ACI ファブリック内での転送

この章は、次の内容で構成されています。

- ACI ファブリック内の転送について （1 ページ）
- 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 （2 ページ）
- ACI で VXLAN （3 ページ）
- 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 3 VNID （5 ページ）
- ポリシー ID と適用 （7 ページ）
- ACI ファブリック ネットワーク アクセス セキュリティ ポリシー モデル（契約） （8 ページ）
- マルチキャスト ツリー トポロジ （14 ページ）
- トラフィック ストーム制御について （16 ページ）
- ストーム制御の注意事項と制約事項 （16 ページ）
- ファブリック ロード バランシング （19 ページ）
- エンドポイントの保持 （22 ページ）
- IP エンドポイントの学習動作 （23 ページ）
- プロキシ ARP について （25 ページ）
- ループ検出 （31 ページ）
- 不正なエンドポイントの検出 （34 ページ）

ACI ファブリック内の転送について

ACI ファブリックは、64,000 以上の専用テナント ネットワークをサポートしています。単一のファブリックは、100 万以上の IPv4/IPv6 エンドポイント、64,000 以上のテナント、および 200,000 以上の 10G ポートをサポートできます。ACI ファブリックにより、物理サービスと仮想サービス間を接続する追加のソフトウェアやハードウェア ゲートウェイを必要とすることなくサービス（物理または仮想）がどこでも可能になり、Virtual Extensible Local Area Network（VXLAN）/VLAN/Network Virtualization using Generic Routing Encapsulation（NVGRE）のカプセル化が正規化されます。

ACI ファブリックは、基盤となる転送グラフからエンドポイント ID ポリシーおよび関連するポリシーを分離します。また、最適なレイヤ 3 およびレイヤ 2 フォワーディングを保証する分散レイヤ 3 ゲートウェイが提供されます。ファブリックは、一般的な場所の制約（あらゆる場所の IP アドレス）なしで標準のブリッジングおよびルーティングのセマンティックをサポートし、IP コントロールプレーンの Address Resolution Protocol（ARP）/Gratuitous Address Resolution Protocol（GARP）に関するフラッディング要件を削除します。ファブリック内のすべてのトラフィックは、VXLAN 内にカプセル化されます。

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。

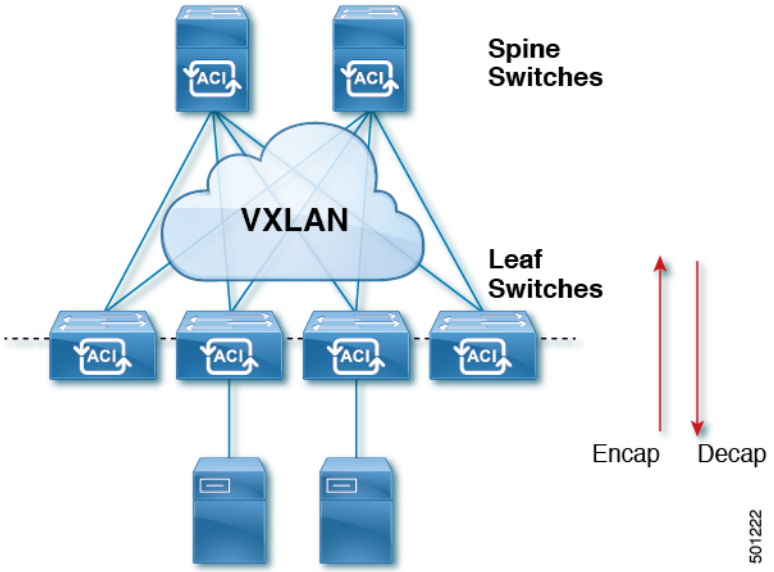
今日のアプリケーション設計は、データセンターのアクセス レイヤを通る、サーバ間の East-West トラフィックを増大させています。このシフトを促進しているアプリケーションには、Hadoop のようなビッグデータの分散処理の設計 、VMware vMotion のようなライブの仮想マシンまたはワークロードの移行、サーバのクラスタ リング、および多層アプリケーションなどが含まれます。

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



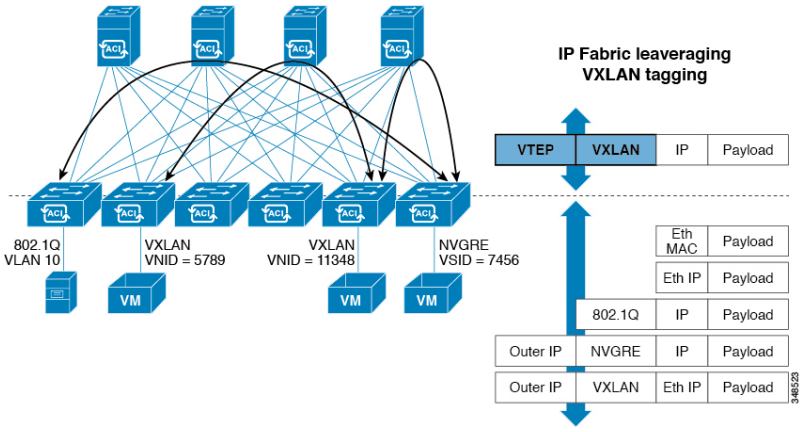
トラフィックがファブリックに入ると、ACIがカプセル化してポリシーを適用し、必要に応じてスパイン スイッチ (最大 2 ホップ) によってファブリックを通過させ、ファブリックを出るときにカプセル化を解除します。ファブリック内では、ACIはエンドポイント間通信でのすべての転送について、Intermediate System-to-Intermediate System プロトコル (IS-IS) および Council of Oracle Protocol (COOP) を使用します。これにより、すべての ACI リンクがアクティブで、ファブリック内での等コストマルチパス (ECMP) 転送と高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

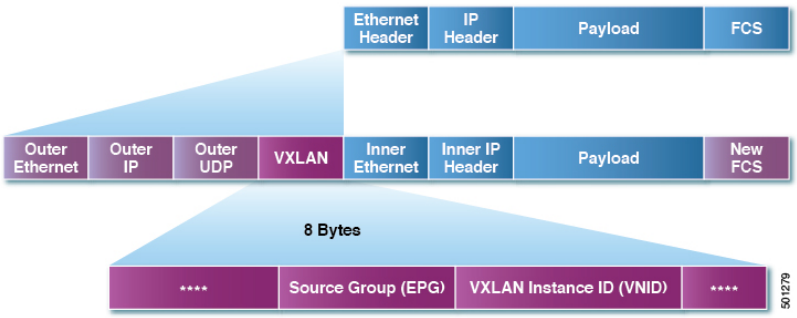
図 *2 : ACI* カプセル化の正規化



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EPGIDが転送から分離されます。次の図に示すように、ACIVXLAN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

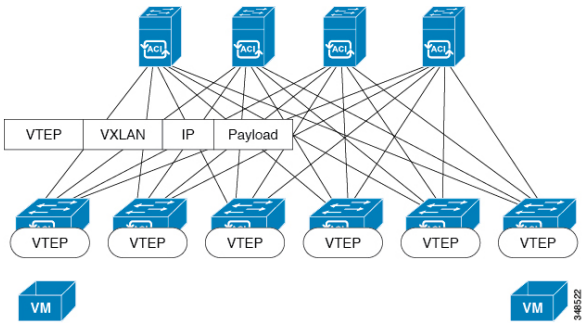
のリスクなしでデータセンター ネットワーク間をオーバーレイ ネットワーク、VXLAN でレイヤ 2 の隣接関係を維持します。

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 3 VNID

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

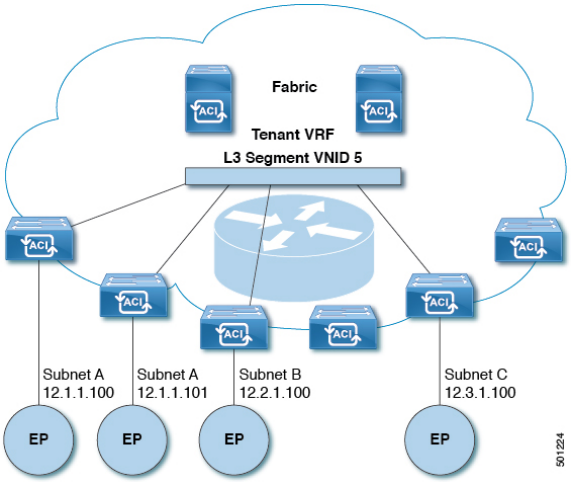
IP インターフェイスには一意の IP アドレスがあります。これは、インフラストラクチャ VLAN として知られる、転送 IP ネットワーク上の VTEP を識別します。VTEP デバイスはこの IP アドレスを使用してイーサネット フレームをカプセル化し、カプセル化されたパケットを、IP インターフェイスを介して転送ネットワークへ送信します。また、VTEP デバイスはリモート VTEP で VXLAN セグメントを検出し、IP インターフェイスを介してリモートの MAC Address-to-VTEP マッピングについて学習します。

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。

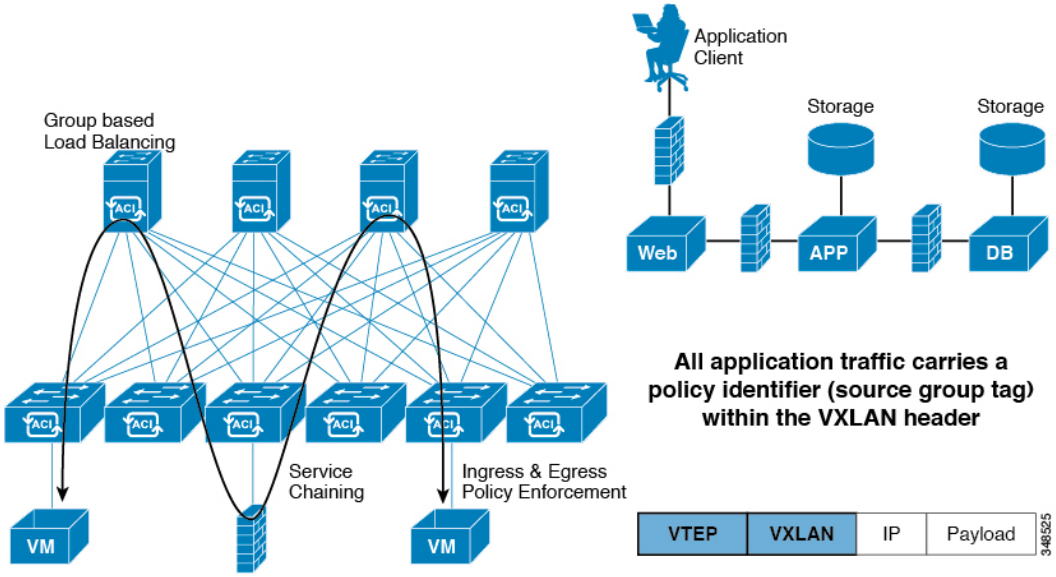
各プラットフォームの適切な MTU 値については、それぞれの設定ガイドを参照してください。

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。

ポリシー ID と適用

アプリケーション ポリシーは、VXLAN パケットで送信される個別のタギング属性を使用して転送から分離されます。ポリシー ID は、ACI ファブリック内のすべてのパケットで送信され、完全に分散した形でポリシーの一貫した適用を行うことができます。次の図は、ポリシー ID を示します。

図 *6 :* ポリシー *ID* と適用



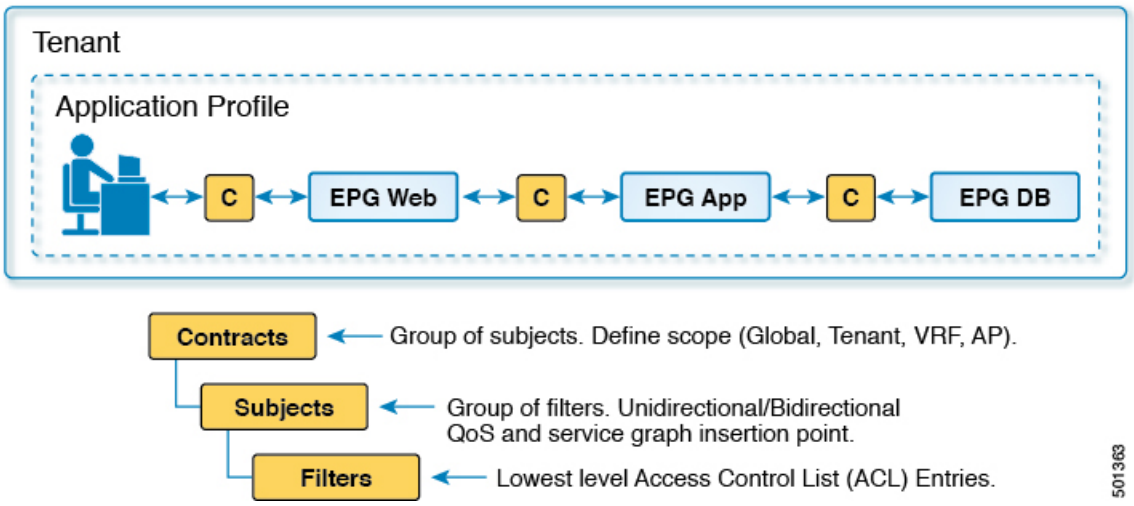
ファブリックおよびアクセスポリシーは、内部のファブリックインターフェイスおよび外部のアクセスインターフェイスの動作を管理します。システムは、デフォルトのファブリックおよびアクセスポリシーを自動的に作成します。ファブリックの管理者（ファブリック全体へのアクセス権がある者）は、要件に応じてデフォルトのポリシーを変更したり、新しいポリシーを作成できます。ファブリックおよびアクセスポリシーにより、さまざまな機能やプロトコルを有効にできます。APICのセレクタにより、ファブリックの管理者は、ポリシーを適用するノードおよびインターフェイスを選択できます。

ACI ファブリック ネットワーク アクセス セキュリティ ポリシー モデル（契約）

ACI のファブリック セキュリティ ポリシー モデルはコントラクトに基づいています。このアプローチにより、従来のアクセス コントロール リスト（ACL）の制限に対応できます。コントラクトには、エンドポイントグループ間のトラフィックで適用されるセキュリティポリシーの仕様が含まれます。

次の図は、 契約のコンポーネントを示しています。

図 7: 契約のコンポーネント



EPG 通信にはコントラクトが必要です。EPG/EPG 通信はコントラクトなしでは許可されません。APICは、コントラクトや関連するEPGなどのポリシーモデル全体を各スイッチの具象モデルにレンダリングします。入力時に、ファブリックに入るパケットはすべて、必要なポリシーの詳細でマークされます。EPGの間を通過できるトラフィックの種類を選択するためにコントラクトが必要とされるので、コントラクトはセキュリティポリシーを適用します。コントラクトは、従来のネットワーク設定でのアクセス コントロール リスト（ACL）によって扱われるセキュリティ要件を満たす一方で、柔軟性が高く、管理が容易な、包括的なセキュリティ ポリシー ソリューションです。

アクセス コントロール リストの制限

従来のアクセス コントロール リスト（ACL）には、ACI ファブリック セキュリティ モデルが対応する多数の制限があります。従来の ACL は、ネットワーク トポロジと非常に強固に結合されています。それらは通常、ルータまたはスイッチの入力および出力インターフェイスごとに設定され、そのインターフェイス、およびそれらのインターフェイスを流れることが予期されるトラフィックに合わせてカスタマイズされます。このカスタマイズにより、それらは多くの場合インターフェイス間で再利用できません。もちろんこれはルータまたはスイッチ間にも当てはまります。

従来の ACL は、非常に複雑で曖昧です。なぜなら、そのリストには、許可された特定の IP アドレス、サブネット、およびプロトコルのリストと、明確に許可されていない多くのものが含まれているためです。この複雑さは、問題が生じるのを管理者が懸念して ACL ルールを削除するのを躊躇するため、維持が困難で、多くの場合は増大するだけということを意味します。複雑さは、それらが通常 WAN と企業間または WAN とデータセンター間の境界などのネットワーク内の特定の境界ポイントでのみ配置されていることを意味します。この場合、ACL のセキュリティのメリットは、エンタープライズ内またはデータセンターに含まれるトラフィック向けには生かされません。

別の問題として、1 つの ACL 内のエントリ数の大幅増加が考えられます。ユーザは多くの場合、一連の送信元が一連のプロトコルを使用して一連の宛先と通信するのを許可する ACL を作成します。最悪の場合、*N* の送信元が *K* のプロトコルを使用して *M* の宛先と対話する場合、ACL に *N*M*K* の行が存在する場合があります。ACL は、プロトコルごとに各宛先と通信する各送信元を一覧表示する必要があります。また、ACL が非常に大きくなる前に多くのデバイスやプロトコルを取得することはありません。

ACI ファブリック セキュリティ モデルは、これらの ACL の問題に処理します。ACI ファブリック セキュリティ モデルは、管理者の意図を直接表します。管理者は、連絡先、フィルタ、およびラベルの管理対象オブジェクトを使用してエンドポイントのグループがどのように通信するかを指定します。これらの管理対象オブジェクトは、ネットワークのトポロジに関連していません。なぜなら、それらは特定のインターフェイスに適用されないためです。それらは、エンドポイントのこれらのグループの接続場所に関係なく、ネットワークが強要しなければならない簡易なルールです。このトポロジの独立性は、これらの管理対象オブジェクトが特定の境界ポイントとしてだけではなくデータセンター全体にわたって容易に配置して再利用できることを意味します。

ACI ファブリック セキュリティ モデルは、エンドポイントのグループ化コンストラクトを直接使用するため、サーバのグループが相互に通信できるようにするための概念はシンプルです。1 つのルールにより、任意の数の送信元が同様に任意の数の宛先と通信することを可能にできます。このような簡略化により、そのスケールと保守性が大幅に向上します。つまり、データセンター全体でより簡単に使用できることにもつながります。

セキュリティ ポリシー仕様を含むコントラクト

ACI セキュリティ モデルでは、コントラクトに EPG 間の通信を管理するポリシーが含まれます。コントラクトは通信内容を指定し、EPG は通信の送信元と宛先を指定します。コントラクトは次のように EPG をリンクします。

EPG 1 ---------- コントラクト ---------- EPG 2

コントラクトで許可されていれば、EPG 1 のエンドポイントは EPG 2 のエンドポイントと通信でき、またその逆も可能です。このポリシーの構造には非常に柔軟性があります。たとえば、EPG 1 と EPG 2 間には多くのコントラクトが存在でき、1 つのコントラクトを使用する EPG が 3 つ以上存在でき、コントラクトは複数の EPG のセットで再利用できます。

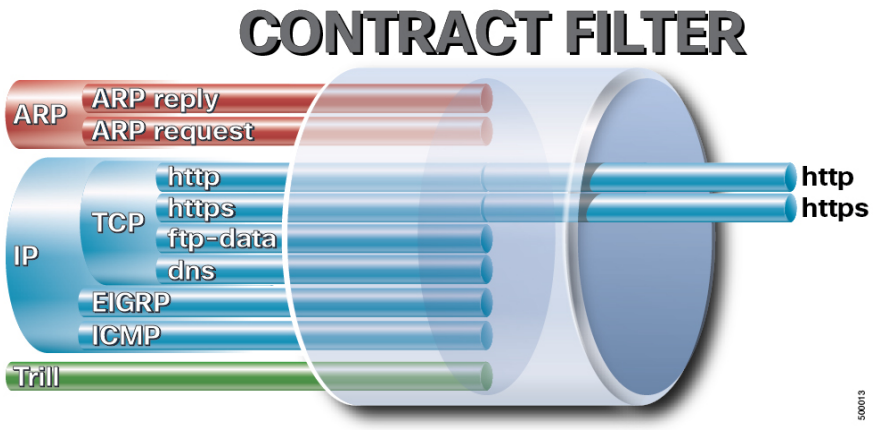
また EPG とコントラクトの関係には方向性があります。EPG はコントラクトを提供または消費できます。コントラクトを提供する EPG は通常、一連のクライアント デバイスにサービスを提供する一連のエンドポイントです。そのサービスによって使用されるプロトコルはコントラクトで定義されます。コントラクトを消費する EPG は通常、そのサービスのクライアントである一連のエンドポイントです。クライアント エンドポイント (コンシューマ) がサーバ エンドポイント (プロバイダー) に接続しようとすると、コントラクトはその接続が許可されるかどうかを確認します。特に指定のない限り、そのコントラクトは、サーバがクライアントへの接続を開始することを許可しません。ただし、EPG 間の別のコントラクトが、その方向の接続を簡単に許可する場合があります。

この提供/消費の関係は通常、EPG とコントラクト間を矢印を使って図で表されます。次に示す矢印の方向に注目してください。

EPG 1 <------- 消費 -------- コントラクト <------- 提供 -------- EPG 2

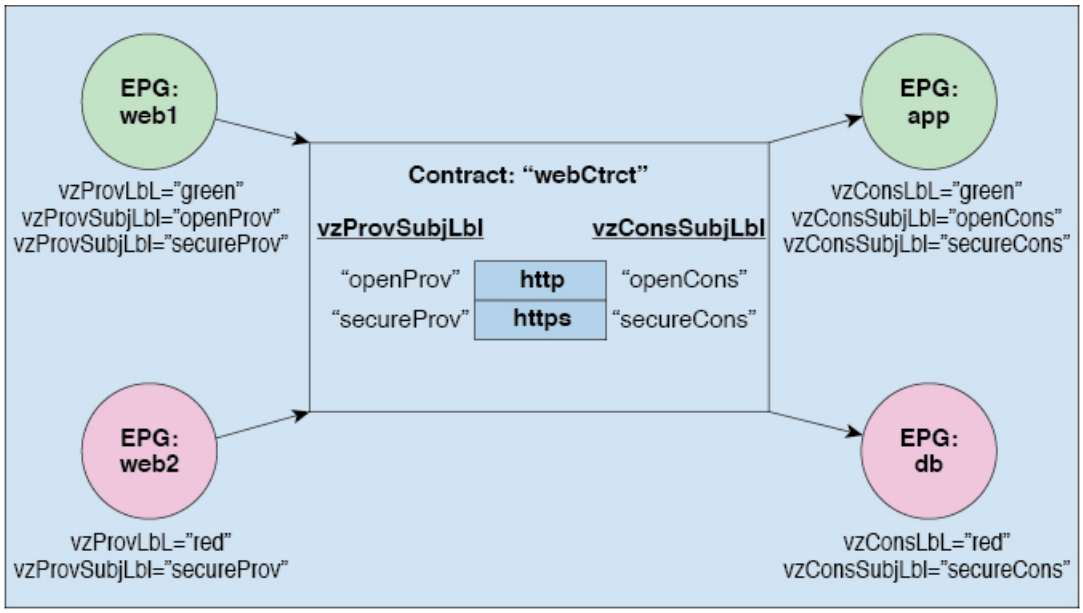
コントラクトは階層的に構築されます。1 つ以上のサブジェクトで構成され、各サブジェクトには 1 つ以上のフィルタが含まれ、各フィルタは 1 つ以上のプロトコルを定義できます。

図 *8:* コントラクト フィルタ



次の図は、コントラクトが EPG の通信をどのように管理するかを示します。

図 *9 : EPG/EPG* 通信を決定するコントラクト



たとえば、TCP ポート 80 とポート 8080 を指定する HTTP と呼ばれるフィルタと、TCP ポート 443 を指定する HTTPS と呼ばれる別のフィルタを定義できます。その後、2 セットの情報カテゴリを持つ webCtrct と呼ばれるコントラクトを作成できます。openProv と openCons are は HTTP フィルタが含まれるサブジェクトです。secureProv と secureCons は HTTPS フィルタが含まれる情報カテゴリです。この webCtrct コントラクトは、Web サービスを提供する EPG とそのサービスを消費するエンドポイントを含む EPG 間のセキュアな Web トラフィックと非セキュアな Web トラフィックの両方を可能にするために使用できます。

これらの同じ構造は、仮想マシンのハイパーバイザを管理するポリシーにも適用されます。EPG が Virtual Machine Manager（VMM）のドメイン内に配置されると、APIC は EPG に関連付けられたすべてのポリシーを VMM ドメインに接続するインターフェイスを持つリーフ スイッチにダウンロードします。VMM ドメインの完全な説明については、『*Application Centric Infrastructure Fundamentals*』の「*Virtual Machine Manager Domains*」の章を参照してください。このポリシーが作成されると、APIC は EPG のエンドポイントへの接続を可能にするスイッチを指定する VMM ドメインにそれをプッシュ（あらかじめ入力）します。VMM ドメインは、EPG 内のエンドポイントが接続できるスイッチとポートのセットを定義します。エンドポイントがオンラインになると、適切な EPG に関連付けられます。パケットが送信されると、送信元 EPG および宛先 EPG がパケットから取得され、対応するコントラクトで定義されたポリシーでパケットが許可されたかどうかが確認されます。許可された場合は、パケットが転送されます。許可されない場合は、パケットはドロップされます。

コントラクトは 1 つ以上のサブジェクトで構成されます。各サブジェクトには 1 つ以上のフィルタが含まれます。各フィルタには 1 つ以上のエントリが含まれます。各エントリは、アクセス コントロール リスト (ACL) の 1 行に相当し、エンドポイント グループ内のエンドポイントが接続されているリーフ スイッチで適用されます。

詳細には、コントラクトは次の項目で構成されます。

- 名前：テナントによって消費されるすべてのコントラクト (common テナントまたはテナント自体で作成されたコントラクトを含む) にそれぞれ異なる名前が必要です。
- サブジェクト：特定のアプリケーションまたはサービス用のフィルタのグループ。
- フィルタ：レイヤ 2 ～ レイヤ 4 の属性 (イーサネット タイプ、プロトコル タイプ、TCP フラグ、ポートなど) に基づいてトラフィックを分類するために使用します。
- アクション：フィルタリングされたトラフィックで実行されるアクション。次のアクションがサポートされます。
  - トラフィックの許可 (通常のコントラクトのみ)
  - トラフィックのマーク (DSCP/CoS) (通常のコントラクトのみ)
  - トラフィックのリダイレクト (サービス グラフによる通常のコントラクトのみ)
  - トラフィックのコピー (サービス グラフまたは SPAN による通常のコントラクトのみ)
  - トラフィックのブロック（禁止コントラクトのみ）

    Cisco APIC リリース 3.2(x) および名前が EX または FX で終わるスイッチでは、標準コントラクトで代わりに件名 [拒否] アクションまたは [コントラクトまたは件名の除外] を使用して、指定のパターンを持つトラフィックをブロックできます。
  - トラフィックのログ（禁止コントラクトと通常のコントラクト）
- エイリアス：(任意) 変更可能なオブジェクト名。オブジェクト名は作成後に変更できませんが、エイリアスは変更できるプロパティです。

このように、コントラクトによって許可や拒否よりも複雑なアクションが可能になります。コントラクトは、所定のサブジェクトに一致するトラフィックをサービスにリダイレクトしたり、コピーしたり、その QoS レベルを変更したりできることを指定可能です。具象モデルでアクセス ポリシーをあらかじめ入力すると、APIC がオフラインまたはアクセスできない場合でも、エンドポイントは移動でき、新しいエンドポイントをオンラインにでき、通信を行うことができます。APIC は、ネットワークの単一の障害発生時点から除外されます。ACI ファブリックにパケットが入力されると同時に、セキュリティ ポリシーがスイッチで実行している具象モデルによって適用されます。

セキュリティ ポリシーの適用

トラフィックは前面パネルのインターフェイスからリーフ スイッチに入り、パケットは送信元 EPG の EPG でマーキングされます。リーフ スイッチはその後、テナント エリア内のパケットの宛先 IP アドレスでフォワーディング ルックアップを実行します。ヒットすると、次のシナリオのいずれかが発生する可能性があります。

1. ユニキャスト（/32）ヒットでは、宛先エンドポイントの EPG と宛先エンドポイントが存在するローカル インターフェイスまたはリモート リーフ スイッチの VTEP IP アドレスが提供されます。

2. サブネット プレフィクス（/32 以外）のユニキャスト ヒットでは、宛先サブネット プレフィクスの EPG と宛先サブネット プレフィクスが存在するローカル インターフェイスまたはリモート リーフ スイッチの VTEP IP アドレスが提供されます。
3. マルチキャスト ヒットでは、ファブリック全体の VXLAN カプセル化とマルチキャスト グループの EPG で使用するローカル レシーバのローカル インターフェイスと外側の宛先 IP アドレスが提供されます。

（注） マルチキャストと外部ルータのサブネットは、入力リーフ スイッチでのヒットを常にもたらします。セキュリティ ポリシーの適用は、宛先 EPG が入力リーフ スイッチによって認識されるとすぐに発生します。

転送テーブルの誤りにより、パケットがスパイン スイッチの転送プロキシに送信されます。転送プロキシはその後、転送テーブル検索を実行します。これが誤りである場合、パケットはドロップされます。これがヒットの場合、パケットは宛先エンドポイントを含む出力リーフ スイッチに送信されます。出力リーフ スイッチが宛先の EPG を認識するため、セキュリティ ポリシーの適用が実行されます。出力リーフ スイッチは、パケット送信元の EPG を認識する必要があります。ファブリック ヘッダーは、入力リーフ スイッチから出力リーフ スイッチに EPG を伝送するため、このプロセスをイネーブルにします。スパイン スイッチは、転送プロキシ機能を実行するときに、パケット内の元の EPG を保存します。

出力リーフ スイッチでは、送信元 IP アドレス、送信元 VTEP、および送信元 EPG 情報は、学習によってローカルの転送テーブルに保存されます。ほとんどのフローが双方向であるため、応答パケットがフローの両側で転送テーブルに入力し、トラフィックが両方向で入力フィルタリングされます。

マルチキャストおよび EPG セキュリティ

マルチキャスト トラフィックでは、興味深い問題が起こります。ユニキャスト トラフィックでは、宛先 EPG はパケットの宛先の検査からはっきり知られています。ただし、マルチキャスト トラフィックでは、宛先は抽象的なエンティティ、マルチキャスト グループです。パケットの送信元はマルチキャスト アドレスではないため、送信元 EPG は以前のユニキャストの例と同様に決定されます。宛先グループの起源はマルチキャストが異なる場所です。

マルチキャスト グループが、ネットワーク トポロジから若干独立しているので、グループ バインディングへの (S, G) および (*, G) の静的設定は受け入れ可能です。マルチキャスト グループが転送テーブルにある場合、マルチキャスト グループに対応する EPG は、転送テーブルにも配置されます。

（注） このマニュアルでは、マルチキャスト グループとしてマルチキャスト ストリームを参照します。

リーフ スイッチは、マルチキャスト ストリームに対応するグループを常に宛先 EPG と見なし、送信元 EPG と見なすことはありません。前述のアクセス コントロール マトリクスでは、マルチキャスト EPG が送信元の場合は行の内容は無効です。トラフィックは、マルチキャスト ストリームの送信元またはマルチキャスト ストリームに加わりたい宛先からマルチキャストストリームに送信されます。マルチキャスト ストリームが転送テーブルにある必要があり、ストリーム内に階層型アドレッシングがないため、マルチキャストトラフィックは、入力ファブリックの端でアクセスが制御されます。その結果、IPv4 マルチキャストは入力フィルタリングとして常に適用されます。

マルチキャスト ストリームの受信側は、トラフィックを受信する前にマルチキャスト ストリームに最初に加わる必要があります。IGMP Join 要求を送信すると、マルチキャスト レシーバは実際に IGMP パケットの送信元になります。宛先はマルチキャスト グループとして定義され、宛先 EPG は転送テーブルから取得されます。ルータが IGMP Join 要求を受信する入力点で、アクセス制御が適用されます。Join 要求が拒否された場合、レシーバはその特定のマルチキャスト ストリームからトラフィックを受信しません。

マルチキャスト EPG へのポリシーの適用は、前述のようにコントラクトのルールに従ってリーフ スイッチにより入力時に発生します。また、EPG バインディングに対するマルチキャスト グループは、APIC によって特定のテナント（VRF）を含むすべてのリーフ スイッチにプッシュされます。

マルチキャスト ツリー トポロジ

ACI ファブリックは、アクセス ポートからのユニキャスト、マルチキャスト、およびブロードキャスト トラフィックの転送をサポートします。エンドポイント ホストからのすべてのマルチデスティネーション トラフィックは、ファブリックにマルチキャスト トラフィックとして伝送されます。

ACI ファブリックは、入力インターフェイスに入るトラフィックを使用可能な中間ステージのスパイン スイッチを介して関連する出力スイッチにルーテッドできる Clos トポロジ（Charles Clos にちなんで名付けられた）に接続されるスパインおよびリーフスイッチで構成されます。リーフスイッチには次の 2 種類のポートがあります。スパイン スイッチに接続するためのファブリック ポートと、サーバー、サービス アプライアンス、ルータ、Fabric Extender（FEX; ファブリック エクステンダ）などを接続するアクセス ポートです。

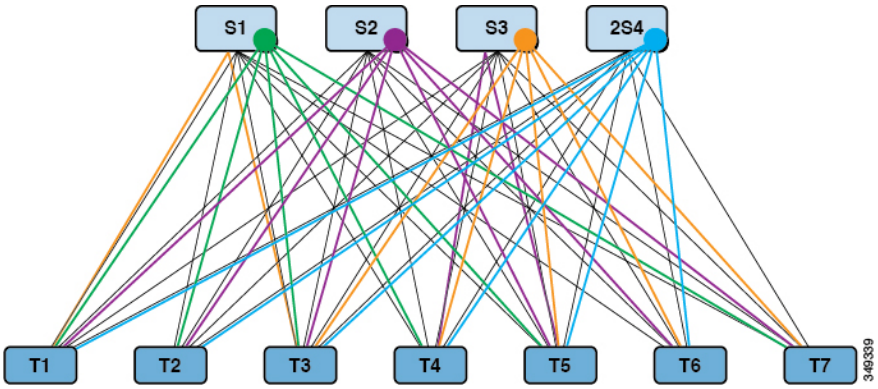
リーフスイッチ（top-of-rack（ToR; トップオブラック）スイッチとも呼ばれます）は、スパイン スイッチ (「エンド オブ ロウ」または「EoR」スイッチとも呼ばれます）に接続されます。リーフスイッチは互いに接続されず、スパイン スイッチはリーフスイッチのみに接続します。この Clos トポロジでは、すべての下位層のスイッチがフルメッシュ トポロジの最上位層のスイッチにそれぞれ接続されます。スパイン スイッチに不具合があると、ACI ファブリック全体のパフォーマンスだけがわずかに低下します。データ パスは、トラフィック負荷がスパイン スイッチ間で均等に分散されるように選択されます。

ACI ファブリックは、Forwarding Tag（FTAG）ツリーを使用してバランス マルチデスティネーション トラフィックをロードします。すべてのマルチデスティネーション トラフィックは、ファブリック内でカプセル化された IP マルチキャスト トラフィックの形式で転送されます。入力リーフは、FTAG をスパインに転送するときにトラフィックに割り当てます。FTAG は接

続先マルチキャストアドレスの一部としてパケットに割り当てられます。ファブリックでは、トラフィックは指定されたFTAGツリーに沿って転送されます。スパインおよび中間リーフスイッチは、FTAG ID に基づいてトラフィックを転送します。転送ツリーは、FTAG ID 1 つにつき 1 つ構築されます。任意の 2 つのノード間で、FTAG 1 つにつきリンク 1 つだけが転送されます。複数の FTAG を使用することで、転送に異なるリンクを使用している各 FTAG でパラレル リンクを使用できます。ファブリック内の FTAG ツリーの数が多いほど、ロード バランシングの効果が大きい可能性があるということになります。ACI ファブリックは、最大 12 個の FTAG をサポートします。

次の図は、4 つの FTAG によるトポロジを示します。ファブリック内のすべてのリーフスイッチは、各 FTAG に直接または中継ノードを介して接続されます。1 つの FTAG が各スパイン ノードに根付いています。

図 *10*：マルチキャスト ツリー トポロジ



リーフスイッチはスパインへの直接接続性がある場合、直接パスを使用してFTAGツリーに接続します。直接リンクがない場合、リーフスイッチは上記の図に示すようにFTAGツリーに接続されている中継ノードを使用します。図には、各スパインが 1 つの FTAG ツリーのルートとして示されていますが、複数のFTAGツリールートを1つのノード上に置くことができます。

ACI ファブリック起動検出プロセスの一環として、FTAG ルートはスパイン スイッチに配置されます。APIC は、各スパイン スイッチをスパインがアンカーする FTAG で構成します。ルートの ID と FTAG の数は構成から取得されます。APIC は、使用される FTAG ツリーの数と各ツリーに対するルートを指定します。FTAG ツリーは、ファブリックでトポロジの変更があるたびに再計算されます。

ルートの配置は誘導される構成で、スパイン スイッチの障害などのランタイム イベントで動的に再度ルート付けされることはありません。通常、FTAG構成は静的です。スパイン スイッチの追加または削除時は、管理者がスパイン スイッチの残りのセットまたは拡張セット間で FTAG を再配布することを決める可能性があるため、FTAG はあるスパインから別のスパインへ再アンカーできます。

トラフィック ストーム制御について

トラフィック ストームは、パケットが LAN でフラッディングする場合に発生するもので、過剰なトラフィックを生成し、ネットワークのパフォーマンスを低下させます。トラフィック ストーム制御ポリシーを使用すると、物理インターフェイス上におけるブロードキャスト、未知のマルチキャスト、または未知のユニキャストのトラフィック ストームによって、レイヤ 2 ポート経由の通信が妨害されるのを防ぐことができます。

デフォルトでは、ストーム制御はACIファブリックでは有効になっていません。ACIブリッジドメイン（BD）レイヤ 2 の未知のユニキャストのフラッディングは BD 内でデフォルトで有効になっていますが、管理者が無効にすることができます。その場合、ストーム制御ポリシーはブロードキャストと未知のマルチキャストのトラフィックにのみ適用されます。レイヤ2の未知のユニキャストのフラッディングが BD で有効になっている場合、ストーム制御ポリシーは、ブロードキャストと未知のマルチキャストのトラフィックに加えて、レイヤ2の未知のユニキャストのフラッディングに適用されます。

トラフィック ストーム制御（トラフィック抑制ともいいます）を使用すると、着信するブロードキャスト、マルチキャスト、未知のユニキャストのトラフィックのレベルを1秒間隔でモニタできます。この間に、トラフィック レベル（ポートで使用可能な合計帯域幅のパーセンテージ、または特定のポートで許可される1秒あたりの最大パケット数として表されます）が、設定したトラフィック ストーム制御レベルと比較されます。入力トラフィックが、ポートに設定したトラフィック ストーム制御レベルに到達すると、トラフィック ストーム制御機能によってそのインターバルが終了するまでトラフィックがドロップされます。管理者は、ストーム制御しきい値を超えたときにエラーを発生させるようにモニタリング ポリシーを設定できます。

ストーム制御の注意事項と制約事項

以下のガイドラインと制約事項に従って、トラフィック ストーム制御レベルを設定してください。

- 通常、ファブリック管理者は以下のインターフェイスのファブリック アクセス ポリシーでストーム制御を設定します。
  - 標準トランク インターフェイス。
  - 単一リーフ スイッチ上のダイレクト ポート チャネル。
  - バーチャル ポート チャネル（2 つのリーフ スイッチ上のポート チャネル）。
- リリース 4.2(1) 以降では、ストーム制御のしきい値に達した場合に、次の制約事項に従って、SNMP トラップをCisco Application Centric Infrastructure（ACI）からトリガーできるようになりました。
  - ストーム制御に関連するアクションには、ドロップとシャットダウンの2つがあります。シャットダウン アクションでは、インターフェイス トラップが発生しますが、ストームがアクティブまたはクリアであることを示すためのストーム制御トラップ

は、シャットダウンアクションによっては決定されません。したがって、ポリシーでシャットダウンアクションが設定されているストーム制御トラップは無視する必要があります。

- ストーム制御ポリシーがオンの状態でポートがフラップすると、統計情報の収集時にクリア トラップとアクティブ トラップが一緒に表示されます。通常、クリア トラップとアクティブ トラップは一緒に表示されませんが、この場合は予期される動作です。

- ポート チャネルおよびバーチャル ポート チャネルでは、ストーム制御値（1 秒あたりのパケット数またはパーセンテージ）はポートチャネルのすべての個別メンバーに適用されます。ポートチャネルのメンバーであるインターフェイスには、ストーム制御を設定しないでください。

（注） Cisco Application Policy Infrastructure Controller（APIC）リリース 1.3(1) およびスイッチ リリース 11.3(1) 以降のスイッチ ハードウェアの場合、ポート チャネル設では、集約ポートのトラフィック抑制は設定値の最大2倍になることがあります。新しいハードウェア ポートは slice-0 と slice-1 の 2 つのグループに内部的にさらに分割されています。スライスマップを確認するには、`vsh_lc` コマンドの `show platform internal hal l2 port gpd` を使用して、`Sl` カラムで `slice 0` または `slice 1` を探します。ポート チャネル メンバーがスライス 0 とスライス 1 の両方に該当する場合、式は各スライスに基づいて計算されるため、許可されるストーム制御トラフィックが設定値の 2 倍になることがあります。

- 使用可能な帯域幅のパーセンテージで設定する場合、値100はトラフィック ストーム制御を行わないことを意味し、値 0.01 はすべてのトラフィックを抑制します。

- ハードウェアの制限およびさまざまなサイズのパケットのカウント方式が原因で、レベルのパーセンテージは概数になります。着信トラフィックを構成するフレームのサイズに応じて、実際に適用されるパーセンテージ レベルと設定したパーセンテージ レベルの間には、数パーセントの誤差がある可能性があります。1 秒あたりのパケット数（PPS）の値は、256 バイトに基づいてパーセンテージに変換されます。

- 最大バーストは、通過するトラフィックがないときに許可されるレートの最大累積です。トラフィックが開始されると、最初の間隔では累積レートまでのすべてのトラフィックが許可されます。後続の間隔では、トラフィックは設定されたレートまでのみ許可されます。サポートされる最大数は65535 KB です。設定されたレートがこの値を超えると、PPS とパーセンテージの両方についてこの値で制限されます。

- 累積可能な最大バーストは 512 MB です。

- 最適化されたマルチキャスト フラッディング（OMF）モードの出力リーフ スイッチでは、トラフィック ストーム制御は適用されません。

- OMF モードではない出力リーフ スイッチでは、トラフィック ストーム制御が適用されます。
- FEX のリーフ スイッチでは、ホスト側インターフェイスにはトラフィック ストーム制御を使用できません。
- Cisco Nexus C93128TX、C9396PX、C9396TX、C93120TX、C9332PQ、C9372PX、C9372TX、C9372PX-E、C9372TX-E の各スイッチでは、トラフィック ストーム制御のユニキャスト/マルチキャストの差別化がサポートされていません。
- Cisco Nexus C93128TX、C9396PX、C9396TX、C93120TX、C9332PQ、C9372PX、C9372TX、C9372PX-E、C9372TX-E の各スイッチでは、トラフィック ストーム制御の SNMP トラップがサポートされていません。
- Cisco Nexus C93128TX、C9396PX、C9396TX、C93120TX、C9332PQ、C9372PX、C9372TX、C9372PX-E、C9372TX-E の各スイッチでは、トラフィック ストーム制御トラップがサポートされていません。
- ストーム制御アクションは、物理イーサネット インターフェイスおよびポート チャネル インターフェイスでのみサポートされます。

  リリース 4.1(1) 以降では、ストーム制御**シャットダウン**オプションがサポートされています。デフォルトの Soak Instance Count を持つインターフェイスに対して**シャットダウン**アクションが選択されると、しきい値を超えるパケットは 3 秒間ドロップされ、ポートは 3 秒間シャットダウンされます。デフォルトのアクションは、**ドロップ**です。**シャットダウン**アクションを選択すると、ユーザーはソーキング間隔を指定するオプションを使用できます。デフォルトのソーキング間隔は 3 秒です。設定可能な範囲は 3 ～ 10 秒です。
- インターフェイスに設定されたデータ プレーン ポリシング（DPP）ポリサーの値がストーム ポリサーの値よりも低い場合、DPP ポリサーが優先されます。DPPポリサーとストーム ポリサーの間に設定されている低い方の値が、設定されたインターフェイスで適用されます。
- リリース 4.2(6) 以降、ストーム ポリサーは、DHCP、ARP、ND、HSRP、PIM、IGMP、および EIGRP プロトコルに対応する、リーフ スイッチのすべての転送制御トラフィックに強制されます。このことは、ブリッジ ドメインが**BD でのフラッディング**または**カプセル化でのフラッディング**のどちらに設定されているかには関係しません。この動作の変更は、EX 以降のリーフスイッチにのみ適用されます。
  - EX スイッチでは、プロトコルの 1 つに対し、スーパーバイザ ポリサーとストーム ポリサーの両方を設定できます。この場合、サーバーが設定されたスーパーバイザ ポリサー レート（制御プレーン ポリシング、CoPP）よりも高いレートでトラフィックを送信すると、ストーム ポリサーはストーム ポリサー レートとして設定されているよりも多くのトラフィックを許可します。着信トラフィック レートがスーパーバイザ ポリサー レート以下の場合、ストーム ポリサーは設定されたストーム トラフィック レートを正しく許可します。この動作は、設定されたスーパーバイザ ポリサーおよびストーム ポリサーのレートに関係なく適用されます。
  - ストーム ポリサーが、指定されたプロトコルのリーフ スイッチで転送されるすべての制御トラフィックに適用されるようになった結果、リーフ スイッチで転送される制

御トラフィックがストーム ポリサー ドロップの対象になります。以前のリリースでは、この動作の変更の影響を受けるプロトコルでは、このようなストームポリサーのドロップは発生しません。

- トラフィック ストーム制御は、PIM が有効になっているブリッジ ドメインまたは VRF インスタンスのマルチキャスト トラフィックをポリシングできません。
- ストーム コントロール ポリサーがポートチャネル インターフェイスに適用されている場合、許可されるレートが設定されているレートを超えることがあります。ポートチャネルのメンバー リンクが複数のスライスにまたがる場合、許可されるトラフィック レートは、構成されたレートにメンバー リンクがまたがるスライスの数を掛けたものに等しくなります。

  ポートからスライスへのマッピングは、スイッチ モデルによって異なります。

  例として、ストーム ポリサー レートが 10Mbps のメンバー リンク port1、port2、および port3 を持つポートチャネルがあるとします。

  - port1、port2、port3 が slice1 に属している場合、トラフィックは 10Mbps にポリシングされます。
  - port1 と port2 が slice1 に属し、port3 が slice2 に属している場合、トラフィックは 20Mbps にポリシングされます。
  - port1 が slice1 に属し、port2が slice2 に属し、port3 が slice3 に属している場合、トラフィックは 30Mbps にポリシングされます。

ファブリック ロード バランシング

ACI ファブリックでは、利用可能なアップリンク リンク間のトラフィックを平衡化するためのロード バランシング オプションがいくつか提供されます。ここでは、リーフからスパインへのスイッチ トラフィックのロード バランシングについて説明します。

スタティック ハッシュ ロードバランシングは、各フローが 5 タプルのハッシュに基づいてアップリンクに割り当てられるネットワークで使用される従来のロードバランシング機構です。このロードバランシングにより、使用可能なリンクにほぼ均等な流量が分配されます。通常、流量が多いと、流量の均等な分配により帯域幅も均等に分配されます。ただし、いくつかのフローが残りよりも多いと、スタティック ロード バランシングにより完全に最適ではない結果がもたらされる場合があります。

ACI ファブリック ダイナミック ロード バランシング（DLB）は、輻輳レベルに従ってトラフィック割り当てを調整します。DLB では、使用可能なパス間の輻輳が測定され、輻輳状態が最も少ないパスにフローが配置されるので、データが最適またはほぼ最適に配置されます。

DLB は、フローまたはフローレットの粒度を使用して使用可能なアップリンクにトラフィックを配置するように設定できます。フローレットは、時間の大きなギャップによって適切に区切られるフローからのパケットのバーストです。パケットの 2 つのバースト間のアイドル間隔が使用可能なパス間の遅延の最大差より大きい場合、2 番目のバースト（またはフローレット）

を1つ目とは異なるパスに沿ってパケットのリオーダーなしで送信できます。このアイドル間隔は、フローレット タイマーと呼ばれるタイマーによって測定されます。フローレットにより、パケット リオーダーを引き起こすことなくロードバランシングに対する粒度の高いフローの代替が提供されます。

DLB 動作モードは積極的または保守的です。これらのモードは、フローレット タイマーに使用するタイムアウト値に関係します。アグレッシブ モードのフローレット タイムアウトは比較的小さい値です。この非常に精密なロード バランシングはトラフィックの分配に最適ですが、パケット リオーダーが発生する場合があります。ただし、アプリケーションのパフォーマンスに対する包括的なメリットは、保守的なモードと同等かそれよりも優れています。保守的なモードのフローレット タイムアウトは、パケットが並び替えられないことを保証する大きな値です。新しいフローレットの機会の頻度が少ないので、トレードオフは精度が低いロードバランシングです。DLB は常に最も最適なロード バランシングを提供できるわけではありませんが、スタティック ハッシュ ロード バランシングより劣るということはありません。

（注） すべての Nexus 9000 シリーズ スイッチには DLB のハードウェア サポートがありますが、DLB 機能は、第 2 世代プラットフォーム（EX、FX、および FX2 サフィックスを持つスイッチ）の現在のソフトウェア リリースでは有効になっていません。

ACI ファブリックは、リンクがオフラインまたはオンラインになったことで使用可能なリンク数が変化すると、トラフィックを調整します。ファブリックは、リンクの新しいセットでトラフィックを再分配します。

スタティックまたはダイナミックのロードバランシングのすべてのモードでは、トラフィックは、Equal Cost Multipath (ECMP) の基準を満たすアップリンクまたはパス上でのみ送信され、これらのパスはルーティングの観点から同等で最もコストがかかりません。

ロードバランシング技術ではありませんが、Dynamic Packet Prioritization（DPP）は、スイッチで DLB と同じメカニズムをいくつか使用します。DPP の設定は DLB 専用です。DPP は、長いフローよりも短いフローを優先します。短いフローは約 15 パケット未満です。短いフローは長いフローよりも遅延の影響を受けやすいため、DPP はアプリケーション全体のパフォーマンスを向上させることができます。

すべての DPP 優先トラフィックには、カスタム QoS 設定にもかかわらず CoS 3 がマークされています。

これらのパケットが同じリーフに入力および出力されると、CoS 値が保持され、フレームが CoS3 マーキングを使用してファブリックから送信されます。

GPRS トンネリング プロトコル（GTP）は、主にワイヤレス ネットワークでデータを配信するために使用されます。Cisco Nexus スイッチは Telcom データセンター内の場所です。パケットがデータセンターの Cisco Nexus 9000 スイッチを介して送信される場合、トラフィックは GTP ヘッダーに基づいてロード バランシングされる必要があります。ファブリックがリンク バンドルを介して外部ルータに接続されている場合、トラフィックはすべてのバンドル メンバー（たとえば、レイヤ 2 ポート チャネル、レイヤ 3 ECMP リンク、レイヤ 3 ポート チャネル、およびポート チャネル上の L3Out）に均等に分散される必要があります。）。GTP トラフィックのロード バランシングは、ファブリック内でも実行されます。

GTP ロード バランシングを実現するために、Cisco Nexus 9000 シリーズ スイッチは 5 タプルのロード バランシング メカニズムを使用します。ロードバランシング メカニズムでは、パケットの送信元 IP、宛先 IP、プロトコル、レイヤ 4 リソース、および宛先ポート（トラフィックが TCP または UDP の場合）フィールドが考慮されます。GTP トラフィックの場合は、これらのフィールドへの一意の値の数が限られていると、トンネルでのトラフィック ロードの均等分散が制限されます。

ロード バランシングにおける GTP トラフィックの極性化を回避するために、GTP ヘッダーのトンネル エンドポイント ID（TEID）が UDP ポート番号の代わりに使用されます。TEID がトンネルごとに異なるため、トラフィックをバンドルの複数のリンク間で均等にロードバランシングすることができます。

GTP ロード バランシングは、GTPU パケットに存在する 32 ビット TEID 値で送信元および宛先ポート情報を上書きします。

GTP トンネルのロード バランシング機能により、次のサポートが追加されます。

- 物理インターフェイスでの IPv4/IPv6 トランスポート ヘッダーによる GTP
- UDP ポート 2152 を使用した GTPU

ACI ファブリックのデフォルト設定では、従来の静的なハッシュが使用されます。スタティックなハッシュ機能により、アップリンク間のトラフィックがリーフ スイッチからスパイン スイッチに分配されます。リンクがダウンまたは起動すると、すべてのリンクのトラフィックが新しいアップリンク数に基づいて再分配されます。

リーフ/スパイン スイッチ ダイナミック ロード バランシング アルゴリズム

次の表に、リーフ/スパイン スイッチ ダイナミック ロード バランシングで使用されるデフォルトの設定不可能なアルゴリズムを示します。

表 *1 : ACI* リーフ/スパイン スイッチ ダイナミック ロード バランシング

Traffic Type	データ ポイントのハッシュ
リーフ/スパイン IP ユニキャスト	• 送信元 MAC アドレス • 宛先 MAC アドレス • 送信元 IP アドレス • 宛先 IP アドレス • プロトコル タイプ • 送信元レイヤ 4 ポート • 宛先レイヤ 4 ポート • セグメント ID（VXLAN VNID）または VLAN ID

Traffic Type	データ ポイントのハッシュ
リーフ/スパイン レイヤ 2	• 送信元 MAC アドレス • 宛先 MAC アドレス • セグメント ID（VXLAN VNID）または VLAN ID

エンドポイントの保持

スイッチでキャッシュ エンドポイントの MAC アドレスと IP アドレスを保持することで、パフォーマンスが向上します。スイッチは、アクティブになるときにエンドポイントについて学習します。ローカル エンドポイントはローカル スイッチにあります。リモート エンドポイントは他のスイッチにありますが、ローカルでキャッシュされます。リーフスイッチは、直接（または直接接続されたレイヤ2スイッチまたはファブリックエクステンダを通じて）接続されたエンドポイント、ローカルエンドポイント、およびファブリックの他のリーフスイッチに接続されたエンドポイント（ハードウェアのリモート エンドポイント）に関する場所とポリシーの情報を保存します。スイッチは、ローカル エンドポイントには 32 Kb エントリ キャッシュを、リモート エンドポイントには 64 Kb エントリ キャッシュを使用します。

リーフスイッチで稼働するソフトウェアは、これらのテーブルを能動的に管理します。ローカル的に接続されたエンドポイントでは、ソフトウェアは各エントリの保持タイマーの期限切れ後にエントリをエージング アウトします。エンドポイント エントリは、エンドポイントのアクティビティが終了するとスイッチ キャッシュからプルーニングされ、エンドポイントの場所が他のスイッチに移動するか、またはライフサイクルの状態がオフラインに変わります。ローカル保持タイマーのデフォルト値は15分です。非アクティブのエントリを削除する前に、リーフスイッチはエンドポイントに 3 つの ARP 要求を送信し、実際になくなっているかを確認します。スイッチが ARP 応答を受信しない場合、エントリはプルーニングされます。リモートで接続されたエンドポイントの場合、スイッチは非アクティブになってから5分後にエントリをエージング アウトします。リモート エンドポイントは、再度アクティブになるとテーブルにすぐに再入力されます。

（注） バージョン 1.3(1g) では、仮想ホストおよびローカル ホストに対してトリガーされるサイレント ホスト トラッキングが追加されています。

エンドポイントが再度キャッシュされるまでリモート リーフスイッチで適用されるポリシー以外にテーブルにリモートエンドポイントがなくても、パフォーマンスのペナルティはありません。

ブリッジドメインのサブネットが *enforced* に構成されている場合、エンドポイント保持ポリシーは次のように動作します。

- ブリッジドメインのサブネットに含まれていない IP アドレスを持つ新しいエンドポイントは学習されません。
- デバイスが追跡に応答しない場合、学習済みのエンドポイントはエンドポイント保持キャッシュからエージアウトします。

この実施プロセスは、サブネットがブリッジドメインで定義されているかどうか、またはサブネットが EPG で定義されているかどうかに関係なく、同じように動作します。

エンドポイントの保持タイマー ポリシーは変更できます。静的エンドポイントの MAC および IP アドレスを設定すると、保持タイマーをゼロに設定することで、スイッチ キャッシュに永久的に保存できます。エントリの保持タイマーをゼロに設定することは、それが自動的に削除されないことを意味します。この操作は慎重に行う必要があります。エンドポイントが移動したりポリシーが変化する場合は、 APIC を介してエントリを手動で最新情報に更新する必要があります。保持タイマーがゼロ以外の場合、この情報は APIC の介入なしで各パケットで確認されほぼ瞬時に更新されます。

エンドポイントの保持ポリシーは、プルーニングがどのように行われるかを決定します。ほとんどの場合、デフォルトのポリシーアルゴリズムが使用されます。エンドポイントの保持ポリシーを変更すると、システムパフォーマンスに影響を与える場合があります。何千ものエンドポイントと通信するスイッチの場合、エージング間隔を短くすると、多数のアクティブなエンドポイントをサポートするのに使用可能なキャッシュ ウィンドウの数が増えます。エンドポイントの数が 10,000 を超える場合は、複数のスイッチにエンドポイントを分散させることを推奨します。

デフォルトのエンドポイント保持ポリシーの変更に関しては、次のガイドラインに従ってください。

- リモート バウンス間隔 =（リモート エージ * 2）+ 30 秒
  - 推奨されるデフォルト値：
    - ローカル エイジ = 900 秒
    - リモート エージ = 300 秒
    - バウンス エージ = 630 秒
- アップグレードに関する考慮事項：リリース 1.0(1k) より前の ACI バージョンにアップグレードする場合は、テナント共通のエンドポイント保持ポリシー（`epRetPol`）のデフォルト値が次のようになっていることを確認してください: バウンス期間 = 660 秒。

IP エンドポイントの学習動作

ACI ブリッジ ドメインがユニキャスト ルーティングを有効にして構成されている場合、MAC アドレスを学習するだけでなく、MAC アドレスに関連付けられた IP アドレスも学習します。

ACI は MAC アドレスを追跡し、ブリッジ ドメインごとに一意である必要があります。ACI では、エンドポイントは単一の MAC アドレスに基づいていますが、任意の数の IP アドレスをブリッジ ドメインの単一の MAC アドレスに関連付けることができます。ACI は、これらの IP アドレスを MAC アドレスにリンクします。MAC アドレスが、IP アドレスのみを持つエンドポイントを表す場合があります。

したがって、ACI は次のようにローカルエンドポイントを学習して保存する場合があります。

- MAC アドレスのみ
- 単一の IP アドレスを持つ MAC アドレス
- 複数の IP アドレスを持つ MAC アドレス

3 番目のケースは、サーバーがプライマリおよびセカンダリ IP アドレスなど、同じ MAC アドレスに複数の IP アドレスを持っている場合に発生します。また、ACI ファブリックがファブリック上のサーバーの MAC アドレスと IP アドレスを学習したが、サーバーの IP アドレスがその後変更された場合にも発生する可能性があります。これが発生すると、ACI は MAC アドレスを保存し、古い IP アドレスと新しい IP アドレスの両方にリンクします。ACI ファブリックがエンドポイントをベース MAC アドレスでフラッシュするまで、古い IP アドレスは削除されません。

ACI でのローカル エンドポイントの移動には、主に 2 つのタイプがあります。

- MAC アドレスが別のインターフェイスに移動する場所
- IP アドレスが別の MAC アドレスに移動する場所

MAC アドレスが別のインターフェイスに移動すると、ブリッジドメインの MAC アドレスにリンクされているすべての IP アドレスも一緒に移動します。ACI ファブリックは、IP アドレスのみが移動した場合（および新しい MAC アドレスを受信した場合）も移動を追跡します。これは、たとえば、仮想サーバーの MAC アドレスが変更され、新しい ESXI サーバー（ポート）に移動された場合に発生する可能性があります。

VRF 内の複数の MAC アドレスに IP アドレスが存在する場合、これは IP フラップが発生したことを示しています（これは、ファブリック転送の決定に悪影響を与える可能性があります）。これは、レガシー ネットワークの 2 つの個別のインターフェイスでの MAC フラッピング、またはブリッジドメインでの MAC フラップに似ています。

IP フラップが発生する可能性のあるシナリオの 1 つは、サーバーのネットワーク情報カード（NIC）ペアがアクティブ/アクティブに設定されているが、その2つが単一の論理リンク（ポート チャネルや仮想ポート チャネルなど）で接続されていない場合です。このタイプのセットアップにより、単一の IP アドレス（仮想マシンの IP アドレスなど）が、ファブリック内の 2 つの MAC アドレス間を常に移動する可能性があります。

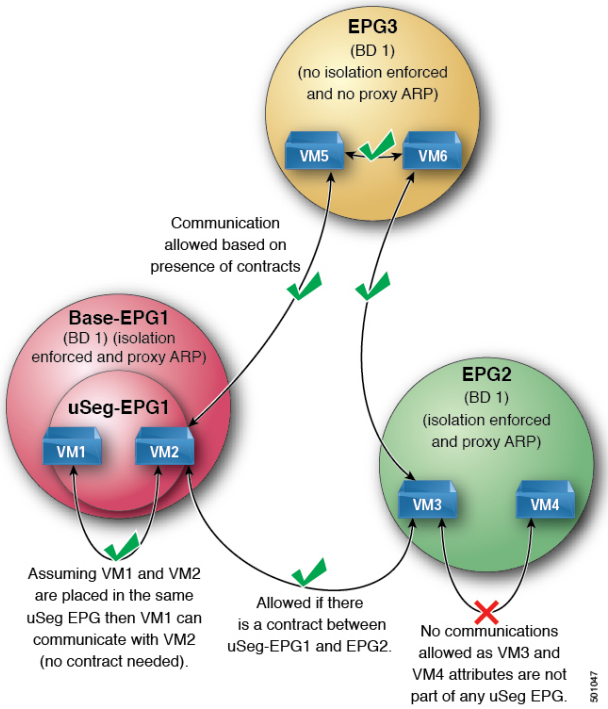
この種の動作に対処するには、NIC ペアを VPC の 2 つのレッグとして構成して、アクティブ/アクティブ セットアップを実現することをお勧めします。サーバー ハードウェアがアクティブ/アクティブ構成（ブレード シャーシなど）をサポートしていない場合、アクティブ/スタンバイ タイプの NIC ペア構成も IP フラッピングの発生を防ぎます。

プロキシ ARP について

Cisco ACI のプロキシ ARP は、ネットワークまたはサブネット内のエンドポイントが、別のエンドポイントの MAC アドレスを知らなくても、そのエンドポイントと通信できるようにします。プロキシ ARP はトラフィックの宛先場所を知っており、代わりに、最終的な宛先として自身の MAC アドレスを提供します。

プロキシ ARP を有効にするには、EPG 内エンドポイント分離を EPG で有効にする必要があります。詳細については、次の図を参照してください。EPG 内エンドポイント分離と Cisco ACI の詳細については、「*Cisco ACI* 仮想化ガイド」を参照してください。

図 *11* : プロキシ *ARP* および *Cisco APIC*



Cisco ACI ファブリック内のプロキシ ARP は従来のプロキシ ARP とは異なります。通信プロセスの例として、プロキシ ARP が EPG で有効になっているとき、エンドポイント A が ARP 要求をエンドポイント B に送信し、エンドポイント B がファブリック内で学習される場合、エンドポイント A はブリッジ ドメイン（BD）MAC からプロキシ ARP 応答を受信します。エンドポイント A が B、エンドポイントの ARP 要求を送信し、エンドポイント B はすでに ACI ファブリック内で学習しない場合は、ファブリックはプロキシ ARP の BD 内で要求を送信します。エンドポイント B は、ファブリックに戻る要求、このプロキシ ARP に応答します。この時点では、ファブリックはプロキシ ARP エンドポイント A への応答を送信しませんが、エンドポイント B は、ファブリック内で学習します。エンドポイント A は、エンドポイント B に別の ARP 要求を送信する場合、ファブリックはプロキシ ARP 応答から送信 BD mac です。

次の例ではプロキシ ARP 解像度がクライアント VM1 と VM2 間の通信の手順します。

1. VM2 通信を VM1 が必要です。

図 *12 : VM2* 通信を *VM1* が必要です。

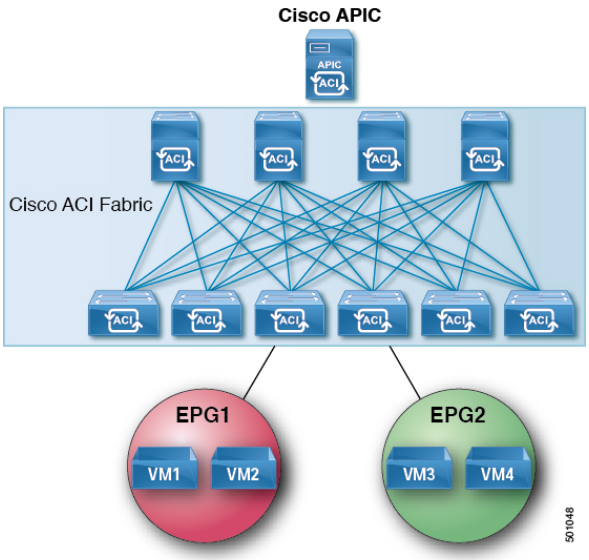


表 *2 : ARP* 表の説明

デバイス	状態
VM1	IP = * MAC = *
ACI ファブリック	IP = * MAC = *
VM2	IP = * MAC = *

2. VM1 は、ブロードキャスト MAC アドレスとともに ARP 要求を VM2 に送信します。

図 *13 : VM1* はブロードキャスト *MAC* アドレスとともに *ARP* 要求を *VM2* に送信します

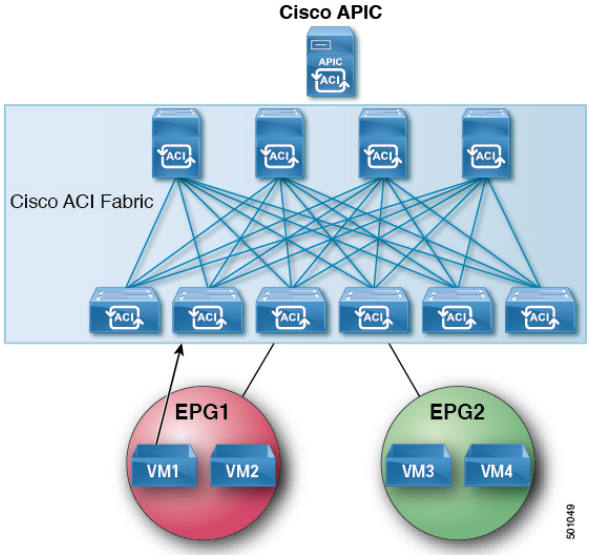


表 *3 : ARP* 表の説明

デバイス	状態
VM1	IP = VM2 IP; MAC = ?
ACI ファブリック	IP = VM1 IP; MAC = VM1 MAC
VM2	IP = * MAC = *

3. ACI ファブリックは、ブリッジ ドメイン（BD）内のプロキシ ARP 要求をフラッディングします。

図 *14 : ACI* ファブリックは *BD* 内のプロキシ *ARP* 要求をフラッディングします

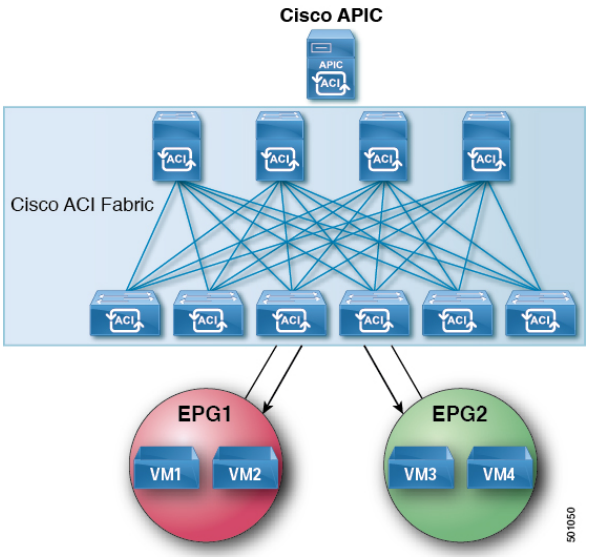


表 *4 : ARP* 表の説明

デバイス	状態
VM1	IP = VM2 IP; MAC = ?
ACI ファブリック	IP = VM1 IP; MAC = VM1 MAC
VM2	IP = VM1 IP; MAC = BD MAC

4. VM2 は、ARP 応答を ACI ファブリックに送信します。

図 *15 : VM2* は *ARP* 応答を *ACI* ファブリックに送信します

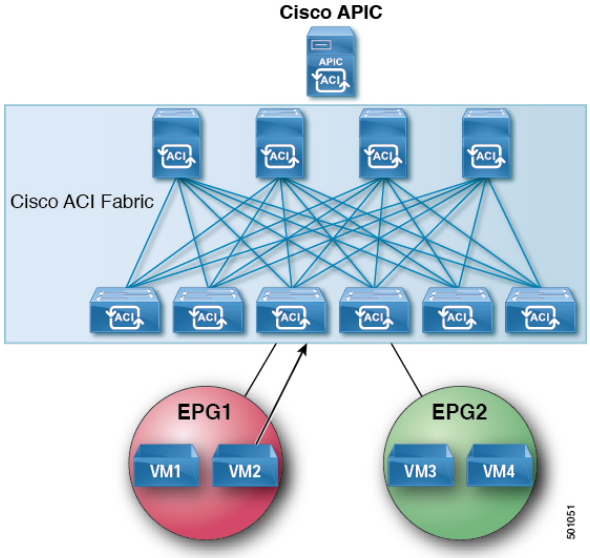


表 *5 : ARP* 表の説明

デバイス	状態
VM1	IP = VM2 IP; MAC = ?
ACI ファブリック	IP = VM1 IP; MAC = VM1 MAC
VM2	IP = VM1 IP; MAC = BD MAC

5. VM2 が学習されます。

図 *16 : VM2* が学習されます

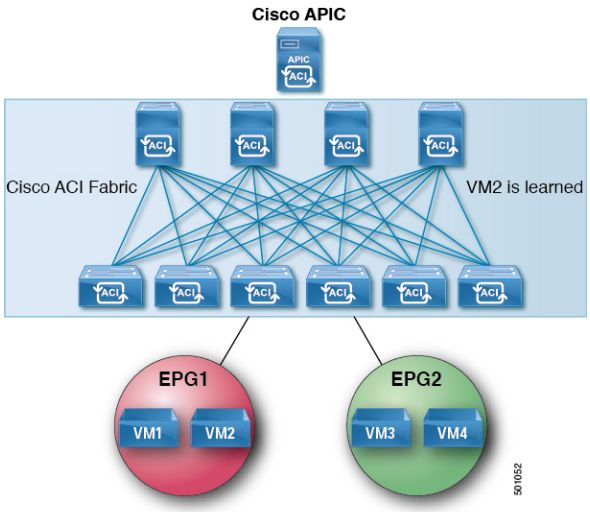


表 *6 : ARP* 表の説明

デバイス	状態
VM1	IP = VM2 IP; MAC = ?
ACI ファブリック	IP = VM1 IP; MAC = VM1 MAC IP = VM2 IP; MAC = VM2 MAC
VM2	IP = VM1 IP; MAC = BD MAC

6. VM1 は、ブロードキャスト MAC アドレスとともに ARP 要求を VM2 に送信します。

図 *17 : VM1* はブロードキャスト *MAC* アドレスとともに *ARP* 要求を *VM2* に送信します

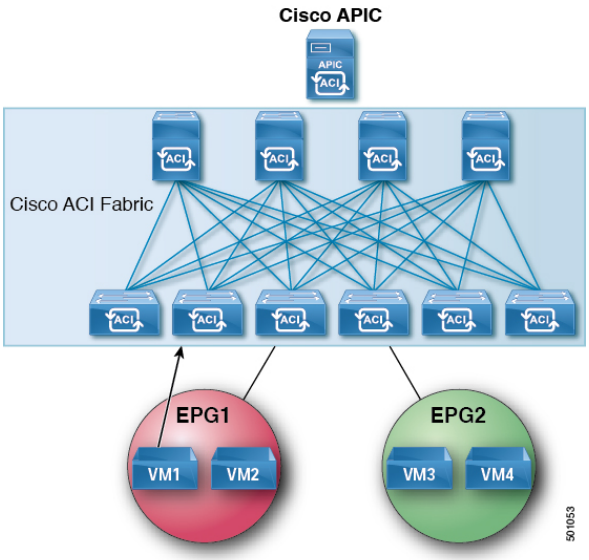


表 *7 : ARP* 表の説明

デバイス	状態
VM1	IP = VM2 IP MAC = ?
ACI ファブリック	IP = VM1 IP; MAC = VM1 MAC IP = VM2 IP; MAC = VM2 MAC
VM2	IP = VM1 IP; MAC = BD MAC

7. ACI ファブリックは、プロキシ ARP VM1 への応答を送信します。

図 *18 : ACI* ファブリック *VM1* にプロキシ *ARP* 応答を送信します。

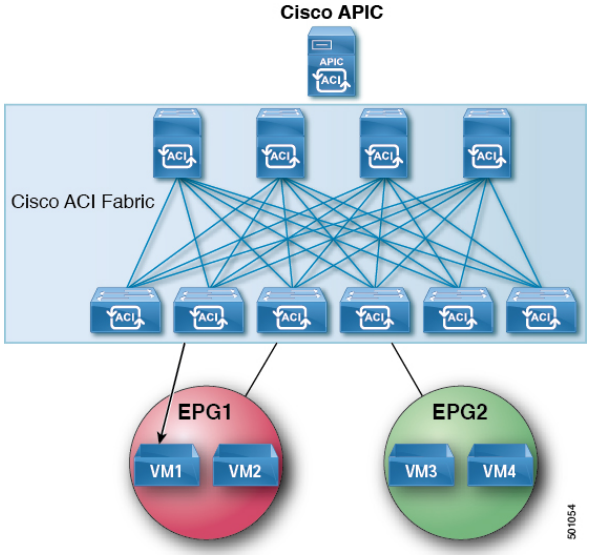


表 *8 : ARP* 表の説明

デバイス	状態
VM1	IP = VM2 IP;MAC = BD MAC
ACI ファブリック	IP = VM1 IP; MAC = VM1 MAC IP = VM2 IP; MAC = VM2 MAC
VM2	IP = VM1 IP; MAC = BD MAC

ループ検出

Cisco Application Centric Infrastructure（ACI）ファブリックは、Cisco ACI アクセス ポートに接続されているレイヤ2ネットワーク セグメントのループを検出できるグローバルなデフォルト ループ検出ポリシーを提供します。これらのグローバル ポリシーはデフォルトで無効になっていますが、ポート レベルのポリシーはデフォルトで有効になっています。グローバル ポリシーを有効にすると、個々のポート レベルで無効にされていない限り、すべてのアクセス ポート、仮想ポート、および仮想ポート チャネル（VPC）でポリシーが有効になります。

Cisco ACI ファブリックは、スパニングツリープロトコル（STP）に参加していません。代わりに、ループを検出するために、ミスケーブル プロトコル（MCP）を実装します。MCP は、外部レイヤ 2 ネットワークで実行されている STP と補完的に機能します。

（注） スパニングツリーを実行し、Cisco ACI ファブリックに接続されている外部スイッチからのインターフェイスは、loop_inc ステータスになる可能性があります。外部スイッチからのポートチャネルをフラッピングすると、問題が解決します。外部スイッチでBDPUフィルタを有効にするか、ループガードを無効にすると、問題を回避できます。

ファブリック管理者は、Cisco ACI ファブリックによって開始された MCP パケットを識別するために MCP が使用するキーを提供します。管理者は、MCP ポリシーがループを識別する方法と、ループに対処する方法（syslog のみ、またはポートを無効にする）を選択できます。

VM の移動などのエンドポイントの移動は正常ですが、頻度が高く、移動の間隔が短い場合は、ループの兆候である可能性があります。個別のグローバルなデフォルトエンドポイント移動ループ検出ポリシーを使用できますが、デフォルトでは無効になっています。管理者は、移動検出ループに対処する方法を選択できます。

また、エラー無効化の回復ポリシーは、管理者が構成できる間隔の後に、検出をループするポートを有効にし、BPDU ポリシーを無効にすることができます。

MCP はネイティブ VLAN モードで実行され、デフォルトでは、送信される MCP BPDU に VLAN タグが付けられません。MCP は、ネイティブ VLAN で送信されたパケットがファブリックによって受信された場合、ケーブル接続の誤りによるループを検出できますが、EPG VLAN の非ネイティブ VLAN にループがある場合は検出されません。リリース 2.0(2) 以降、Cisco Application Policy Infrastructure Controller（APIC）は構成された EPG 内のすべての VLAN で MCP BPDU の送信をサポートしているため、それらの VLAN 内のループが検出されます。新しい MCP 構成モードでは、送信される PDU に各 EPG VLAN ID を持つ 802.1Q ヘッダーを追加することにより、物理ポートが属するすべての EPG VLAN で MCP PDU が送信されるモードで動作するように MCP を構成できます。

3.2(1) リリース以降、Cisco ACI ファブリックは 100 ミリ秒から 300 秒の送信頻度でより高速なループ検出を提供します。

5.2(3) リリース以降では、構成にインターフェイスごとに 256 を超える VLAN がある場合、障害 F4268 が生成されます。構成にリーフスイッチごとに 2000 を超える論理ポート（ポート x VLAN）がある場合、障害 F4269 が生成されます。

（注） VLAN ごとの MCP は、インターフェイスごとに 256 の VLAN でのみ実行されます。256 を超える VLAN がある場合、最初の数値の 256 VLAN が選択されます。

MCP は、Fabrix Extender（FEX）ホスト インターフェイス（HIF）ポートではサポートされていません。

Mis-cabling プロトコルのモード

MCP は 2 つのモードで動作できます。

- 非厳格モード：MCP対応ポートが*UP*の場合、データトラフィックとコントロールプレーン トラフィック（STP、MCP プロトコル パケットなど）が受け入れられます。MCP はリンクのループを監視し、ループが検出されると、リンクはエラー ディセーブルになります。このモードでは、グローバル MCP インスタンス ポリシーに従って、パケットが 2 秒間隔で送信されます。このモードでのループ検出のデフォルト時間は 7 秒です。

- 厳格モード：リリース 5.2(4) 以降、MCP は厳格モードをサポートします。MCP が有効になっているポートが起動するとすぐに、ループをチェックするために MCP パケットが短期間、アグレッシブな間隔で送信されます。これは早期ループ検出フェーズと呼ばれ、データトラフィックを受け入れる前に、リンク（ポートに接続されている）にループがないかどうかがチェックされます。ループが検出されると、ポートはエラーディセーブルになり、シャットダウンされます。ループが検出されない場合、ポートはデータトラフィックの転送を開始します。MCPは、グローバルMCPインスタンス ポリシーに従って、非アグレッシブタイマーを使用してパケットの送信を開始します。

  すでに *UP* 状態のポートで MCP 厳格モードが構成されている場合、MCP はこのポートで早期ループ検出を実行しません。MCP 厳格モード構成のポートをフラップして、すぐに有効にします。

  MCP 厳格モードのイベント シーケンス：

  1. MCP対応ポートが起動すると、初期遅延タイマーが開始されます。リンクレベルの制御パケット（LLDP、CDP、STPなど）のみが受け入れられ、転送されます。この期間中、データトラフィックは受け入れられません。初期遅延タイマーは、外部 L2 ネットワークでSTPがコンバージするための時間です。デフォルト値は0ですが、トポロジと外部ネットワークでのスパニングツリーの構成方法によっては、STP が収束してループを切断するまでの時間を確保するために、初期遅延を 45 ～ 60 秒に設定することもできます（必要な場合）。外部 L2 ネットワークで STP が有効になっていない場合は、初期遅延タイマーを 0 に設定する必要があります。

  2. 猶予期間タイマーは、初期遅延タイマーが期限切れになった後に開始されます。この間ポートは、ループ検出に使用される MCP パケットをアグレッシブに送信します。この間に、早期のループ検出が行われます。ループが検出されると、ポートはエラーディセーブルになります。デフォルト値は 3 秒です。この期間中、データトラフィックは受け入れられません。

     猶予タイマー期間中、MCP は次のグローバル MCP インスタンス ポリシー構成を上書きします。

     - ループが検出されると、GUI で **[ポート無効化（Port Disable）]** チェックボックスが選択されていなくても、MCP はポートをエラー ディセーブルにします。
     - MCP 増倍率が 1 より大きい値に設定されている場合でも、単一の MCP フレームを受信するとループと見なされます。

  3. 猶予期間タイマーが期限切れになり、ループが検出されなくなると、ポートは転送ステートに移行し、データトラフィックが受け入れられます。MCPパケットは、グローバルな送信頻度構成に従って、非アグレッシブな間隔で送信されます。

MCP 厳格モードのガイドラインおよび制約事項

次のガイドラインおよび制約事項に従って、厳格モード MCP を構成します。

- MCP 厳格モードは FEX ポートではサポートされません。
- MCP 厳格モードは QinQ エッジ ポートではサポートされません。
- ポートで MCP 厳格モードが有効になっている場合、vPC 高速コンバージェンスはサポートされません。 vPC トラフィックは、収束に時間がかかります。
- 厳格モードの MCP 制御パケットは、APIC リリース 5.2(4) より前のバージョンを実行しているリーフスイッチではデコードできません。したがって、厳格ループ検出を機能させるには、MCP に参加するすべてのリーフスイッチに 5.2(4) の最小バージョンが必要です。
- APIC バージョン 5.2(4) より前のリリースにファブリックをダウングレードする前に、MCP 対応ポートで厳格モードを無効にします。厳格モードが無効になっていない場合、厳格ループ検出は以前のバージョンのスイッチでは機能しません。
- リーフスイッチのポートで厳格モードが有効になっている場合、そのポートで特定の VLAN が有効になっていない場合でも、STP BPDU は受け入れられます。リモート側で設定不備があると、外部 L2 スイッチで意図しない STP 状態が発生する可能性があります。
- 厳格モードが有効になっているか、送信頻度が 2 秒未満の場合は、MCP CoPP バーストと CoPP レートを 5000 に設定します。
- 2 つの MCP 厳格対応ポートが同時に 起動 すると、どちらかの側でループが検出されたときに両方のポートがエラー無効になる可能性があります。

不正なエンドポイントの検出

不正なエンドポイントの制御ポリシーについて

不正なエンドポイントは、リーフ スイッチを頻繁に攻撃し、異なるリーフ スイッチ ポートにパケットを繰り返し挿入し、802.1Q タグを変更する（エンドポイントの移動をエミュレートする）ことで、学習されたクラスと EPG ポートを変更します。誤設定により頻繁に IP アドレスと MAC アドレスが変更 (移動する) されることになります。

ファブリックの急速な移動などで、大きなネットワークの不安定状態、高い CPU 使用率、まれなケースでは、大量かつ長期のメッセージおよびトランザクション サービス (MTS) バッファ消費のため、エンドポイント マッパー (EPM) および EPM クライアント (EPMC) がクラッシュすることになります。また、このような頻繁な移動により、EPM および EPMC ログが非常にすばやくロール オーバーされ、無関係なエンドポイントのデバッグを妨害する可能性があります。

不正なエンドポイントの制御機能は脆弱性にすばやく対処します。

- 急速に移動する MAC および IP エンドポイントの特定。

- エンドポイントを一時的に静的にして、エンドポイントを隔離することによって移動を停止します。
- 3.2(6) リリースより前：**不正 EP 検出間隔**のエンドポイントを静的に維持し、不正エンドポイントとの間のトラフィックをドロップします。この時間が経過すると、不正な MAC アドレスまたは IP アドレスが削除されます。
- 3.2(6) リリース以降：**不正な EP 検出間隔**のエンドポイントを静的に維持（この機能はトラフィックをドロップしなくなりました）。この時間が経過すると、不正な MAC アドレスまたは IP アドレスが削除されます。
- ホスト トラッキング パケットを生成して、影響を受ける MAC または IP アドレスをシステムが再学習できるようにします。
- 修正アクションを有効にするための障害の発生。

不正なエンドポイント制御ポリシーはグローバルに設定されており、他のループ防止方法とは異なり、個々のエンドポイント レベルの機能です (IP および MAC アドレス)。ローカルまたはリモートの移動を区別していません。いかなる種類のインターフェイスの変更も、エンドポイントを隔離する必要があるかどうかを決定する際に移動と見なされます。

不正なエンドポイント制御機能は、デフォルトで無効になっています。

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。