peer-type fabric-external`	
`address-family l2vpn evpn`	
`send-community`	
`send-community extended`	
`rewrite-evpn-rt-asn`	

To provide some context for the configuration for a shared border, the following sample shows the settings required to exchange overlay information. The underlay must be reachable between the BGW and the shared border: specifically between the loopback interfaces that provide the VTEP and the overlay peering function.

Shared border to BGW: eBGP underlay

The configuration for a shared border to a BGW with an eBGP underlay is shown here.

`interface Ethernet1/3`	Define site-external underlay interfaces facing the external Layer 3 core with the BGW present.
`mtu 9216`	Adjust the MTU setting for the interface to a value that accommodates your environment (the minimum value is 1500 bytes plus VXLAN encapsulation).

`ip address 10.55.41.2/30 tag 54321`	Point-to-point IP addressing is used for site-external underlay routing (point-to-point IP addressing with /30 is shown here). The IP address is extended with a tag to allow easy selection for redistribution.

`router bgp 65099`	Define the BGP routing instance with a shared-border-specific autonomous system.
`address-family ipv4 unicast`	**Note:** The BGP router ID matches the loopback0 IP address. Activate the IPv4 unicast global address family (VRF default) to redistribute the required loopback and, if needed, the IP addresses of the physical interfaces within BGP.
`redistribute direct route-map RMAP-REDIST-DIRECT`	Enable BGP multipathing (**maximum-paths**). **Note:** The redistribution from the locally defined interfaces (direct) to BGP is performed through route-map classification. Only IP addresses in VRF default that are extended with the matching tag of the route map are redistributed.
`maximum-paths 4`	

`neighbor 10.55.41.1 remote-as 65520`	The neighbor configuration for the IPv4 unicast global address family (VRF default) facilitates shared-border underlay routing. The eBGP neighbor configuration is performed by specifically selecting the source interface for this eBGP peering.
`update-source Ethernet1/3`	
`address-family ipv4 unicast`	

Shared border to BGW: eBGP overlay

The configuration of a shared border to a BGW with an eBGP overlay is shown here.

`router bgp 65099`	Define the BGP routing instance with a site-specific autonomous system.
`address-family ipv4 unicast`	Configure the neighbor with the EVPN address family (L2VPN EVPN) for the site-external overlay control plane facing the BGW. eBGP neighbor configuration is performed by specifying the source interface to loopback0. This setting allows underlay ECMP reachability from BGW loopback0 to shared-border loopback0.
`redistribute direct route-`	**Note:** Site-external EVPN peering is always considered to use eBGP with the next hop the BGW.

`map RMAP-REDIST-DIRECT`
`  maximum-paths 4`
 	
`neighbor 10.101.101.41 remote-as 65520`
`  update-source loopback0`
`  ebgp-multihop 5`
`  address-family l2vpn evpn`
`  rewrite-evpn-rt-asn`
`  send-community both`
	With the BGW potentially multiple routing hops away, you must increase the BGP session TTL setting to an appropriate value (**ebgp-multihop**).
	The autonomous system portion of the automated route target (ASN:VNI) can be rewritten for the site-external network (**rewrite-evpn-rt-asn**) without the need to modify any configuration settings on the BGWs. The route-target rewrite helps ensure that the ASN portion of the automated route target matches the destination autonomous system.

Note: In the shared-border deployment, the BGW of every site must have connectivity to the shared border. Otherwise, routes that VXLAN BGP EVPN learns from a shared border to a BGW will not be advertised to remote sites because the shared border and the remote site BGWs are considered site-external devices.

`interface loopback 51`
`  vrf member BLUE`
`  ip address 10.55.55.1/32`
	Note: In cases where only Layer 3 extension is configured on the BGW, special in the case of Shared Border, an additional loopback interface is required. The loopback interface must be present in the same VRF instance on all BGW and with an individual IP address per BGW. Ensure the loopback interfaces IP address is redistributed into BGP EVPN, specially towards Site-External.

Legacy site integration

For migration and integration purposes, existing non-VXLAN BGP EVPN sites (legacy sites) require connectivity with VXLAN BGP EVPN sites. For integration, a Layer 3–only connectivity model can be used. This approach would allow routing exchange between the different networks, similar to the external

connectivity approach through VRF-lite. Depending on the VRF awareness and number of VRF instances, this option can be acceptable, but the configuration complexity will increase with the number of VRF instances. If Layer 2 extension with same IP subnet between the legacy site and VXLAN EVPN is required, the complexity and dependencies increase, and you must consider IEEE 802.1q trunks for Layer 2 extension, VRF-aware routing for Layer 3, and first-hop gateway consistency.

VXLAN EVPN Multi-Site architecture simplifies legacy site integration and consistently provides the required Layer 2 and Layer 3 extension. Alternative approaches are documented as part of multifabric designs and EVPN-to-Overlay Transport Virtualization (OTV) interoperation solutions. For details, see the "For more information" section at the end of this document.

Similar to the process in the shared-border scenario, the integration of a legacy site is achieved by positioning a set of VTEPs external to the VXLAN BGP EVPN sites (a pair of vPC BGWs). The attributes for a site-external VTEP for such an integration are similar to those for a BGW (VXLAN BGP EVPN, ingress replication for BUM, BUM control, etc.), with the addition of a classic Ethernet multihoming approach (vPC) to connect to the legacy network infrastructure (Figure 24).

Note: vPC is not required by the EVPN Multi-Site architecture but is needed to provide resilient and loop-free connectivity to the legacy site.

Special considerations for Layer 2 extensions apply to BUM control and failure isolation, because the legacy site BGW (vPC BGWs) uses some different (and simplified) configurations given the absence of site-internal VTEPs. The EVPN Multi-Site BUM enforcement feature can be useful. You can apply storm control on the VPC BGW Ethernet interfaces connecting to the site-internal switches. This traditional approach works, but does not allow you to enforce BUM control in an aggregated way. Depending on the number of connections to the legacy network, the BGW may end up allowing more BUM traffic than is desired across the EVPN Multi-Site overlay. When using the BUM enforcement feature within the legacy site BGW, you can enforce aggregated rate limiting based on the well-known BUM traffic classes. This approach allows simpler deployment as well as additional control right before traffic traverses the EVPN Multi-Site overlay.

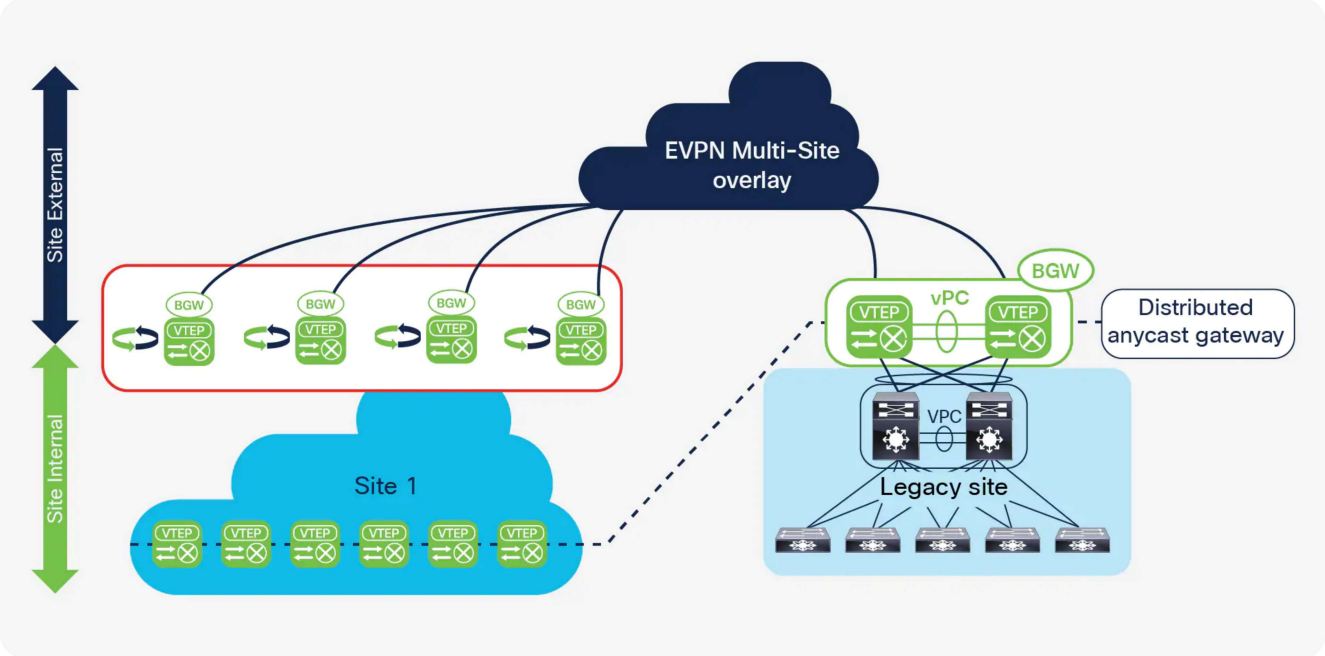


Figure 24.

Legacy site integration

Additional considerations apply to first-hop gateway use and placement. VXLAN BGP EVPN uses the Distributed Anycast Gateway (DAG) as a first-hop gateway, whereas the legacy sites likely use a First-Hop Redundancy Protocol (FHRP) such as Hot Standby Router Protocol (HSRP), Virtual Router Redundancy Protocol (VRRP), or Gateway Load-Balancing Protocol (GLBP). The co-existence of these different first-hop gateway approaches is not supported today, and hence you need to achieve alignment between the legacy sites and VXLAN BGP EVPN sites. For legacy site integration, the BGW is allowed to operate in a vPC domain and to offer the first-hop gateway functions (in this case, DAG). This capability provides a first-hop gateway for the legacy site and helps ensure seamless endpoint mobility between legacy sites and VXLAN BGP EVPN sites.

Note: As of Cisco NX-OS 7.0(3)I7(1), the coexistence of different first-hop gateway modes (such as HSRP and DAG) is not supported for the same network. This restriction applies generally to VXLAN BGP EVPN deployments and is not specific to VXLAN EVPN Multi-Site architecture.

For more information on the use of vPC BGWs to integrate legacy networks with VXLAN EVPN fabrics, including a detailed description of the supported use cases and configuration exmaples, please refer to the “NextGen DCI with VXLAN EVPN Multi-Site Using vPC Border Gateways White Paper” available in the “For more information” section at the end of this document.

Network services integration

Network services integration is a big topic, especially when multiple sites are present and you need to distribute firewalls and load balancers across them. The EVPN Multi-Site BGW generally supports connection of network services (L4-L7 services) such as firewalls, load balancers, and Intrusion Detection System (IDS) and Intrusion Prevention System (IPS) applications. As of Cisco NX-OS 7.0(3)I7(1), all connectivity to the BGW must be implemented through a Layer 3 physical interface or subinterface. If the desired network services deployment can be achieved through routing and routing redundancy, EVPN Multi-Site architecture also supports these connectivity models. For cases in which Layer 2 redundancy, for instance, the use of vPC, is required, connectivity to the EVPN Multi-Site BGW is not currently supported. Also, connectivity models that use SVI and interface VLANs and IEEE 802.1q tagged Layer 2 interfaces (trunks) are not supported on the BGW.

To deploy network services in this cases, you can use a site-internal VTEP (that is, a services VTEP). Subsequent software releases will extend the capabilities to a BGW.

Network services deployment with EVPN Multi-Site architecture is covered in a separate document.

Verification and show commands

After you set up a VXLAN BGP EVPN Multi-Site environment, you need the tools necessary to verify the current state. This section explores the available show commands and their expected output. All output is based on the topology shown in Figure 25.

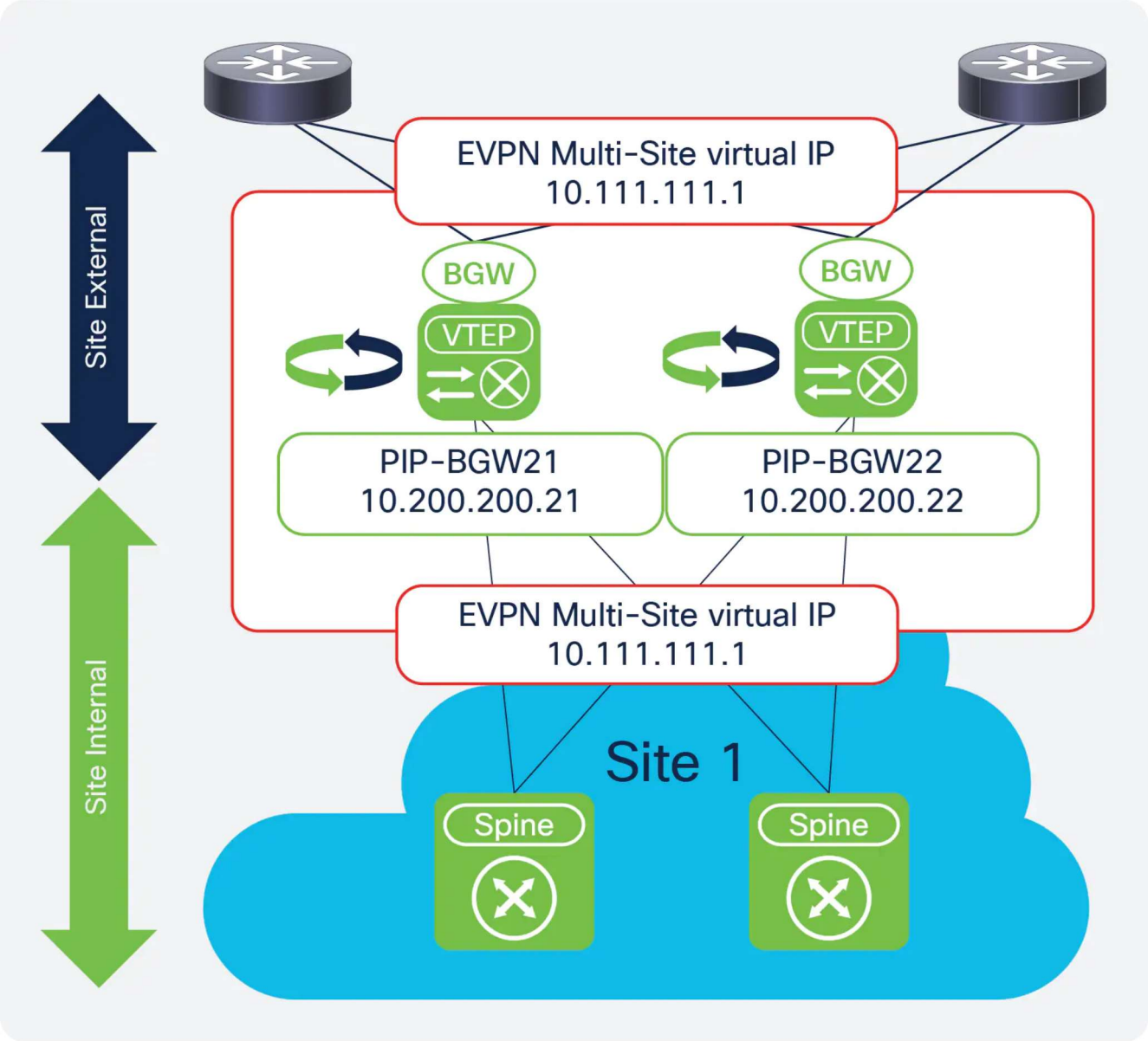


Figure 25.

Show commands and verification

In addition to the **show** commands presented in this section, VXLAN OAM (NGOAM) works consistently for single-site and EVPN Multi-Site architecture. End-to-end VXLAN OAM is supported as of Cisco NX-OS 7.0(3)I7(1).

VTEP interface status

EVPN Multi-Site architecture provides additional status information about the BGW VTEP. The output now includes EVPN Multi-Site architecture configured and elapsed delay-restore time, the virtual router MAC address, and the virtual IP address and status.

```
BGW21-N93180EX# show nve interface nve 1 detail

Interface: nve1, State: Up, encapsulation: VXLAN

 VPC Capability: VPC-VIP-Only [not-notified]

 Local Router MAC: 00a3.8e9d.9267

 Host Learning Mode: Control-Plane
```

```
 Source-Interface: loopback1 (primary: 10.200.200.21, secondary: 0.0.0.0)
 Source Interface State: Up
 IR Capability Mode: No
 Virtual RMAC Advertisement: No
 NVE Flags:
 Interface Handle: 0x49000001
 Source Interface hold-down-time: 180
 Source Interface hold-up-time: 30
 Remaining hold-down time: 0 seconds
 Multi-Site delay-restore time: 180 seconds
 Multi-Site delay-restore time left: 0 seconds
 Virtual Router MAC: 0200.0a6f.6f01
 Interface state: nve-intf-add-complete
 unknown-peer-forwarding: disable
 down-stream vni config mode: n/a
 Multisite bgw-if: loopback100 (ip: 10.111.111.1, admin: Up, oper: Up)
 Multisite bgw-if oper down reason:
Nve Src node last notif sent: None
Nve Mcast Src node last notif sent: None
Nve MultiSite Src node last notif sent: Port-up


BGW21-N93180EX#
```

The EVPN Multi-Site delay-restore function can be triggered either by interface status tracking or by the launch of the BGW itself. The status of the EVPN Multi-Site virtual IP address indicates whether the relevant IP address is active for advertising through the underlay routing protocol.

If all site-external interfaces are down, the EVPN Multi-Site virtual IP address is moved to the operational Down state, and the reasons are shown.

```
BGW21-N93180EX# show nve multisite dci-links
Interface      State
---------      -----
Ethernet1/1    Down
Ethernet1/2    Down


BGW21-N93180EX#
BGW21-N93180EX# show nve interface nve 1 detail
```

```
…
 Multisite bgw-if: loopback100 (ip: 10.111.111.1, admin: Up, oper: Down)
  Multisite bgw-if oper down reason:   DCI isolated.
```

Similarly, if all site-internal interfaces are down, the EVPN Multi-Site virtual IP address is moved to the operational Down state, and the reasons are shown.

```
BGW21-N93180EX# show nve multisite fabric-links
Interface      State
---------      -----
Ethernet1/53   Down
Ethernet1/54   Down

BGW21-N93180EX#
BGW21-N93180EX# show nve interface nve 1 detail
…
 Multisite bgw-if: loopback100 (ip: 10.111.111.1, admin: Up, oper: Down)
  Multisite bgw-if oper down reason:   FABRIC isolated.
```

In addition to verification of the state, control-plane protocol actions are performed as described in the "Failure scenarios" section.

Site-internal and site-external interface status

With EVPN Multi-Site interface tracking, the BGW function and advertisement and participation are controlled. The output provided as part of the interface tracking allows verification of the state.

```
BGW21-N93180EX# show nve multisite dci-links
Interface      State
---------      -----
Ethernet1/1    Down
Ethernet1/2    Up

BGW21-N93180EX# show nve multisite fabric-links
Interface      State
---------      -----
Ethernet1/53   Up
Ethernet1/54   Up

BGW21-N93180EX#
```

Designated forwarder election status

The designated-forwarder election status can be viewed per BGW and per VLAN and L2VNI. The output shows the status of the overall configured local VLANs (active VLANs), the VLANs for which the local BGW is the designated forwarder (designated-forwarder VLANs), and the mapped Layer 2 VNIs (active VNIs). In addition, a list of all the BGWs viable for designated-forwarder election is shown (designated-forwarder list).

```
BGW21-N93180EX# show nve ethernet-segment


ESI: 0300.0000.0000.0100.0309
   Parent interface: nve1
  ES State: Up
  Port-channel state: N/A
  NVE Interface: nve1
   NVE State: Up
   Host Learning Mode: control-plane
 Active Vlans: 1,10,2003
  DF Vlans: 10
  Active VNIs: 30010,50001
  CC failed for VLANs:
  VLAN CC timer: 0
  Number of ES members: 2
  My ordinal: 0
  DF timer start time: 00:00:00
  Config State: N/A
 DF List: 10.200.200.21 10.200.200.22
  ES route added to L2RIB: True
  EAD/ES routes added to L2RIB: False
  EAD/EVI route timer age: not running
--------------------------------------
```

Note: As of Cisco NX-OS 7.0(3)I7(1), the Layer 3 VNI is always shown as active on all BGWs because designated-forwarder election is not performed. The same status applies for the VLAN that is mapped to the L3VNI.

Designated-forwarder message exchange

In addition to the designated-forwarder election status, you can display the specific designated-forwarder election messages. For EVPN Multi-Site architecture, BGP EVPN Route Type 4 is used to

perform designated-forwarder election. The output shows all the BGP EVPN route Type 4 instances that are learned on a given node with the relevant Ethernet Segment (ES) as the site ID and the origin's BGW PIP address.

```
BGW21-N93180EX# show bgp l2vpn evpn route-type 4

BGP routing table information for VRF default, address family L2VPN EVPN

Route Distinguisher: 10.100.100.21:27001   (ES [0300.0000.0000.0100.0309 0])

BGP routing table entry for [4]:[0300.0000.0000.0100.0309]:[32]:
[10.200.200.21]/136, version 59722

Paths: (1 available, best #1)

Flags: (0x000002) on xmit-list, is not in l2rib/evpn


  Advertised path-id 1

  Path type: local, path is valid, is best path

  AS-Path: NONE, path locally originated

    10.200.200.21 (metric 0) from 0.0.0.0 (10.100.100.21)

      Origin IGP, MED not set, localpref 100, weight 32768

      Extcommunity: ENCAP:8 RT:0000.0000.0001


  Path-id 1 advertised to peers:

    10.52.52.52        10.53.53.53        10.100.100.201     10.100.100.202

BGP routing table entry for [4]:[0300.0000.0000.0100.0309]:[32]:
[10.200.200.22]/136, version 59736

Paths: (1 available, best #1)

Flags: (0x000012) on xmit-list, is in l2rib/evpn, is not in HW


  Advertised path-id 1

  Path type: internal, path is valid, is best path

             Imported from 10.100.100.22:27001:[4]:[0300.0000.0000.0100.0309]:
[32]:[10.200.200.22]/136

  AS-Path: NONE, path sourced internal to AS

    10.200.200.22 (metric 3) from 10.100.100.201 (10.100.100.201)

      Origin IGP, MED not set, localpref 100, weight 0

      Extcommunity: ENCAP:8 RT:0000.0000.0001

      Originator: 10.100.100.22 Cluster list: 10.100.100.201


  Path-id 1 not advertised to any peer
```

The important part of this output is not its detailed information, but the fact that one BGP EVPN route type 4 prefix must exist for each BGW at the local site. Thus, in the case of two BGWs, you need two prefixes in every BGW: one local to the BGW and one received remotely.

The preceding example shows a site with two BGWs. The BGW with PIP address 10.200.200.21 is local to the show output, and the BGW with PIP address 10.200.200.22 is local to the site and the prefix was received by the BGP EVPN.

For more information

Additional documentation about EVPN Multi-Site architecture and related topics can be found at the sites listed here.

Configuration guides and examples

Configuring VXLAN EVPN Multi-Site architecture (Cisco Nexus 9000 Series Switches):
https://www.cisco.com/c/en/us/td/docs/switches/datacenter/nexus9000/sw/7-x/vxlan/configuration/guide/b_Cisco_Nexus_9000_Series_NX-OS_VXLAN_Configuration_Guide_7x/b_Cisco_Nexus_9000_Series_NX-OS_VXLAN_Configuration_Guide_7x_chapter_01100.html

Configuring VXLAN BGP EVPN (Cisco Nexus 9000 Series Switches):
https://www.cisco.com/c/en/us/td/docs/switches/datacenter/nexus9000/sw/7-x/vxlan/configuration/guide/b_Cisco_Nexus_9000_Series_NX-OS_VXLAN_Configuration_Guide_7x/b_Cisco_Nexus_9000_Series_NX-OS_VXLAN_Configuration_Guide_7x_chapter_0100.html

VXLAN EVPN configuration example (Cisco Nexus 9000 Series Switches):
https://communities.cisco.com/community/technology/datacenter/data-center-networking/blog/2015/05/19/vxlanevpn-configuration-example

Cisco programmable fabric with VXLAN BGP EVPN configuration guide:
https://www.cisco.com/c/en/us/td/docs/switches/datacenter/pf/configuration/guide/b-pf-configuration.html

Solution overviews

Building hierarchical fabrics with VXLAN EVPN Multi-Site architecture:
https://www.cisco.com/c/dam/en/us/products/collateral/switches/nexus-9000-series-switches/at-a-glance-c45-739422.pdf

VXLAN innovations: VXLAN EVPN Multi-Site architecture (part 2 of 2):
https://blogs.cisco.com/datacenter/vxlan-innovations-vxlan-evpn-multi-site-part-2-of-2

Design considerations and related references

The magic of superspines and RFC-7938 with overlays:
https://learningnetwork.cisco.com/blogs/community_cafe/2017/10/17/the-magic-of-super-spines-and-rfc7938-with-overlays-guest-post

draft-sharma-multi-site-evpn - Multi-site EVPN based VXLAN using BGWs

https://tools.ietf.org/html/draft-sharma-multi-site-evpn

RFC-7432 (BGP MPLS-based Ethernet VPN): https://tools.ietf.org/html/rfc7432

draft-ietf-bess-evpn-overlay (network virtualization overlay solution using EVPN): https://tools.ietf.org/html/draft-ietf-bess-evpn-overlay

draft-ietf-bess-evpn-inter-subnet-forwarding (integrated routing and bridging in EVPN): https://tools.ietf.org/html/draft-ietf-bess-evpn-inter-subnet-forwarding

draft-ietf-bess-evpn-prefix-advertisement - IP Prefix Advertisement in EVPN

https://tools.ietf.org/html/draft-ietf-bess-evpn-prefix-advertisement

RFC-7947 (Internet exchange BGP route server): https://tools.ietf.org/html/rfc7947

BRKDCN-2035 (VXLAN BGP EVPN-based multipod, multifabric, and multisite architecture): https://www.ciscolive.com/online/connect/sessionDetail.ww?SESSION_ID=95611

BRKDCN-2125 (overlay management and visibility with VXLAN): https://www.ciscolive.com/online/connect/sessionDetail.ww?SESSION_ID=95613

Building data centers with VXLAN BGP EVPN (Cisco NX-OS perspective): https://www.ciscopress.com/store/building-data-centers-with-vxlan-bgp-evpn-a-cisco-nx-9781587144677

VXLAN BGP EVPN multifabric: https://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-738358.html

VXLAN BGP EVPN and OTV interoperation (Cisco Nexus 7000 Series and 7700 platform switches): https://www.cisco.com/c/en/us/td/docs/switches/datacenter/nexus7000/sw/vxlan/config/cisco_nexus7000_vxlan_config_guide_8x/cisco_nexus7000_vxlan_config_guide_8x_chapter_01001.html