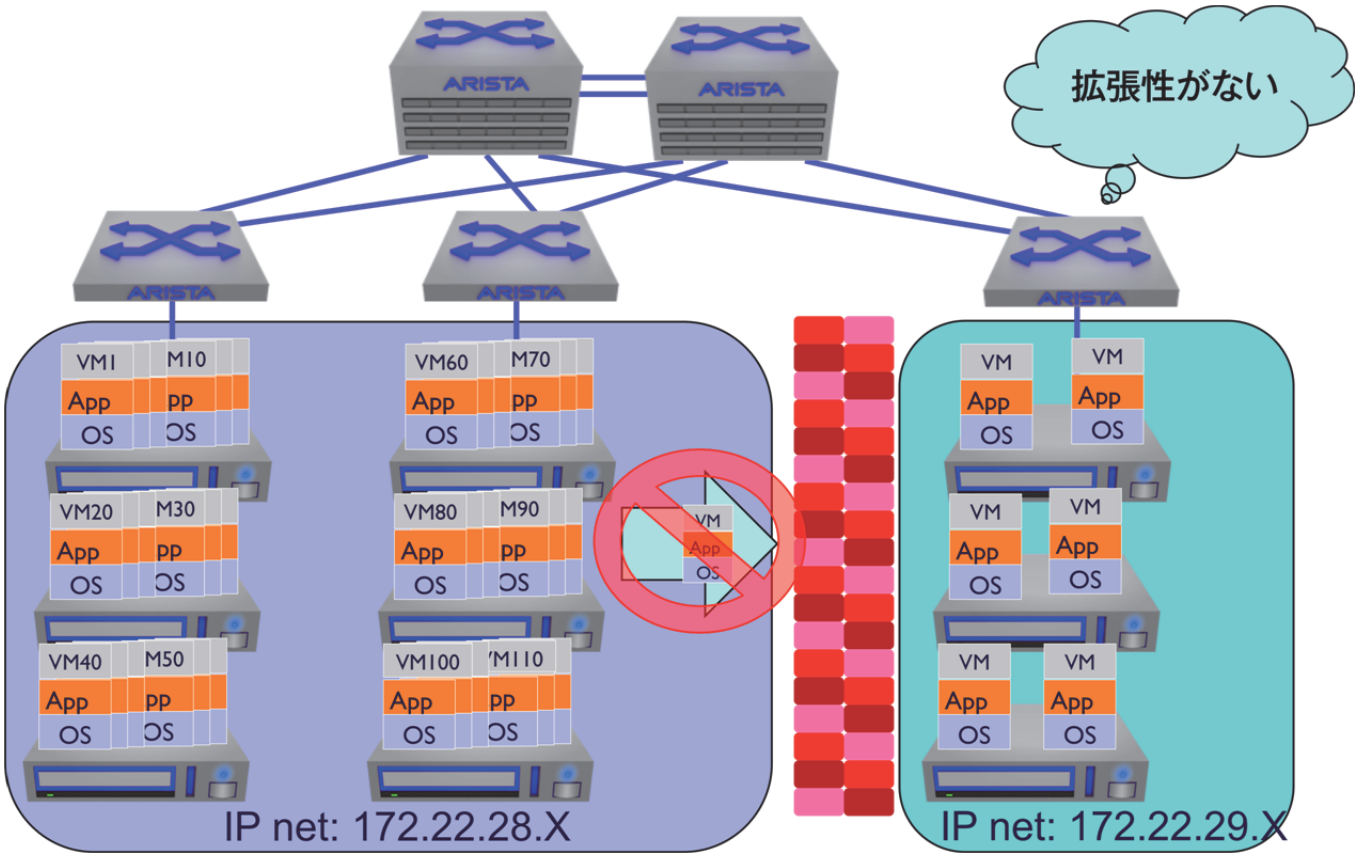
ホワイトペーパー

VXLAN（Virtual Extensible LAN）の概要

この文書では、VXLANのしくみについて概要を説明します。また、仮想インフラの導入時に、いつどの部分でVXLANを使えばよいかという判断材料も提示します。VXLANは、アリスタ、Broadcom、Intel、VMwareをはじめとする各社で仕様を策定した規格です。仮想データセンターの拡張性を向上させます。

VMwareのvSphereをはじめとする仮想化の大きなメリットの1つは、データセンター内のサーバー間で仮想マシン（VM）を移動できることです。VMを実行したままでも移動が可能です。ステートフルvMotionやライブvMotionと呼ばれるこの機能によって、サーバーの管理とプロビジョニングがシンプルになります。VMの機能や可用性にも影響は及びません。vMotionをサポートするためには、VMは常にしかるべきIPサブネット内に存在しなくてはなりません。これによって、ネットワーク上の他のユーザーとVMとの間のネットワーク接続が保証されます。

残念ながら、IPサブネットの壁により、VMの移動範囲は、vSwitchが同じサブネットに属するvSphereサーバーのクラスタに限定されます。たとえば、使用率が低いサーバーにVMを移動しようとシステム管理者が考えたときには、そのVMのネットワーク接続がvMotionによって中断することがないかどうか確認しなくてはなりません。これは、小規模なサブネットのクラスタでは通常は問題とはなりませんが、サブネット、VM、サーバーの数が増えてくると、vMotionを制限するIPサブネットの壁にぶつかります。



VXLANの用途：

VXLANのレイヤ2トンネリング機能は、IPサブネットの壁を克服するものです。データセンターのサブネット分割の方法に関係なく、データセンターのどのサーバーにもVMを移動できるようになります。これによって

管理者は、信頼性の高いL3アーキテクチャをデータセンターに採用しつつ、データセンターのすべてのサーバー間でVMの移動をサポートできます。

使用例:

- 利用者向けのクラウドをプロビジョニングするホスティング・プロバイダ
- IPアドレス空間が一杯になった後もデータセンターのネットワーク・アーキテクチャを維持したいと考えるVMファーム
- 802.1qのVLANの制約を超えてマルチテナント・サービスを拡張する必要があるクラウド・サービス・プロバイダ

VXLANとは根本的に、レイヤ2の複数の(サブ)ネットワークを集約してレイヤ3のインフラにトンネリングするためのしくみです。VXLANの基本的な用途は、複数のレイヤ3ネットワークのドメインを接続して共通のレイヤ2ドメインに見せかけることです。このしくみによって、別々のネットワーク上にある仮想マシン同士が、同じレイヤ2サブネット内にあるかのように通信できます。

VTEP(Virtual Tunnel End Point)を使用して複数の仮想ネットワークを伝送

VXLANの実装:

VXLANを使用するには、以下をサポートしたネットワーク・インフラが必要です。

- マルチキャストのサポート:IGMPとPIM
- レイヤ3ルーティング・プロトコル:OSPF、BGP、IS-IS

ほとんどの部分では、ネットワーク・デバイスはVXLANトラフィックを透過的に処理します。つまり、IPにカプセル化されたトラフィックのスイッチングやルーティングは、通常のIPトラフィックと特に変わりません。VXLANゲートウェイは、Virtual Tunnel End Point(VTEP)とも呼ばれ、カプセル化と非カプセル化のサービスをVXLANに対して一元的に提供しています。VTEPには、ハイパーバイザ内の仮想ブリッジ、VXLANに対応したVMアプリケーション、VXLANに対応したスイッチング・ハードウェアのいずれかを使用できます。既存のデータセンター・インフラでネットワークを仮想化するうえで、VTEPは鍵となります。

そんなに多くは使えません…。

VXLANの各ネットワーク・セグメントは、VXLAN Network Identifier（VNI）という24ビットの一意のIDと関連付けられています。24ビットのアドレス空間であることから、仮想ネットワークの数は、802.1Qの上限である4096を超えて、最大1670万まで拡張できます。ただし、マルチキャストとネットワーク・ハードウェアの制限により、多くの環境では、実際に使用可能な仮想ネットワークの数は少なくなります。L2の論理ドメインが同じであるVMは、同じサブネットを使用し、共通のVNIが割り当てられます。L2とVNIのこのマッピングによって、VM間の通信が可能となります。　なお、VXLANを使用した場合でも、レイヤ3のアドレッシングの方法は変わりません。L2の物理ネットワークに適用されるIPアドレッシングの規則が、仮想ネットワークにも適用されます。

VXLANは、VMのMACアドレスとVNIの組み合わせによってVMを一意に識別します。このしくみが興味深いのは、重複するMACアドレスがデータセンターのドメイン内に共存できることです。重複するMACアドレスが同じVNIで共存してさえいなければ問題ありません。

VNIサブネット上の仮想マシンには、VXLANを利用するための特別な設定は必要ありません。カプセル化/非カプセル化やVNIのマッピングは、ハイパーバイザに組み込まれたVTEPが対処するからです。また、VXLAN対応のスイッチング・プラットフォームも同様で、802.1q対応のネットワーク・デバイスのカプセル化/非カプセル化のオーバーヘッドに対処します。VTEPに対しては、レイヤ2またはIPサブネットからVNIネットワークへのマッピングと、VNIからIPマルチキャスト・グループへのマッピングの設定が必要です。前者のマッピングでは、VNI/MACのトラフィック・フローのためのフォワーディング・テーブルがVTEPに作成されます。後者のマッピングでは、オーバーレイ・ネットワーク全体に対するブロードキャスト/マルチキャスト機能をVTEPがエミュレートできるようになります。VTEPの構成を同期する処理は、RANCIDのような一般的な構成管理ツールで自動化するか、またはVMware vCenter OrchestratorやOpen vSwitchなどのシステムを使って管理できます。

VXLANのフレームのカプセル化と転送：

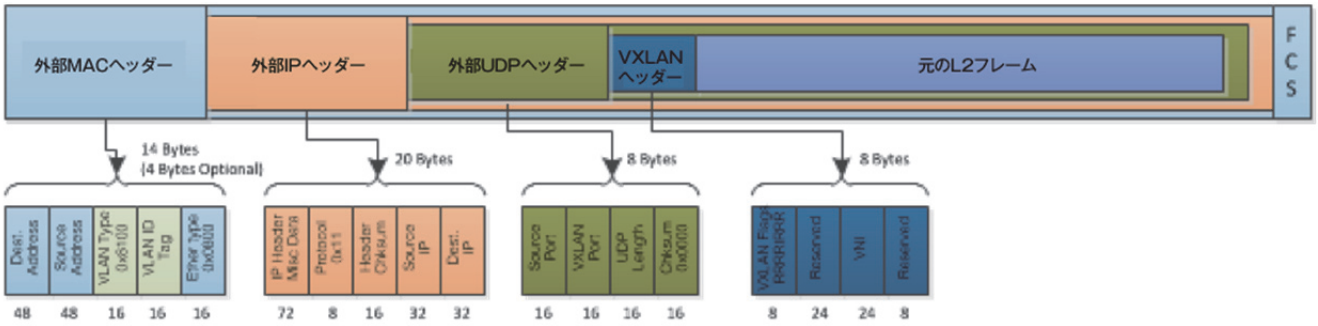
ここまでに挙げた要素を使って、VTEPは次のような形で転送規則を適用します。

1) 送信元と宛先のMACアドレスが同じホスト上にある場合は、トラフィックはvSwitchを通じてローカルでスイッチングされ、VXLANのカプセル化/非カプセル化は実行されない。
2) 宛先のMACアドレスが同じESXホスト上にない場合は、送信元のVTEPが適切なVXLANヘッダーを付けてフレームをカプセル化し、ローカルのテーブルに基づいて宛先のVTEPに転送する。宛先のVTEPは、VXLANヘッダーを外して内部フレームを取り出したうえで、宛先のVMに届ける。
3) 不明なユニキャストやブロードキャスト/マルチキャストのトラフィックの場合は、ローカルのVTEPはVXLANヘッダーを付けてフレームをカプセル化し、作成時点でVNIに割り当てられたVNIのマルチキャスト・アドレスにそのフレームをマルチキャストする。これには、すべてのARPやBoot-p/DHCP要求などが含まれる。

他のホストのVTEPは、マルチキャスト・フレームを受信して、ユニキャスト・トラフィックと同じように処理する（上記の2を参照）。

このトンネリング方法の実装は、MPLSやOTVといった他の手法に比べて比較的シンプルです。管理者は、VNIまたはIPマッピングとマルチキャスト・アドレスさえ設定すれば済むからです。残りはVTEPが管理します。

フレームの構造の詳細は次のとおりです。



VXLANヘッダーの構造

イーサネット・ヘッダー：

宛先アドレス – 宛先のVTEPが同じサブネット上の場合は、そのMACアドレスを設定します。宛先のVTEPが別のサブネット上にある場合は、ネクスト・ホップ・デバイス（通常はルーター）のアドレスを設定します。

VLAN – VXLANの実装ではオプションです。デフォルトでは、802.1QのTagged Prototocol Identifier（TPID）のEthertypeである0X8100で、該当するVLAN IDタグを持ちます。

Ethertype – IPv4のペイロード・パケットを示す0X0800に設定します。現時点ではIPv6はサポートしていませんが、今後の導入に向けた検討が進んでいます。

IPヘッダー：

プロトコル – 0x11に設定して、UDPパケットであることを示します。

送信元IP – 送信元のVTEPのIPアドレスに設定します。

宛先IP – 宛先のVTEPのIPアドレスに設定します。未知/未学習の場合や、ブロードキャスト/マルチキャスト・アドレスの場合は、VXLANはマルチキャスト・グループを使ってネットワークのブロードキャストをシミュレートします。処理の概略は次のとおりです。

a) 宛先IPを、送信元の仮想マシンのVNIに対応するIPマルチキャスト・グループに置き換える。
b) フレームがマルチキャストされ、同じVNIマルチキャスト・グループのすべてのVTEPが受信する。各VTEPはこのフレームを取り出し、送信元のID とVNIのマッピングを今後の使用のために学習したうえで、フレームの種類とローカルのフォワーディング・テーブルの情報に基づいて、パケットを転送または破棄する。
c) 対象の仮想マシンが属するVTEPが、仮想マシンの応答をカプセル化して、送信元のVTEPに転送する。
d) 送信元のVTEPが応答を受信し、IDとVNIのマッピングをやはり今後の使用のためにキャッシュする。

UDPヘッダー：

送信元ポート – 転送を行うVTEPが設定します。この値には、インナーのイーサネット・ヘッダーから算出したハッシュ値も使用できます。これにより、ポート・チャネルやECMPのハッシュ・アルゴリズムがこの値をトラフィックの負荷分散に活用できます。

VXLANポート – VXLANのIANAポート。ベンダー固有。

UDPチェックサム – 送信元のVTEPが0x0000に設定するものとします。受信側のVTEPが0x0000とは異なるチェックサムを受信した場合、フレームをチェックし、チェックサムが正しくない場合には破棄する必要があります。

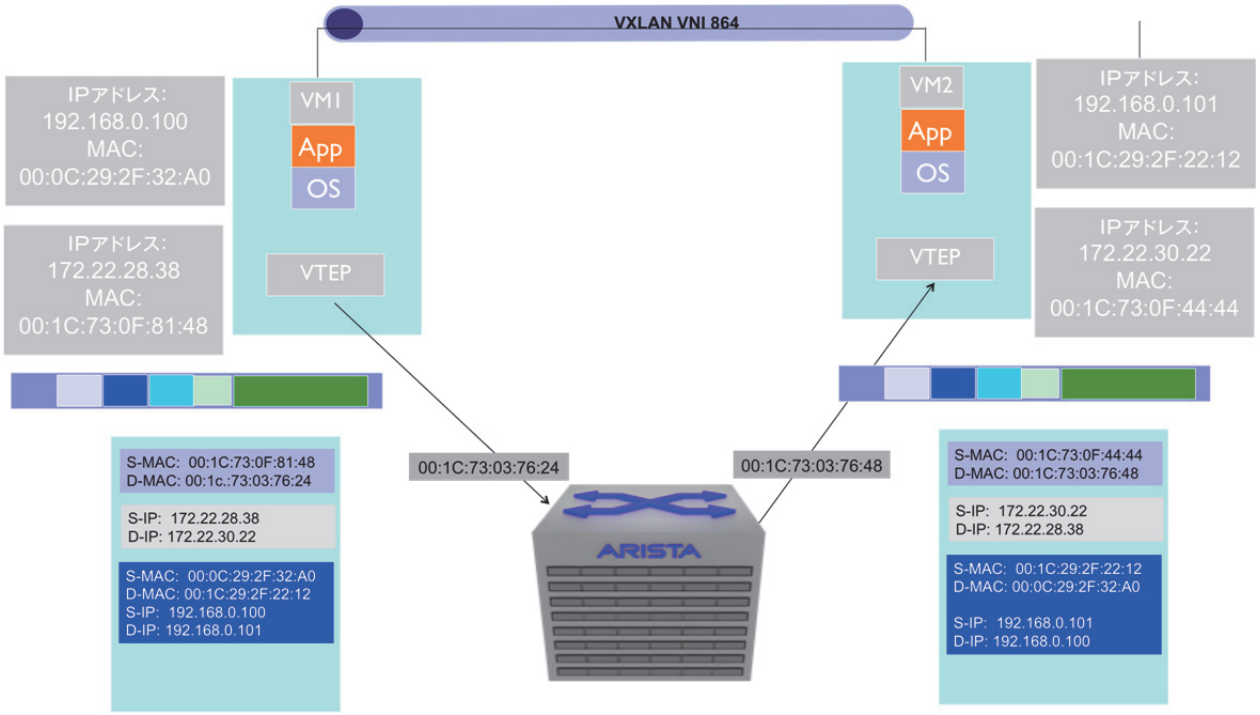
VXLANヘッダー:

VXLANフラグ – ビット3(VNIビット)を除き、予約済みのビットはすべてゼロに設定します。VNIビットは、有効なVNIについては1に設定します。

VNI – 24ビットのフィールドで、VXLANのネットワークIDを示します。

Reserved – 24ビットと8ビットの予約済みフィールドで、ゼロに設定します。

VXLANのパケットの流れ:



VXLAN:VMからVMへの通信

それぞれIPサブネットが異なる別々のホスト上にあるVM1とVM2の間でセッションが開始された場合、パケットの流れは次のようになります。ここでは、開始時点の状態と想定します。つまり、対応付けが未学習であるものとします。

- VM1は、192.168.0.101に対応するMACアドレスを要求するARPパケットを送信する。
- VTEP1は、このARPをマルチキャスト・パケットにカプセル化し、VNI 864に該当するグループにマルチキャストする。
- VNI 864に該当するすべてのVTEPがこのパケットを受信し、VTEP1とVM1のMACの対応付けを自らのテーブルに追加する。
- VTEP2は、マルチキャスト・パケットを受信し、フレームを取り出して、VNI 864内のポート・グループにフラッディングする。
- VM2は、ARPを受信し、自らのMACアドレスを伝える応答をVM1に送信する。

- VTEP2は、この応答をユニキャストのIPパケットとしてカプセル化し、VTEP1に転送する。この応答をユニキャストで送信できるのは、カプセル化されて届いた最初のARPによって、VTEP2がVTEP1とVM1のMACアドレスの対応付けを学習しているため。
- VTEP1は、応答のパケットを受信し、中身を取り出して、VM1に転送する。

この時点で、VM1とVM2の間の通信が確立され、該当するすべての状態マシンに関連付けがプログラムされました。これ以降、送信元が192.168.0.100、宛先が192.160.0.101のユニキャスト・トラフィックについては、VTEP1はそのパケットに次のようなヘッダーを付加します。

a. VNI VXLANヘッダーは864。
b. 標準のUDPヘッダー。UDPチェックサムは0x0000に設定し、VXLANの宛先ポートはベンダーに応じた適切なIANAポートに設定。
c. 送信先IPはVTEP2のIPアドレスに設定。プロトコルIDはUDP（0x011）に設定。
d. 標準のMACヘッダーに、ネクスト・ホップのMACアドレスを指定。（上の例の場合、ネクスト・ホップは、MACアドレスが00:13:73:0C:76:24のルーターのインターフェイス）。

VTEP2は、間にあるルーター経由でこのパケットを受信します。フレームを取り出す処理はUDPヘッダーの値により開始されます。VTEP2はフレームをvSwitchに送り、VNI 864に対応付けられたポート・グループに届けます。こうしてVM2にフレームが送られ、処理が行われます。応答のトラフィックも、上の例を逆にたどる形ですべて同様に処理されます。

導入上の考慮事項:

ネットワークのデータグラムのペイロードと帯域幅の使用率:

VXLANのカプセル化のヘッダーによって、イーサネット・フレームの全体のサイズは50バイト大きくなります。したがって、ジャンボ・フレームをサポートするインフラが不可欠です。また、VXLANのトラフィックに対応するうえで、帯域幅の使用が増すことも考慮する必要があります。パケット・サイズが大きくなった複数のネットワークが混在すると、帯域幅の消費が増えることから、VXLANは、10Gb以上のネットワーク・テクノロジで導入するのが賢明です。

標準のIPデータグラムを使用すると、VXLANで長距離間のvMotionやHigh Availability（HA）の導入の選択肢を提供しやすくなります。VXLANのフレームは、中のパケットの情報も加味してパケット・ヘッダーに可変性を加え、負荷分散アルゴリズムの助けとします。ただし、ネットワーク・デザイン上の関心が、障害復旧の用途や、データセンターのリモート・ミラーリングの用途におけるVXLANの活用にある場合には、VMware vMotion/HAのハートビートのラウンド・トリップの遅延が10ミリ秒を超えないことが重要です。トラフィックの優先順位サービスを備えた高帯域幅で低レイテンシーのスイッチングを活用したネットワーク・デザインによって、これらの要件を満たすことができ、仮想データセンターを拡張できます。

マルチキャストの要件:

前述のとおり、VXLANネットワーク内では、ブロードキャスト、未知のユニキャスト、マルチキャストをシミュレートするために、IPマルチキャスト・サービスを使用しています。これはVXLANの要件です。また、要件ではないものの、現在の構成では、マルチキャスト・グループとVNIを1対1で対応づけることが推奨されています。このようにすると、MACテーブルの更新の情報は、必要なVTEPのみに送られます。1つのみのマルチキャスト・アドレスをすべてのVNIで使用する方法も可能ですが、この場合、必要としないVTEPにもアドレスが実質的にフラッディングされ、ネットワーク内で不要なトラフィック・フローが発生することになります。
PIMのsparseモード、denseモード、BIDIRは、いずれもVXLANをサポートするマルチキャスト機能を提供します。管理者によっては、PIMに懸念を抱く向きもあるかもしれません。特に、CPUに負荷をかけるPIMの処理に起因するネットワーク障害を経験したことのある方です。しかし、ここで留意する必要があるのは、現在のスイッチング・プラットフォームはPIMをハードウェアでサポートしており、ネットワークのパフォーマンスや信頼性に悪影響を及ぼすことなく、PIMの大規模な展開に対応できるということです。

ARPキャッシュとMACテーブルの検討事項:

VXLANネットワーク上のVMは、VTEPを通じて非仮想ネットワークとの通信を行います。VTEPには、仮想ファイアウォールやVMware vShieldのようなソフトウェア・アプライアンスか、VXLAN対応スイッチを使用できます。いずれにせよ、VTEPがルーティング・サービスを提供する場合、不要なARPを防ぐためには、サービス対象となる仮想ネットワーク上のVMの数に対応できる規模のARPキャッシュが必要です。

VXLANのフレームにカプセル化されたVMのトラフィックは、サーバーのVTEPのMAC IDを使用します。これによって、データセンターの物理スイッチのMACアドレスのエントリは減ります。理想としては、VXLANの物理ネットワークは、VTEPのMACアドレスと、データセンター内のホストの管理インターフェイスのMACアドレスのみを学習すればよいという状況が理想的です。しかし、小規模なVMの展開であればその方法を適用できる可能性もあるものの、VMとサーバー・クラスタをサブネットに分割して、サーバー1台あたり最大で数十というVMでも維持可能なトラフィック・ボリュームに対応しておく方が賢明です。

VLANとVXLANの比較表:

機能/拡張性	802.1Q VLAN	VXLAN
仮想ネットワークの数	4K:スパニング・ツリーの規模による制限	1600万以上:ネットワークで使用するマルチキャスト・ルーターがサポートするマルチキャスト・グループ数による制限
ネットワークの規模	802.1Q VLANが許容する範囲	PIMマルチキャスト・グループが許容する範囲
ネットワークのパケット・サイズ	1.5Kまたは9K	VXLANヘッダー分として50バイトを追加
マルチキャストの要件	なし	PIM、SM、DM、またはBIDIR。グループ数に応じて仮想ネットワーク数が決定
ルーティングのサポート	802.1Q対応のルーターまたはスイッチ	VMware vShield、vEdge対応のルーターまたはスイッチ、およびVTEP対応のルーター
ARPキャッシュ	VLANあたりのVM数を制限	VMwareまたはVTEPのキャッシュがVNIあたりのVM数を制限
MACテーブル	スイッチのMACテーブルに対するVMのMACアドレス数が制限	スイッチのMACテーブルに対するVTEPのMACアドレス数が制限

まとめ:

VXLANは、レイヤ3ネットワークの境界を越えてレイヤ2サブネットを拡張するための強力な手段です。トラフィックをカプセル化してL3ゲートウェイの先に延伸することによって、VMの可搬性やvMotionの制限の問題を打開でき、異なるIPサブネット上に置かれているサーバーでVMをホストできます。またVXLANでは、データセンター・インフラ全体にわたって複数のサブネットをオーバーレイできます。仮想ネットワーク数を制限するのは、基盤のネットワークの物理的な帯域幅と、VXLANネットワークのブロードキャスト/マルチキャスト・トラフィックをシミュレートするためのマルチキャスト・サブネット数のみです。適切なハードウェアを使用すれば、ネットワークの安定性を犠牲にすることなく、802.1QのVLANの4Kという上限数をVXLANによって打破できます。カプセル化したトラフィックのルーティングには、既に確立されているIPトランスポートを使用します。したがって、リンク・アグリゲーション、ループの検出と破棄、経路探索は、OSPF、BGP、IS-ISといった実証済みのプロトコルで解決できます。VXLANは、既存のインフラにそのまま適用でき、インフラの変更は必要ありません。VMware、Intel、Broadcom、アリスタ、Open vSwitchなど各社がサポートしていることから、相互運用性が保証され、ベンダー・ロックインを回避できます。VXLANを使用したシステムによって、ネットワーク管理者はクラウド仮想化を新しいレベルにまで拡大し、これまでより多くのユーザーに経済的に対応できます。

詳細については、www.arista.com/jpを参照してください。