



Cisco ACI 転送

- ・ [ファブリック内での転送 \(1 ページ\)](#)

ファブリック内での転送

ACI ファブリックは現代のデータセンタートラフィックフローを最適化する

Cisco ACI アーキテクチャは、従来のデータセンター設計から来る制限を解放して、最新のデータセンターで増大する East-West トラフィックの需要に対応します。

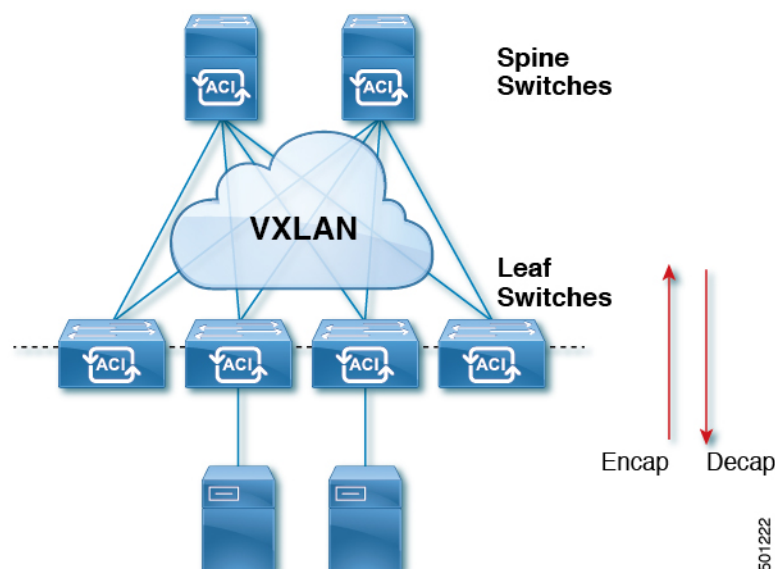
今日のアプリケーション設計は、データセンターのアクセスレイヤを通る、サーバ間の East-West トラフィックを増大させています。このシフトを促進しているアプリケーションには、Hadoop のようなビッグデータの分散処理の設計、VMware vMotion のようなライブの仮想マシンまたはワークロードの移行、サーバのクラスタリング、および多層アプリケーションなどが含まれます。

North-South トラフィックは、コア、集約、およびアクセスレイヤ、またはコラプストコアとアクセスレイヤが重要となる、従来型のデータセンター設計を推進します。クライアントデータは WAN またはインターネットで受信され、サーバの処理を受けた後、データセンターを出ます。このような方式のため、WAN またはインターネットの帯域幅の制限により、データセンターのハードウェアは過剰設備になりがちです。ただし、スパニングツリープロトコルが、ループをブロックするために要求されます。これは、ブロックされたリンクにより利用可能な帯域幅を制限し、トラフィックが準最適なパスを通るように強制する可能性があります。

従来のデータセンター設計においては、IEEE 802.1Q VLAN がレイヤ 2 境界の論理セグメンテーションまたはブロードキャストドメインを提供します。ただし、ネットワークリンクの VLAN の使用は効率的ではありません。データセンターネットワークでデバイスの配置要件は柔軟性に欠け、VLAN の最大値である 4094 の VLAN が制限となり得ます。IT 部門とクラウドプロバイダが大規模なマルチテナントデータセンターを構築するようになるにつれ、VLAN の制限は問題となりつつあります。

スパインリーフアーキテクチャは、これらの制限に対処します。ACI ファブリックは、外界からは、ブリッジングとルーティングが可能な単一のスイッチに見えます。レイヤ3のルーティングをアクセスレイヤに移動すると、最新のアプリケーションが必要としている、レイヤ2の到達可能性が制限されます。仮想マシンワークロードモビリティや一部のクラスタリングのソフトウェアのようなアプリケーションは、送信元と宛先のサーバ間がレイヤ2で隣接していることを必要とします。アクセスレイヤでルーティングを行えば、トランクダウンされた同じVLANの同じアクセススイッチに接続したサーバだけが、レイヤ2で隣接します。ACIでは、VXLANが、基盤となるレイヤ3ネットワークインフラストラクチャからレイヤ2のドメインを切り離すことにより、このジレンマを解決します。

図 1: ACI ファブリック



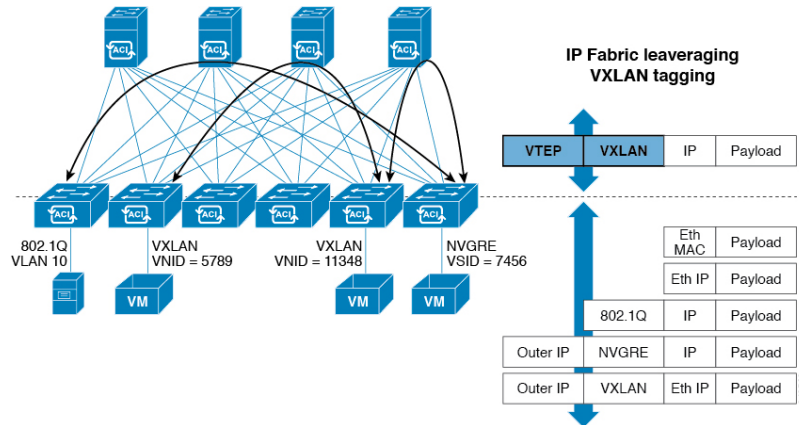
トラフィックがファブリックに入ると、ACIがカプセル化してポリシーを適用し、必要に応じてスパインスイッチ(最大2ホップ)によってファブリックを通過させ、ファブリックを出るときにカプセル化を解除します。ファブリック内では、ACIはエンドポイント間通信でのすべての転送について、Intermediate System-to-Intermediate System プロトコル (IS-IS) および Council of Oracle Protocol (COOP) を使用します。これにより、すべての ACI リンクがアクティブで、ファブリック内での等コストマルチパス (ECMP) 転送と高速再コンバージョンが可能になります。ファブリック内と、ファブリックの外部のルータ内でのソフトウェア定義ネットワーク間のルーティング情報を伝播するために、ACIはマルチプロトコル Border Gateway Protocol (MP-BGP) を使用します。

ACI で VXLAN

VXLANは、レイヤ2オーバーレイの論理ネットワークを構築するレイヤ3のインフラストラクチャ上でレイヤ2のセグメントを拡張する業界標準プロトコルです。ACIインフラストラクチャレイヤ2ドメインが隔離ブロードキャストと障害ブリッジドメインをオーバーレイ内に存在します。このアプローチは大きすぎる、障害ドメインの作成のリスクなしで大きくなるデータセンターネットワークを使用できます。

すべてのトラフィック、ACIファブリックはVXLANパケットとして正規化されます。入力でACI VXLANパケットで外部VLAN、VXLAN、およびNVGREパケットをカプセル化します。次の図は、ACIカプセル化の正規化を示します。

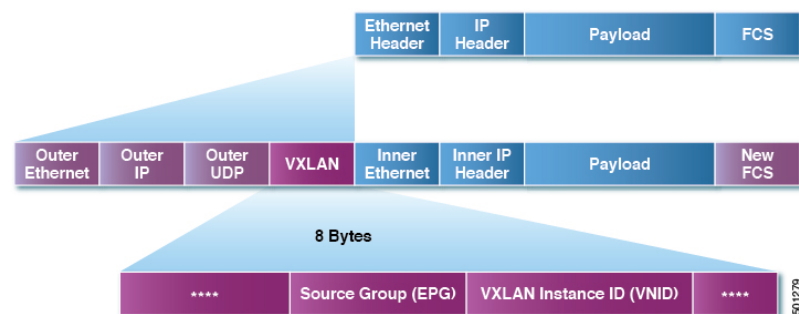
図 2: ACIカプセル化の正規化



ACIファブリックでの転送は、カプセル化のタイプまたはカプセル化のオーバーレイネットワークによって制限または制約されません。ACIブリッジドメインのフォワーディングポリシーは、必要な場合に標準のVLAN動作を提供するために定義できます。

ファブリック内のすべてのパケットにACIポリシー属性が含まれているため、ACIは完全に分散された方法でポリシーを一貫して適用できます。ACIにより、アプリケーションポリシーのEPG IDが転送から分離されます。次の図に示すように、ACI VXLANヘッダーは、ファブリック内のアプリケーションポリシーを特定します。

図 3: ACI VXLANのパケット形式



ACI VXLANパケットには、レイヤ2のMACアドレスとレイヤ3 IPアドレスの送信元と宛先フィールド、ファブリック内の効率的な拡張性の転送を有効にします。ACI VXLANパケットヘッダーの送信元グループフィールドは、パケットが属するアプリケーションポリシーエンドポイントグループ (EPG) を特定します。VXLAN インスタンス ID (VNID) は、テナントの仮想ルーティングおよび転送 (VRF) ドメインファブリック内で、パケットの転送を有効にします。VXLANヘッダーで24ビットVNIDフィールドでは、同じネットワークで一意的レイヤ2のセグメントを最大16個の拡張アドレス空間を提供します。この拡張アドレス空間は、大規模なマルチテナントデータセンターを構築する柔軟性IT部門とクラウドプロバイダーを提供します。

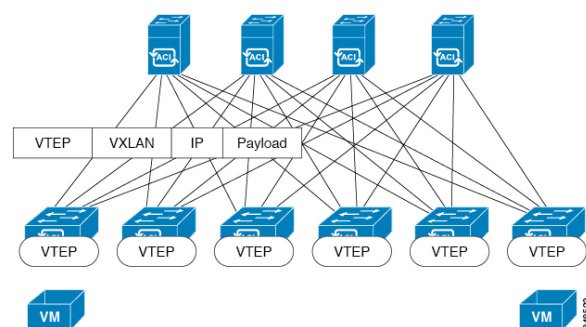
VXLANを有効にACIファブリック全体にわたってスケールでの仮想ネットワークインフラストラクチャのレイヤ3のアンダーレイレイヤ2を展開します。アプリケーションエンドポイントホスト柔軟に配置できます、アンダーレイインフラストラクチャのレイヤ3バウンダリのリスクなしでデータセンターネットワーク間をオーバーレイネットワーク、VXLANでレイヤ2の隣接関係を維持します。

サブネット間のテナントトラフィックの転送を促進するレイヤ3VNID

ACIファブリックは、ACIファブリックVXLANネットワーク間のルーティングを実行するテナントのデフォルトゲートウェイ機能を備えています。各テナントに対して、ファブリックはテナントに割り当てられたすべてのリーフスイッチにまたがる仮想デフォルトゲートウェイを提供します。これは、エンドポイントに接続された最初のリーフスイッチの入力インターフェイスで提供されます。各入力インターフェイスはデフォルトゲートウェイインターフェイスをサポートします。ファブリック全体のすべての入力インターフェイスは、特定のテナントサブネットに対して同一のルータのIPアドレスとMACアドレスを共有します。

ACIファブリックは、エンドポイントのロケータまたはVXLANトンネルエンドポイント(VTEP)アドレスで定義された場所から、テナントエンドポイントアドレスとその識別子を切り離します。ファブリック内の転送はVTEP間で行われます。次の図は、ACIで切り離されたIDと場所を示します。

図4: ACIによって切り離されたIDと場所



VXLANはVTEPデバイスを使用してテナントのエンドデバイスをVXLANセグメントにマッピングし、VXLANのカプセル化およびカプセル化解除を実行します。各VTEP機能には、次の2つのインターフェイスがあります。

- ブリッジングを介したローカルエンドポイント通信をサポートするローカルLANセグメントのスイッチインターフェイス
- 転送IPネットワークへのIPインターフェイス

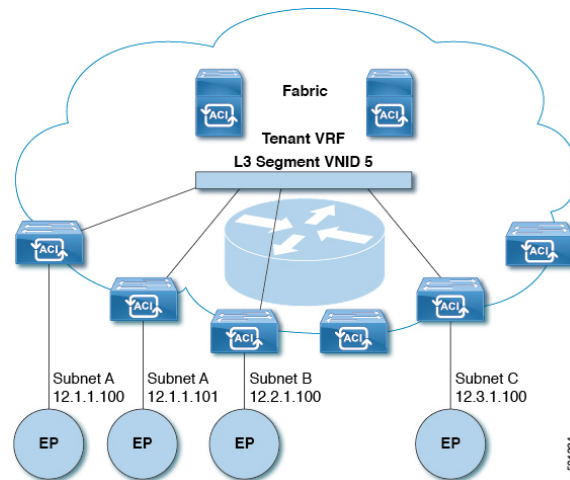
IPインターフェイスには一意のIPアドレスがあります。これは、インフラストラクチャVLANとして知られる、転送IPネットワーク上のVTEPを識別します。VTEPデバイスはこのIPアドレスを使用してイーサネットフレームをカプセル化し、カプセル化されたパケットを、IPインターフェイスを介して転送ネットワークへ送信します。また、VTEPデバイスはリモートVTEPでVXLANセグメントを検出し、IPインターフェイスを介してリモートのMAC Address-to-VTEP マッピングについて学習します。

ACI の VTEP は分散マッピング データベースを使用して、内部テナントの MAC アドレスまたは IP アドレスを特定の場所にマッピングします。VTEP はルックアップの完了後に、宛先リーフスイッチ上の VTEP を宛先アドレスとして、VXLAN 内でカプセル化された元のデータ パケットを送信します。宛先リーフスイッチはパケットをカプセル化解除して受信ホストに送信します。このモデルにより、ACI はスパニングツリー プロトコルを使用することなく、フルメッシュでシングル ホップのループフリー トポロジを使用してループを回避します。

VXLAN セグメントは基盤となるネットワーク トポロジに依存しません。逆に、VTEP 間の基盤となる IP ネットワークは、VXLAN オーバーレイに依存しません。これは送信元 IP アドレスとして開始 VTEP を持ち、宛先 IP アドレスとして終端 VTEP を持っており、外部 IP アドレス ヘッダーに基づいてパケットをカプセル化します。

次の図は、テナント内のルーティングがどのように行われるかを示します。

図 5: ACI のサブネット間のテナントトラフィックを転送するレイヤ 3 VNID



ACI はファブリックの各テナント VRF に単一の L3 VNID を割り当てます。ACI は、L3 VNID に従ってファブリック全体にトラフィックを転送します。出力リーフスイッチでは、ACI によって L3 VNID からのパケットが出力サブネットの VNID にルーティングされます。

ACI のファブリック デフォルト ゲートウェイに送信されてファブリック入力に到達したトラフィックは、レイヤ 3 VNID にルーティングされます。これにより、テナント内でルーティングされるトラフィックはファブリックで非常に効率的に転送されます。このモデルを使用すると、たとえば同じ物理ホスト上の同じテナントに属し、サブネットが異なる 2 つの VM 間では、トラフィックが (最小パス コストを使用して) 正しい宛先にルーティングされる際に経由する必要があるは入力スイッチ インターフェイスのみです。

ACI ルート リフレクタは、ファブリック内での外部ルートの配布にマルチプロトコル BGP (MP-BGP) を使用します。ファブリック管理者は自律システム (AS) 番号を提供し、ルート リフレクタにするスパインスイッチを指定します。



- (注) Cisco ACI は IP フラグメンテーションをサポートしていません。したがって、外部ルータへのレイヤ3 Outside (L3Out) 接続、または Inter-Pod Network (IPN) を介したマルチポッド接続を設定する場合は、インターフェイス MTU がリンクの両端で適切に設定することを推奨します。Cisco ACI、Cisco NX-OS、および Cisco IOS などの一部のプラットフォームでは、設定可能な MTU 値はイーサネットヘッダー (一致する IP MTU、14-18 イーサネットヘッダーサイズを除く) を考慮していません。また、IOS XR などの他のプラットフォームには、設定された MTU 値にイーサネットヘッダーが含まれています。設定された値が 9000 の場合、Cisco ACI、Cisco NX-OS および Cisco IOS の最大 IP パケットサイズは 9000 バイトになりますが、IOS-XR のタグなしインターフェイスの最大 IP パケットサイズは 8986 バイトになります。
- 各プラットフォームの適切な MTU 値については、それぞれの設定ガイドを参照してください。
- CLI ベースのコマンドを使用して MTU をテストすることを強く推奨します。たとえば、Cisco NX-OS CLI で、コマンド、`ping 1.1.1.1 df-bit packet-size 9000 source-interface ethernet 1/1` を使用してください。

翻訳について

このドキュメントは、米国シスコ発行ドキュメントの参考和訳です。リンク情報につきましては、日本語版掲載時点で、英語版にアップデートがあり、リンク先のページが移動/変更されている場合がありますことをご了承ください。あくまでも参考和訳となりますので、正式な内容については米国サイトのドキュメントを参照ください。