Cisco ACI 転送

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。

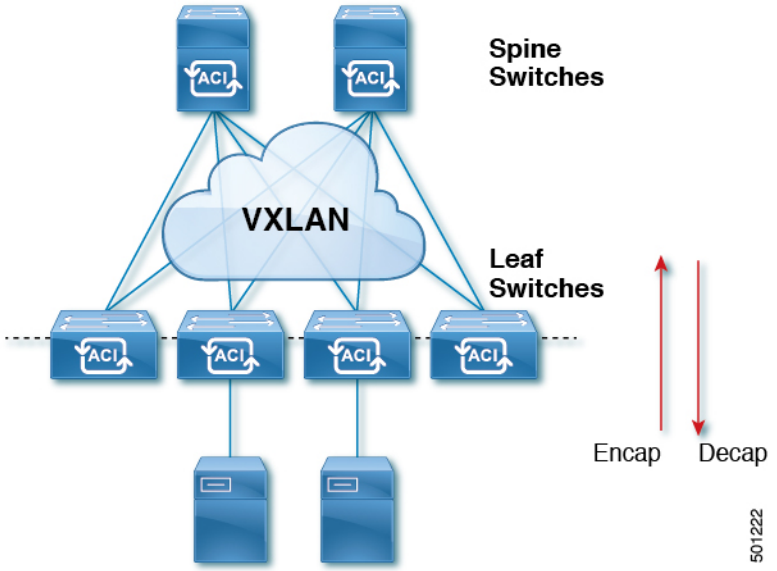
今日のアプリケーション設計は、データセンターのアクセス レイヤを通る、サーバ間の East-West トラフィックを増大させています。このシフトを促進しているアプリケーションには、Hadoop のようなビッグデータの分散処理の設計 、VMware vMotion のようなライブの仮想マシンまたはワークロードの移行、サーバのクラスタ リング、および多層アプリケーションなどが含まれます。

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



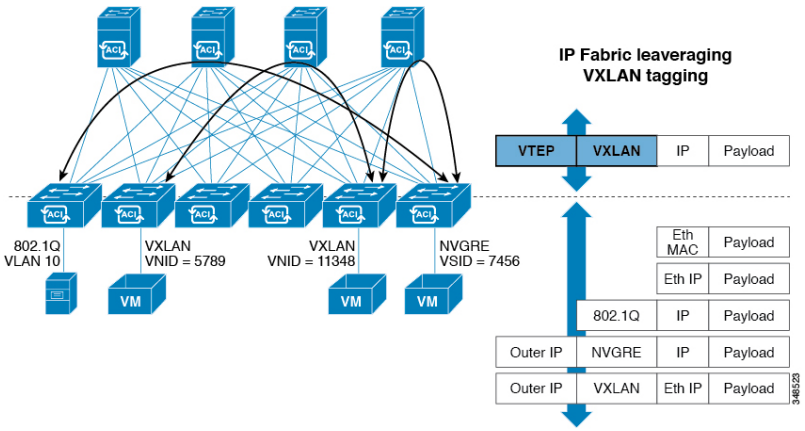
トラフィックがファブリックに入ると、ACIがカプセル化してポリシーを適用し、必要に応じてスパイン スイッチ (最大 2 ホップ) によってファブリックを通過させ、ファブリックを出るときにカプセル化を解除します。ファブリック内では、ACIはエンドポイント間通信でのすべての転送について、Intermediate System-to-Intermediate System プロトコル (IS-IS) および Council of Oracle Protocol (COOP) を使用します。これにより、すべての ACI リンクがアクティブで、ファブリック内での等コストマルチパス (ECMP) 転送と高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。

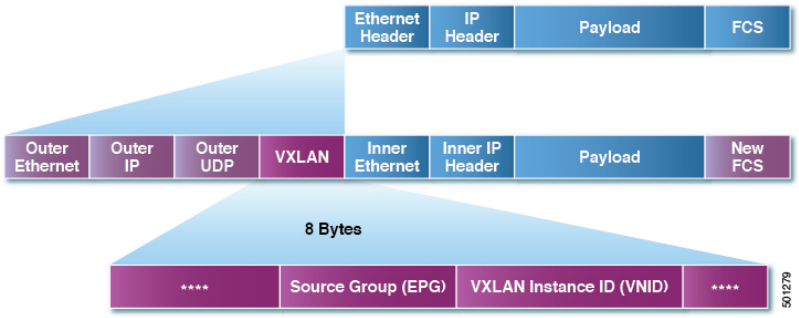
図 *2 : ACI* カプセル化の正規化



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



ACIVXLAN 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と拡張性の転送を有効にします。ACIVXLAN 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。

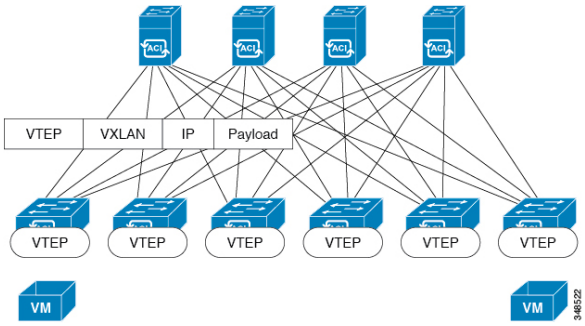
VXLANを有効にACIファブリック全体にわたってスケールでの仮想ネットワーク インフラストラクチャのレイヤ 3 のアンダーレイ レイヤ 2 を展開します。アプリケーション エンドポイント ホスト柔軟に配置できます、アンダーレイ インフラストラクチャのレイヤ 3 バウンダリのリスクなしでデータセンター ネットワーク間をオーバーレイ ネットワーク、VXLANでレイヤ 2 の隣接関係を維持します。

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

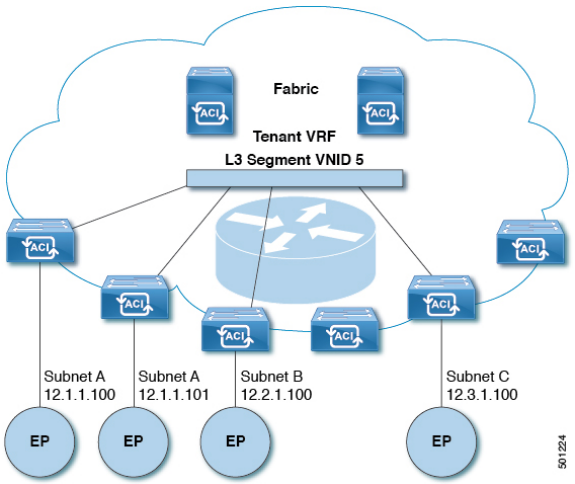
IPインターフェイスには一意のIPアドレスがあります。これは、インフラストラクチャ VLAN として知られる、転送 IP ネットワーク上の VTEP を識別します。VTEP デバイスはこの IP アドレスを使用してイーサネット フレームをカプセル化し、カプセル化されたパケットを、IP インターフェイスを介して転送ネットワークへ送信します。また、VTEP デバイスはリモート VTEP で VXLAN セグメントを検出し、IP インターフェイスを介してリモートの MAC Address-to-VTEP マッピングについて学習します。

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。

IGP プロトコル パケット（EIGRP、OSPFv3）は、インターフェイス MTU サイズに基づいてコンポーネントによって構築されます。Cisco ACIでは、CPU MTU サイズがインターフェイス MTU サイズよりも小さく、構築されたパケットサイズが CPU MTU より大きい場合、パケットはカーネルによってドロップされます（特に IPv6）。このような制御パケットのドロップを回避するには、コントロールプレーンとインターフェイスの両方で常に同じ MTU 値を設定します。

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。