JANOG37

VXLANチュートリアル

シスコシステムズ合同会社

テクニカルソリューションズアーキテクト

大平　伸一

Jan.20. 2016

Agenda

- VXLAN基本知識
- VXLAN BridgingとRouting
- 標準化プロトコル VXLAN EVPN
- VXLAN EVPNの活用例
- VXLAN EVPNのコンフィグ例
- VXLAN 参考情報

VXLAN 基本知識

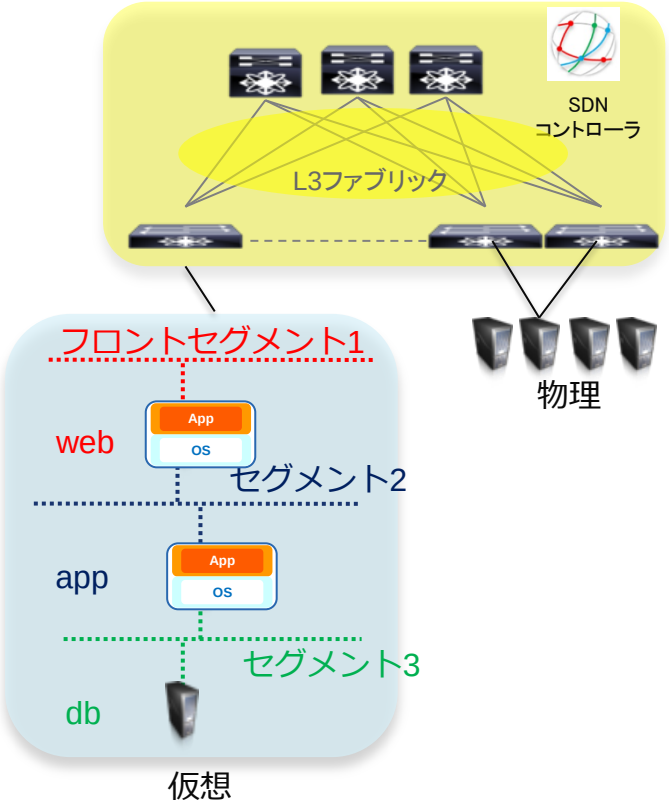
近年のデータセンターネットワークの要件

■データセンターネットワークの新たな要件

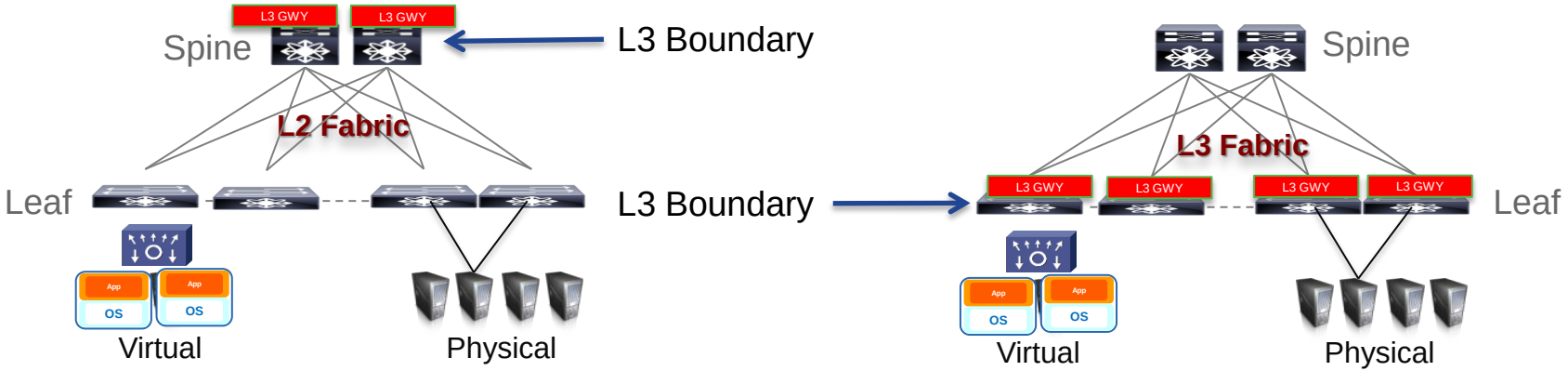
- サーバ仮想化によりマイクロセグメントの増大への対応
- ビジネスの俊敏性向上のためネットワークの自動化
- ネットワーク変更に要するコストおよび複雑性の削減

■VXLAN/SDNによってこれらの要件を満たす

- VXLAN(L3 VNI)によりユニークなテナントの識別し、マイクロセグメントはVXLAN(L2 VNI)で管理
- SDNコントローラとVXLANファブリックにより、ネットワークの自動化を容易に実現
- サーバの追加、削除、移動に伴うネットワーク変更はコントローラーによって自動追従



L2ファブリックとL3ファブリックの比較



■L2ファブリック

1. Spineスイッチで集中的なルーティング処理
2. SpineスイッチでMACテーブル管理やARP処理を行うので拡張性に難あり
3. マルチテナントの識別は、VLANに依存
4. ARPリクエストはファブリックワイドにフラッティング

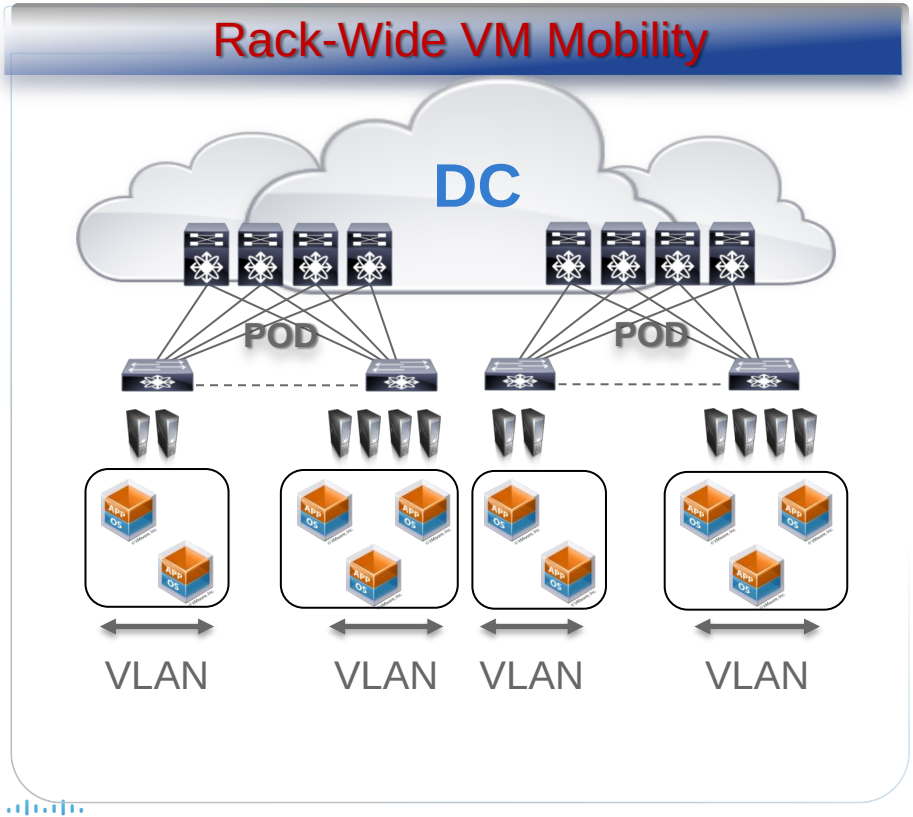
■L3 ファブリック

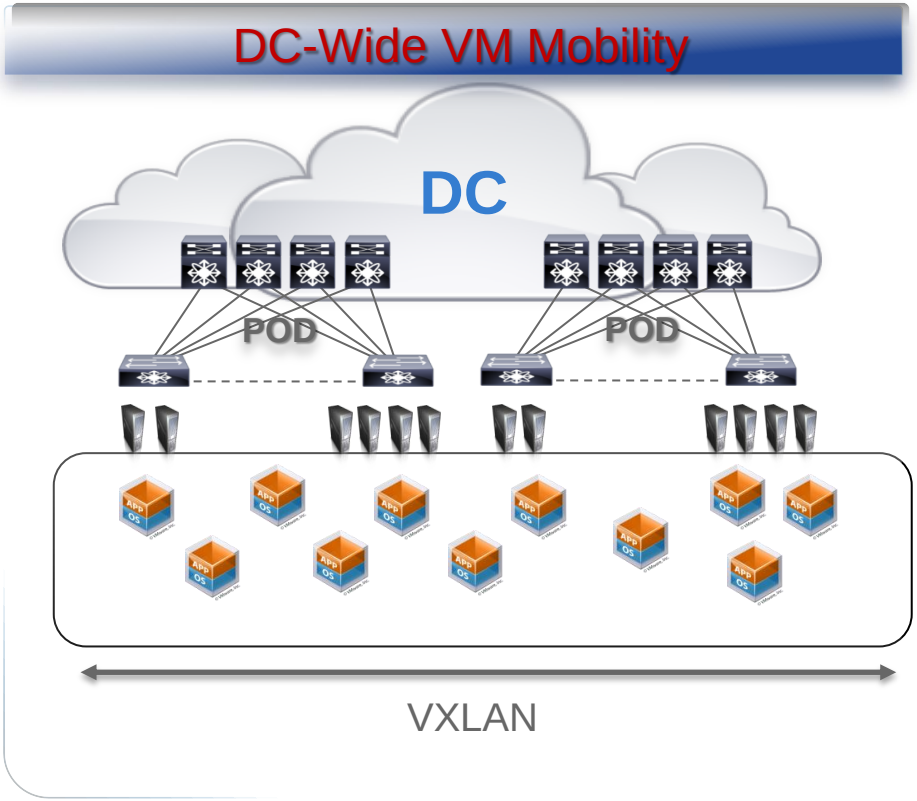
1. 複数のLeafで分散型でルーティング処理
2. 複数のLeafで分散してMACテーブル管理やARP処理などを行うので拡張性に優れる
3. マルチテナントの識別は、VXLAN（VNI）を識別子とするので高い拡張性を保持
4. LeafでARPリクエストの代理応答でき効率的なファブリックを提供

レイヤ3 ファブリックの価値

- L3ファブリックの利点
  - テナントの識別にVXLAN VNIを活用し効率的に管理
  - テナント内のマイクロセグメンテーションに対応
  - テナント間でIPアドレスの重複を許容
  - Leafでの分散 L3 ゲートウェイにより高い拡張性
  - 既存を活かし段階的なネットワーク仮想化技術を導入可能
  - SDNコントローラをOptionとして選択でき、ネットワークを抽象化することで自動化やリソース管理の最適化を実現

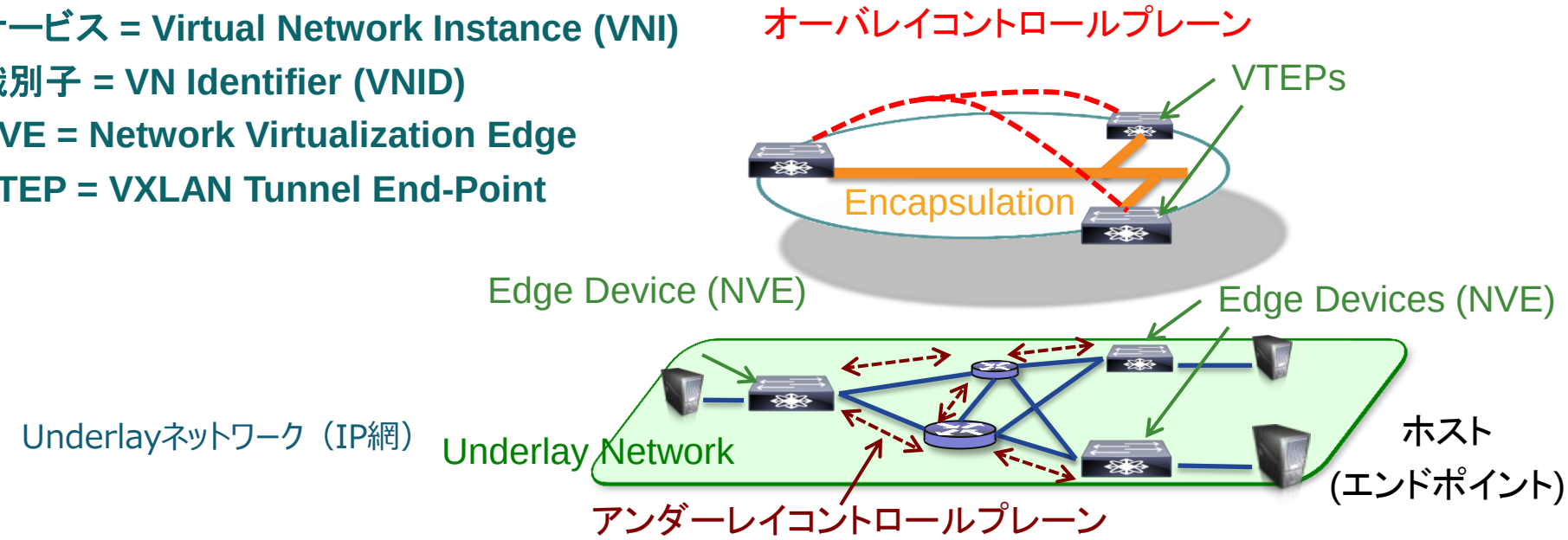
VXLANが可能にするもの





VXLANの用語

サービス = Virtual Network Instance (VNI)
識別子 = VN Identifier (VNID)
NVE = Network Virtualization Edge
VTEP = VXLAN Tunnel End-Point

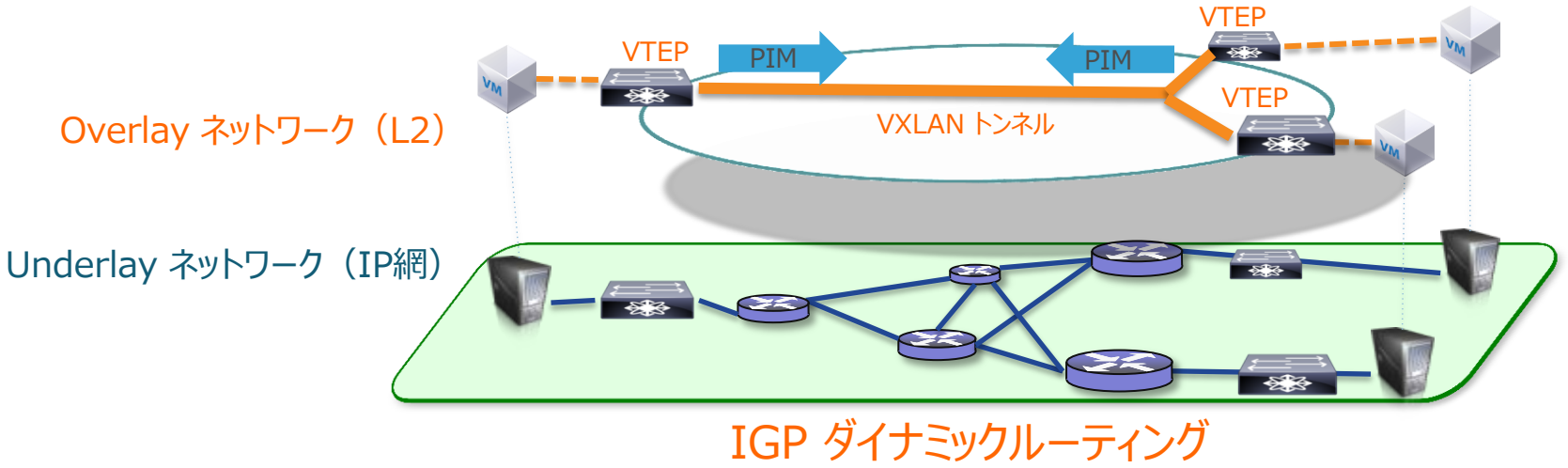


- **テナントネットワーク作成時、Underlayネットワーク構成を意識しない**
  ＝通信経路の最適化や可用性、ボトルネック回避などはアンダーレイに依存

これまでのVXLANの特長

- OverlayネットワークとUnderlayネットワークの分離

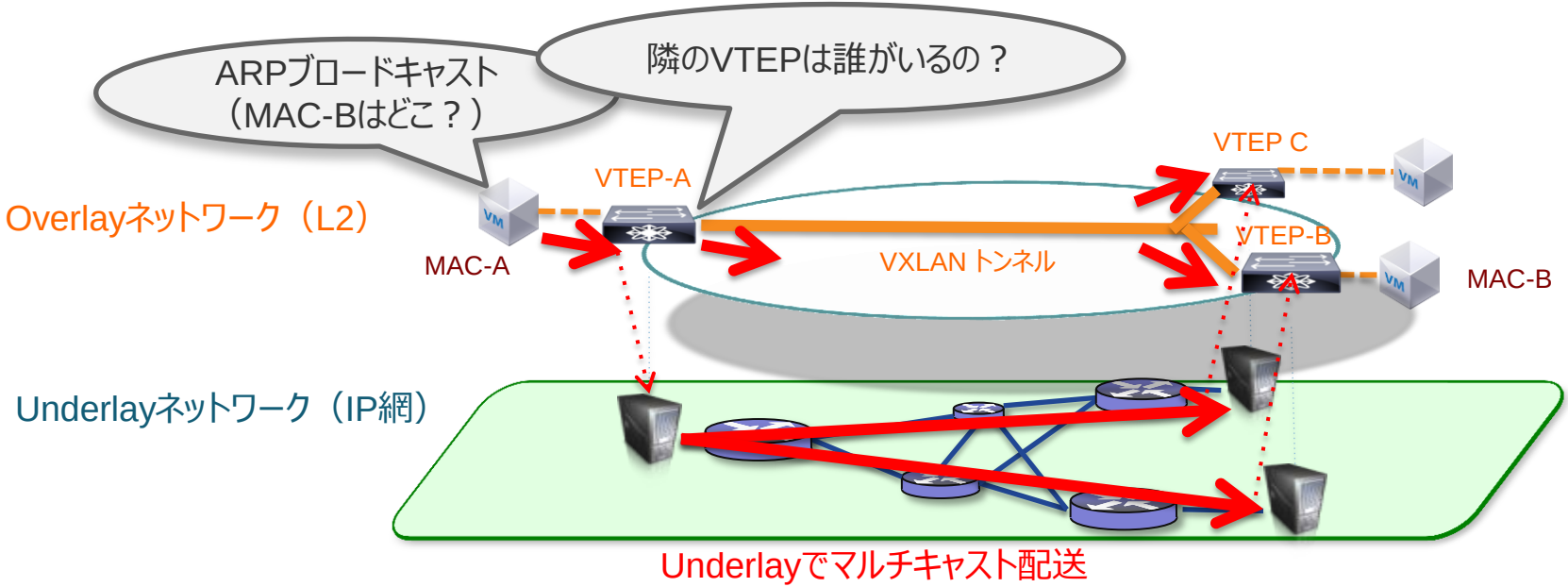
- Underlay ネットワークの構成を意識しない
- しかし、通信経路の最適化や可用性、ボトルネック回避などはUnderlay に依存
- PIM Multicast Routingで ホストへの到達可能なUndelayネットワークが必要



これまでのVXLANの特長

- BUMトラフィックやVTEP検出にマルチ キャストが必要

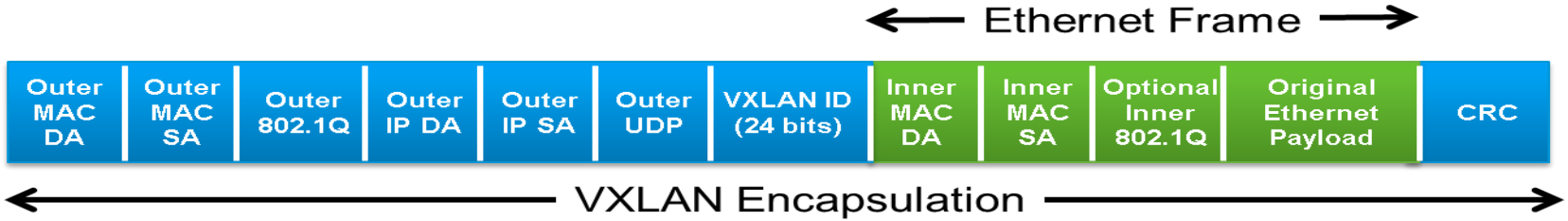
- 従来のVXLANには、そもそもコントロール プレーンの概念がないため、ネイバー VTEPやサーバ ノードを検出するため、非効率なフラッティング パケットが飛び交う



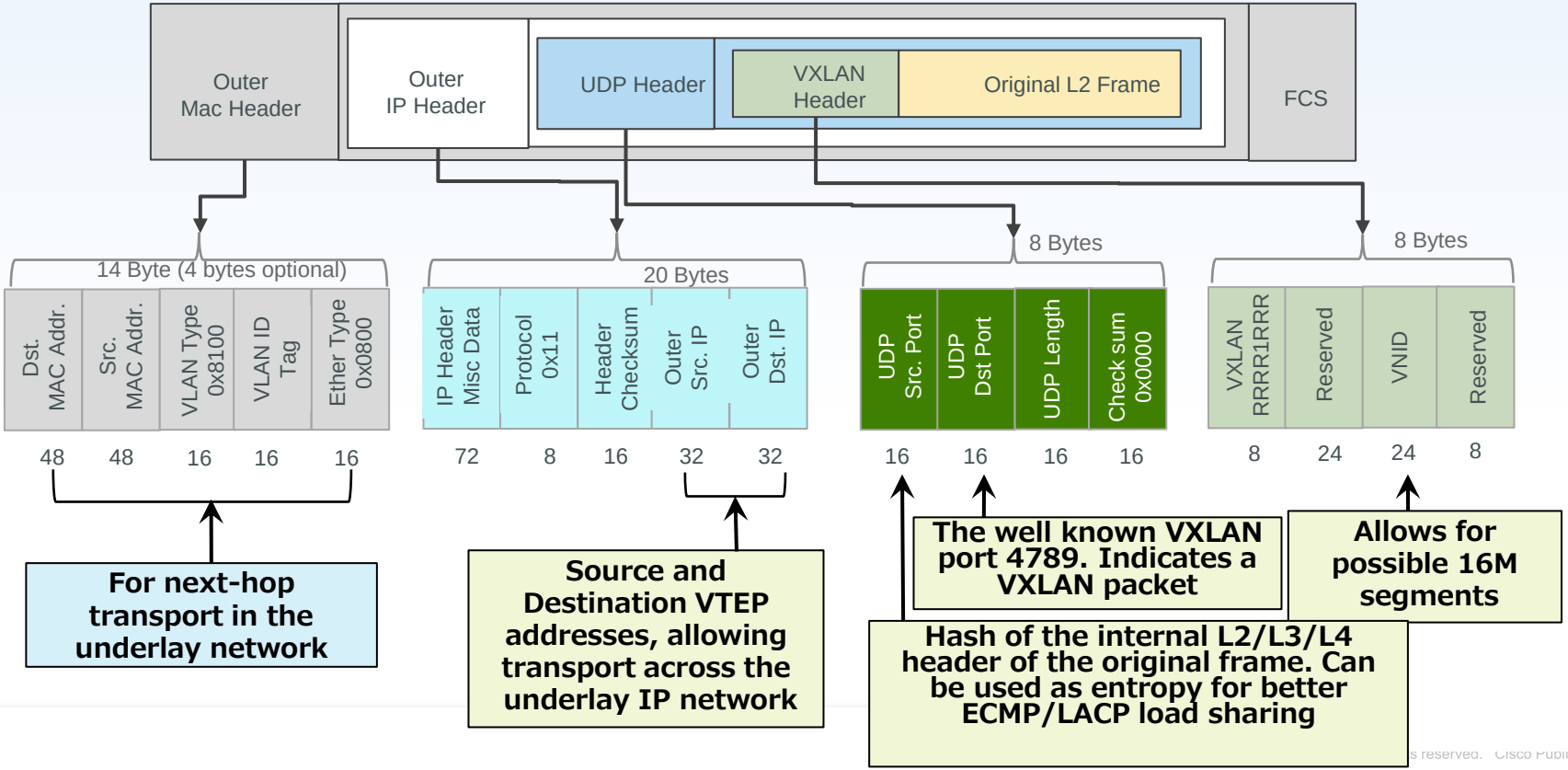
VXLANのヘッダー構造

- Virtual eXtensible Local Area Network (VXLAN)
- VLANより拡張性(eXtensible)のあるレイヤー2セグメント
- レイヤー2フレームをIP/UDPでカプセル化
  - (50 Byteのオーバーヘッド)
- RFC7348 にて標準化完了済

Cisco,VMWare,Citrix,Redhat,Broadcom,Arista



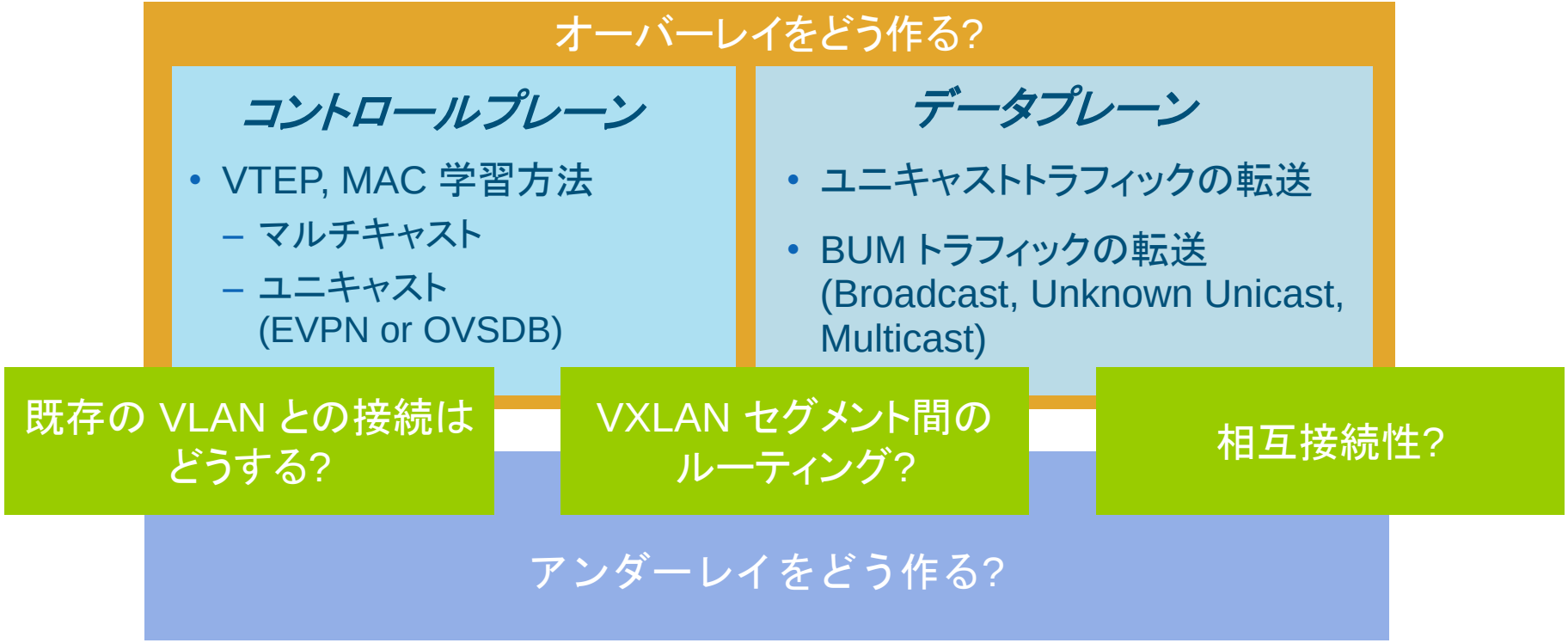
VXLANのヘッダー構造



VXLAN は、Control Planeベースに発展！！

- 過去: VXLANよりほかの Overlay技術
  - Data-Plane のみ (Multicast based Flood&Learn)
- 現在: VXLANは拡張性のあるDC Fabric
  - Control-Plane, 有効なVTEP の検出, Multicast や Unicast (Ingress Replication)
- 将来: VXLAN for DCI – DC間接続
  - DCI 向けに機能拡張 (ARP caching/suppress, マルチホーミング, 障害ドメインの隔離、ループ検出など.)

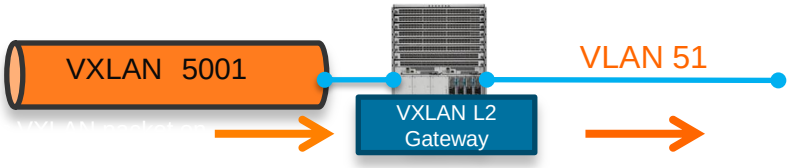
VXLAN ネットワークを作るときに考えるべきこと



VXLAN BridgingとRouting

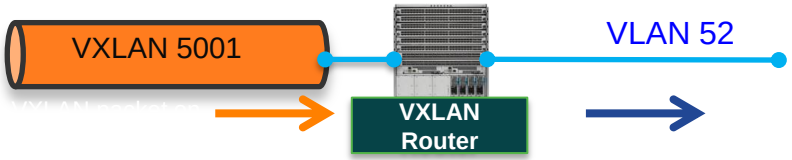
VXLAN Gateway: Bridging と Routingの違い

VXLAN to VLAN Bridging (L2 Gateway)



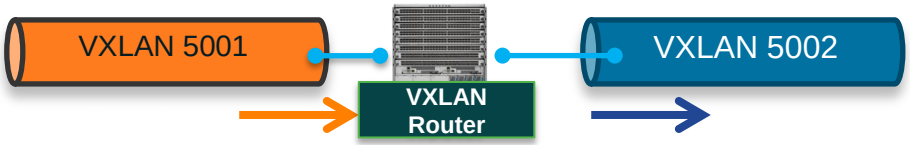
VXLAN to VLAN Routing (L3 Gateway)

Decap、Routing

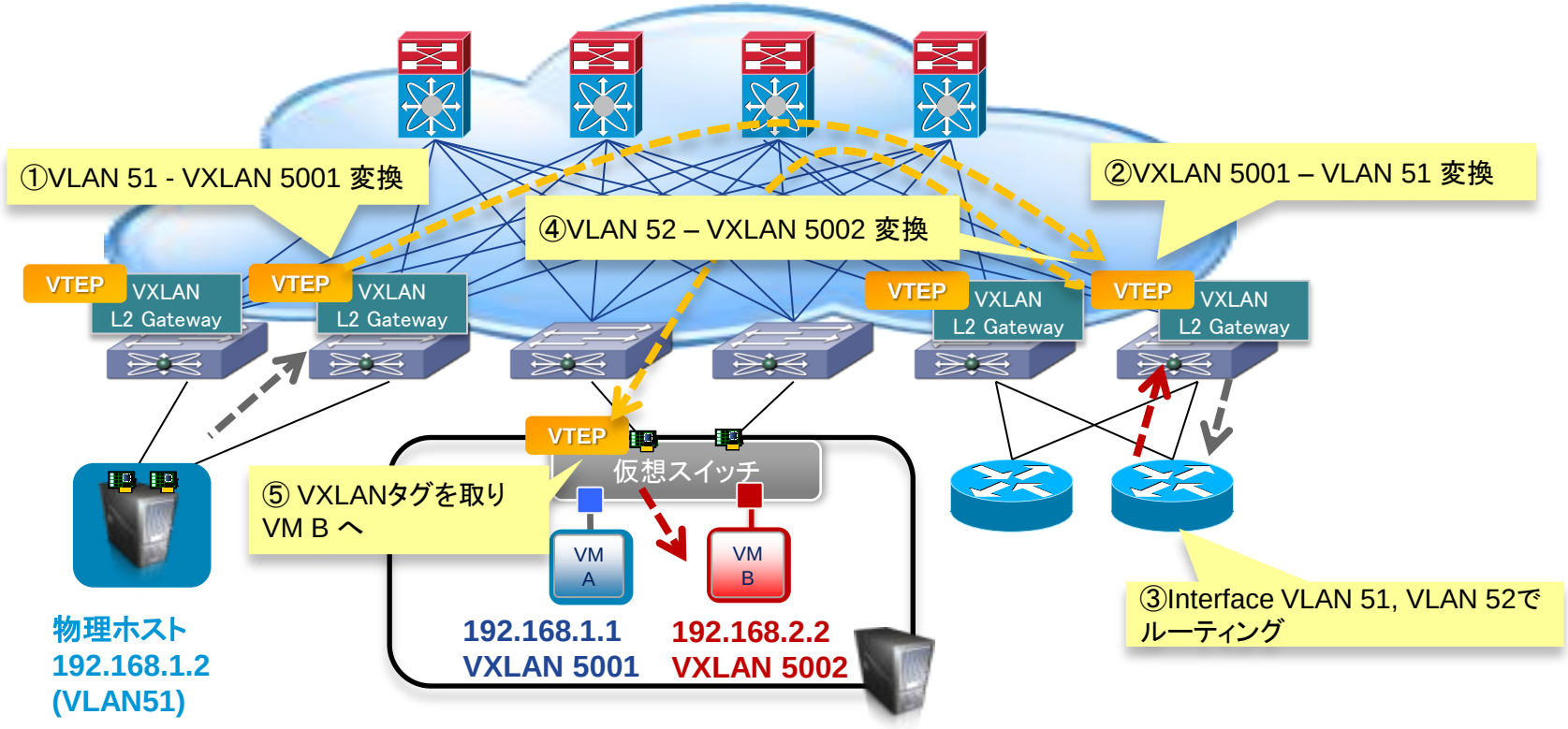


VXLAN to VXLAN Routing (L3 Gateway)

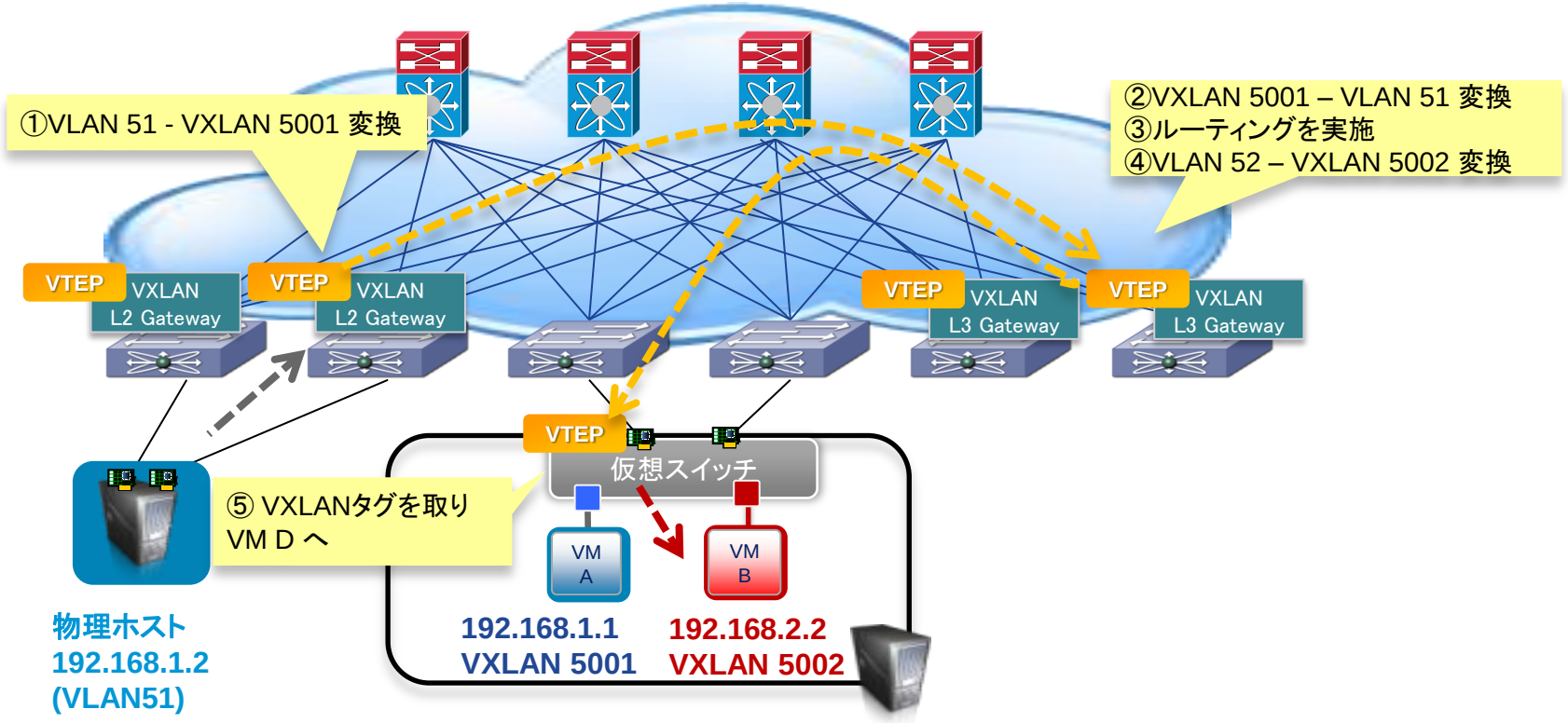
Decap, Routing, Encap



VXLAN L3 Gateway がないと…

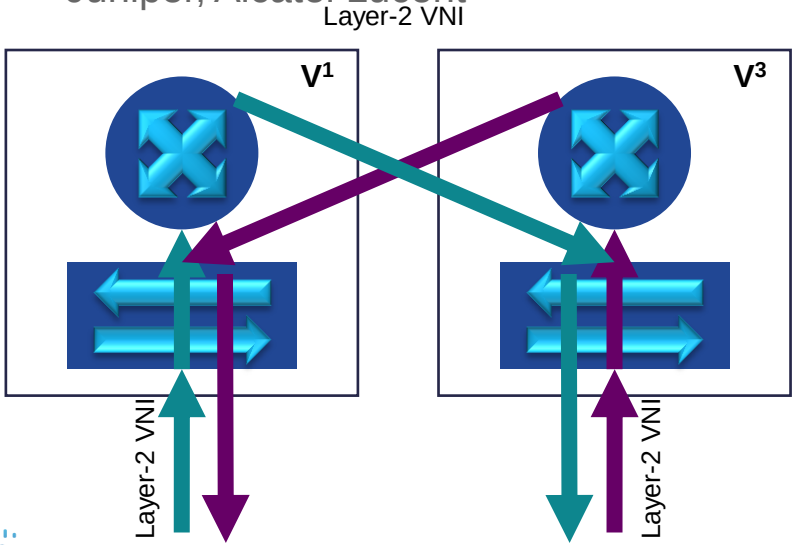


VXLAN L3 Gateway があるとき…

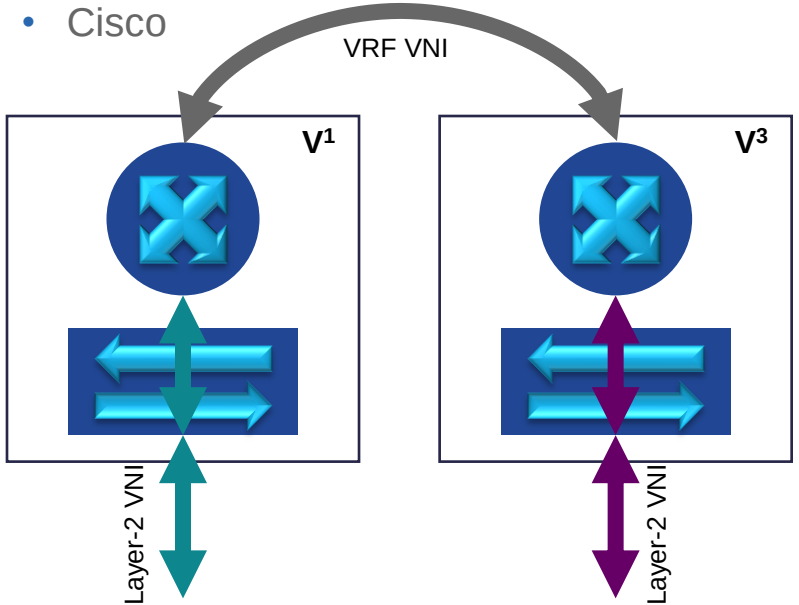


Asymmetric vs Symmetric IRBの違い (Integrated Route/Bridge)

- Asymmetric IRB
  - ルーティング, ブリッジング は Ingress VTEP で実施
  - Egress VTEP では ブリッジング のみ
  - Juniper, Alcatel Lucent



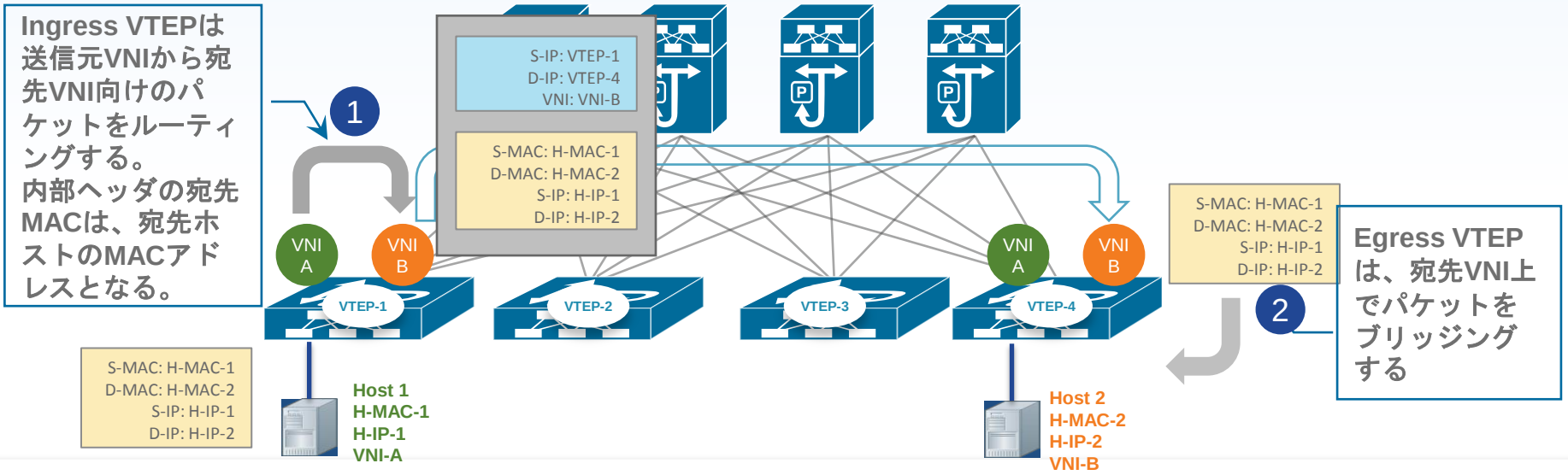
- Symmetric IRB
  - Ingress, Egress VTEPの両方で ルーティング とブリッジングを実施
  - Cisco



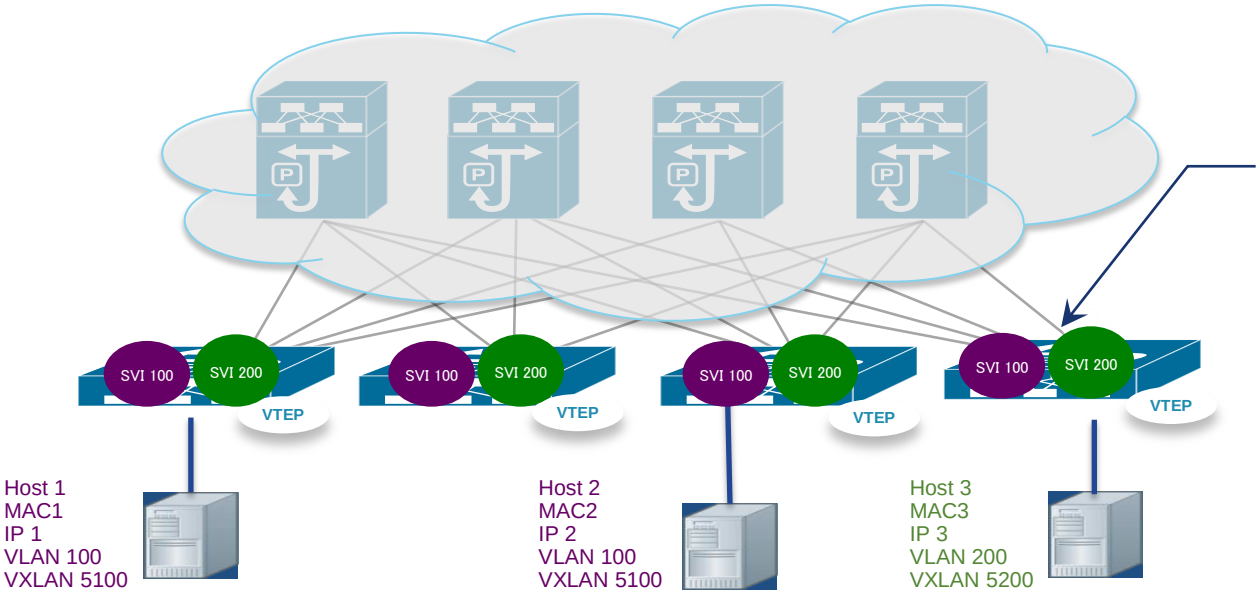
Asymmetric IRB

Asymmetric

- **Ingress VTEPではルーティングとブリッジングを実施**
- **Egress VTEPではブリッジングのみを実施**
- **Source VNIと Destination VNIの両方が、Ingress VTEP上に必要になるので拡張性に乏しい**



Asymmetric IRBの場合 VTEP VNIメンバーシップ



全てのVTEP に全てのVNIの設定が必要になる

全てのVTEPは、全てのVNIに対するMACテーブルの維持管理が必要になる。

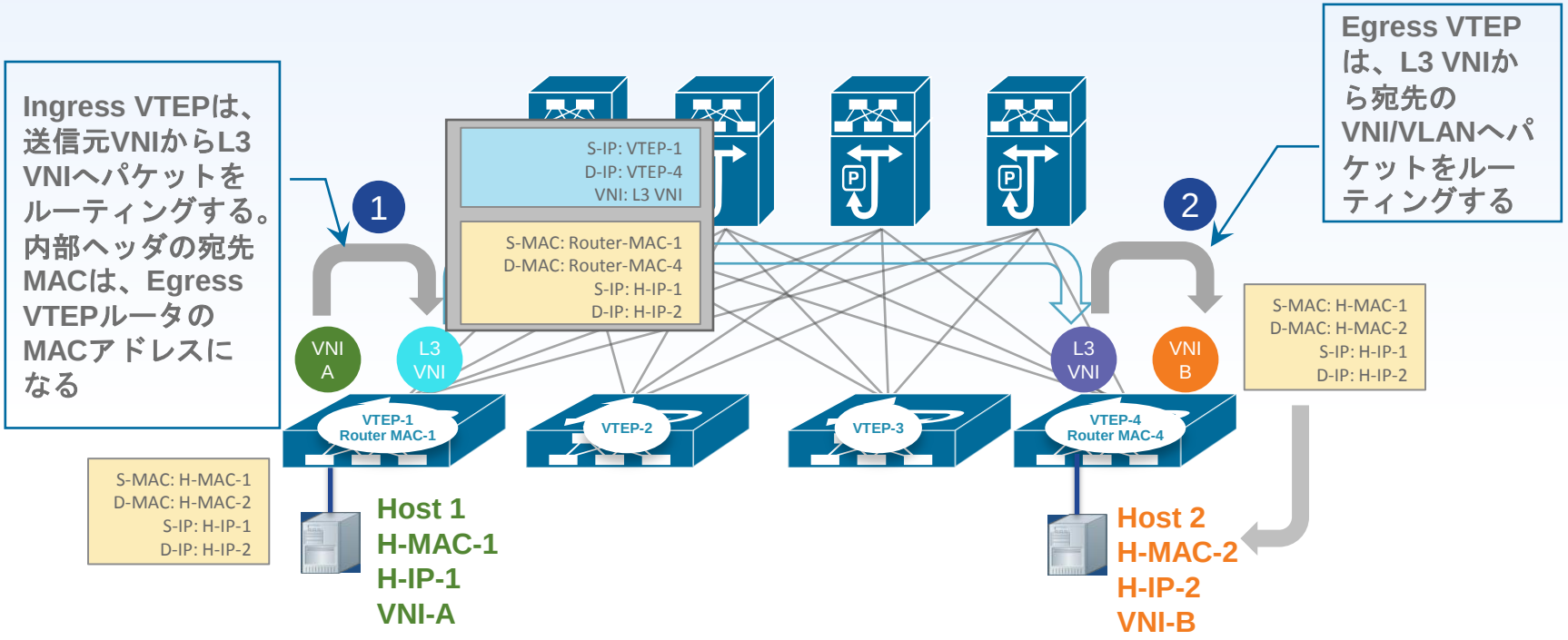
つまりローカルホストと関係ないVNIの設定も必要となる。

1. 1つのVNIに属する全VTEPは、ローカルホストのために仮想IPゲートウェイになることができる
2. 縦方向のルーティングトラフィックの転送を最適化

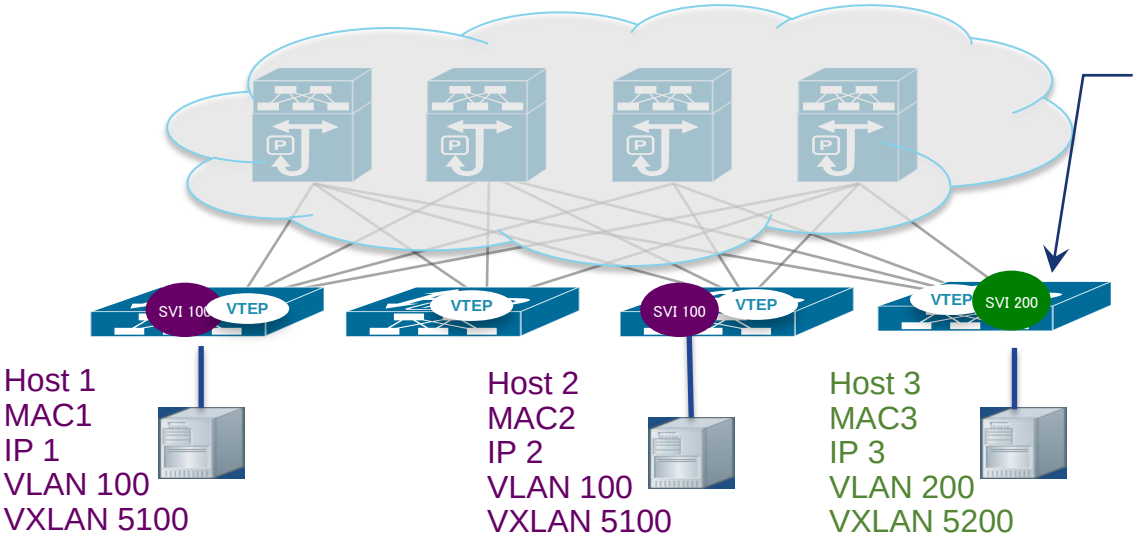
Symmetric IRB

- Ingress VTEPと Egress VTEPの両方でルーティングが可能
- Layer-3 VNI
  - テナントVPNの識別子
  - テナントVRF毎に一つ
- VTEP Router MAC
- Ingress VTEP は、Layer-3 VNIへパケットをルーティング
- Egress VTEP は、宛先のLayer-2 VNIへパケットルーティング

Symmetric IRB



Symmetric IRBの場合 VTEP VNIメンバーシップ



全VTEPは、ローカルホストを持つVNIを定義するだけで良い。

VTEPは、ローカルホストが存在しないVNIに対するMACテーブルの維持管理をする必要がない。

1. ARP および MAC テーブルの利用の最適化
2. VTEPは、ローカルホストが存在するVNIのみの定義が必要になるだけです。

Symmetric IRB vs Asymmetric IRBのまとめ

- Symmetric IRB は、VTEP上でARPとMACテーブルの効率的な使用を提供
- Symmetric IRB は、エンドホストに対して拡張性がある
- Symmetric IRB は、VXLANネットワークのVNIのトータル数でスケールし易い

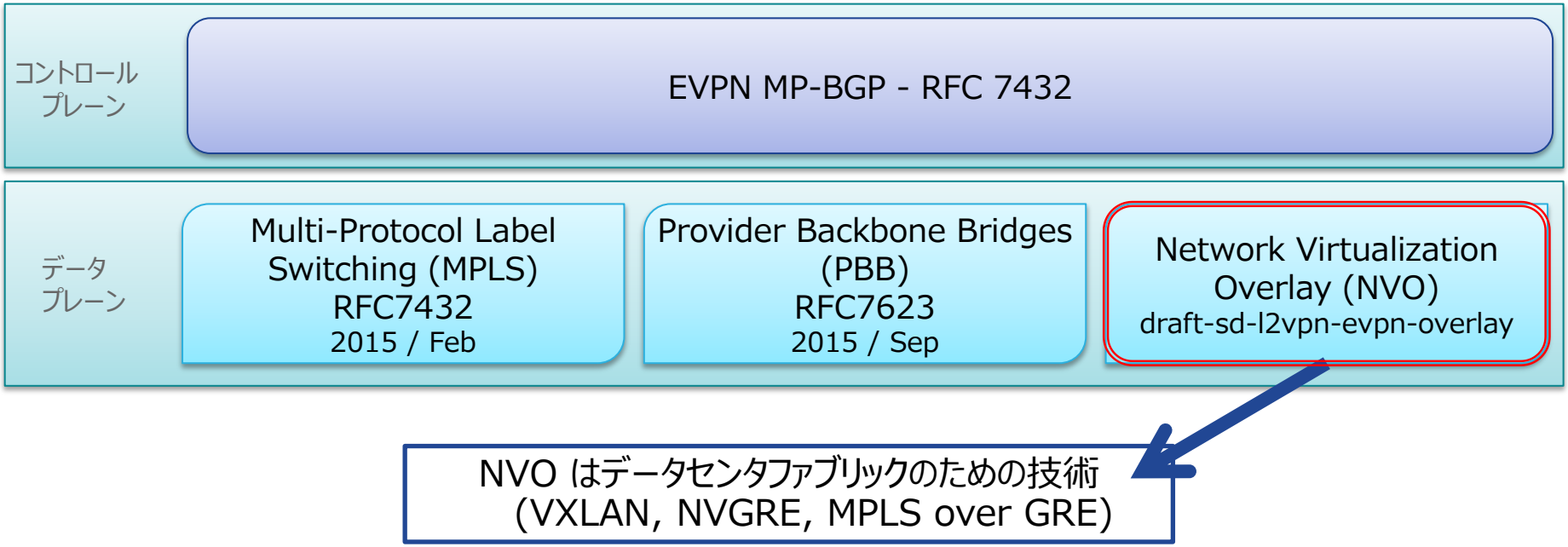
マルチベンダーでの相互接続:

- いくついかベンダーはAsymmetric IRBを実装
- 複数のベンダー間にて Symmetric IRBが最も最適な実装であると合意済み
- Ciscoは、すでにSymmetric IRBでNXOSやIOS-XRに実装済み

標準化プロトコル
VXLAN EVPN

EVPN – Ethernet VPN

EVPNは、次世代のL2VPNソリューションとして、標準化が進められている技術。
L2(MAC)やL3(IP)の情報をEVPN MP-BGP(コントロールプレーン)経由でアドバタイズ

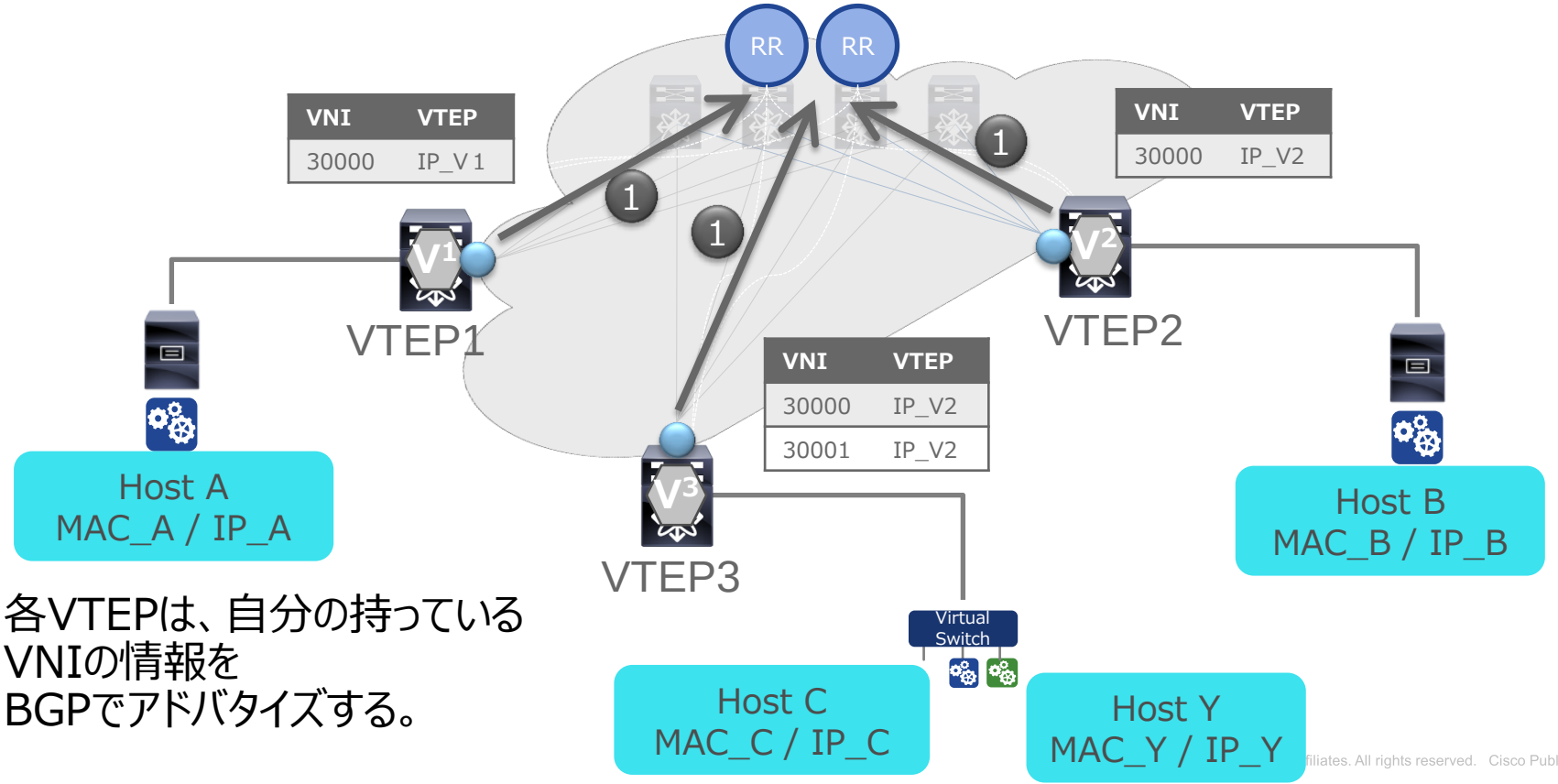


EVPNによるVXLANの改善点

1. VTEPの情報をMP-BGP経由で学ぶことができる
2. ホストの情報(L2/L3)情報をMP-BGP経由で学ぶことができる
3. VMが移動したことによるヘアピントラフィックを防ぐことができる。
4. MP-BGPを使ったマルチテナントのルート制御を実現する(Route-target)
5. Multicastを排除したアンダーレイを構築可能(Ingress Replication)
6. ARPを抑制する機能(Suppress-arp)
7. VTEPの認証を可能にする(MP-BGPのネイバー認証)

VTEP ピアの学習方法

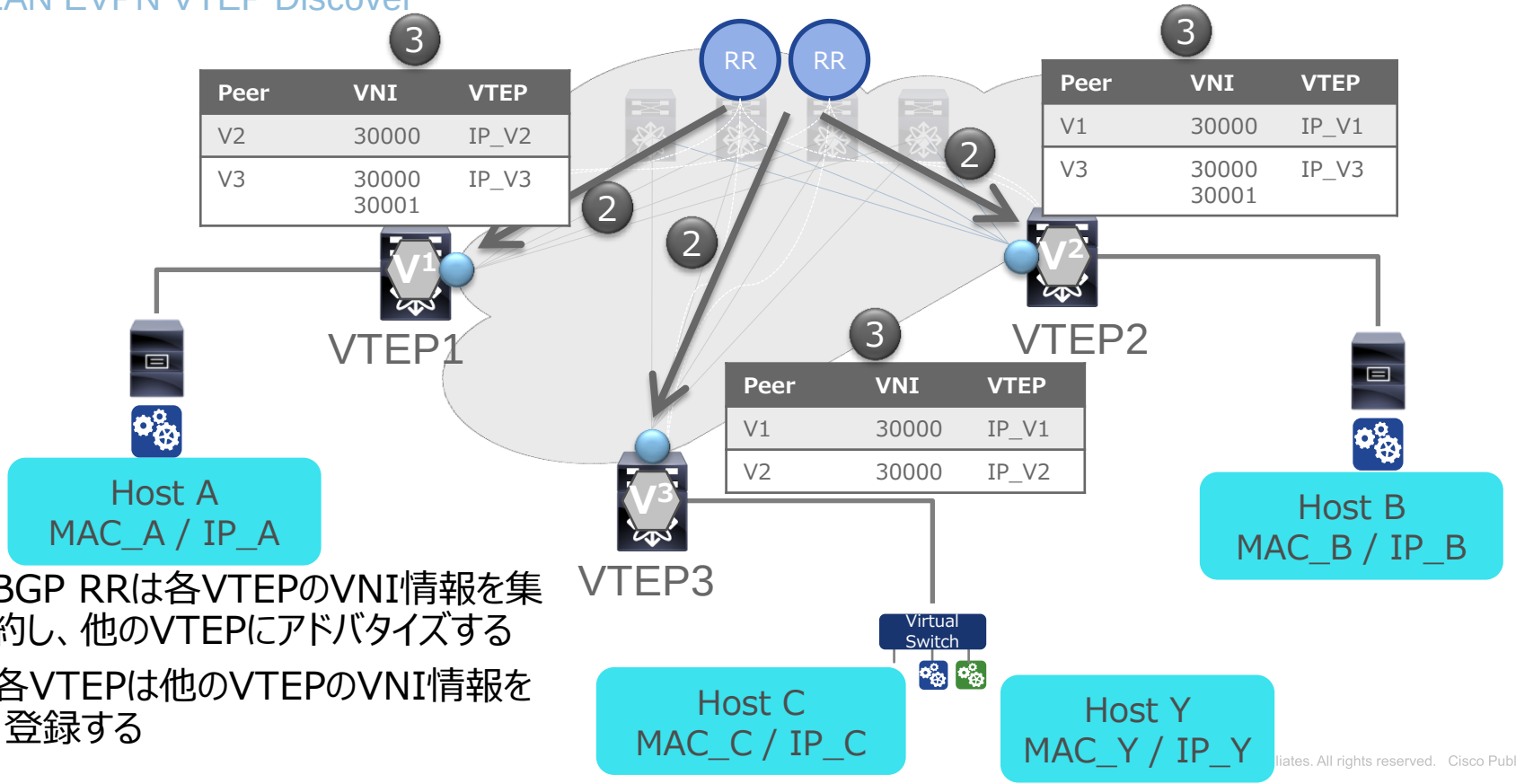
VXLAN EVPN VTEP Discover



1 各VTEPは、自分の持っている VNIの情報を BGPでアドバタイズする。

VTEP ピアの学習方法(つづき)

VXLAN EVPN VTEP Discover

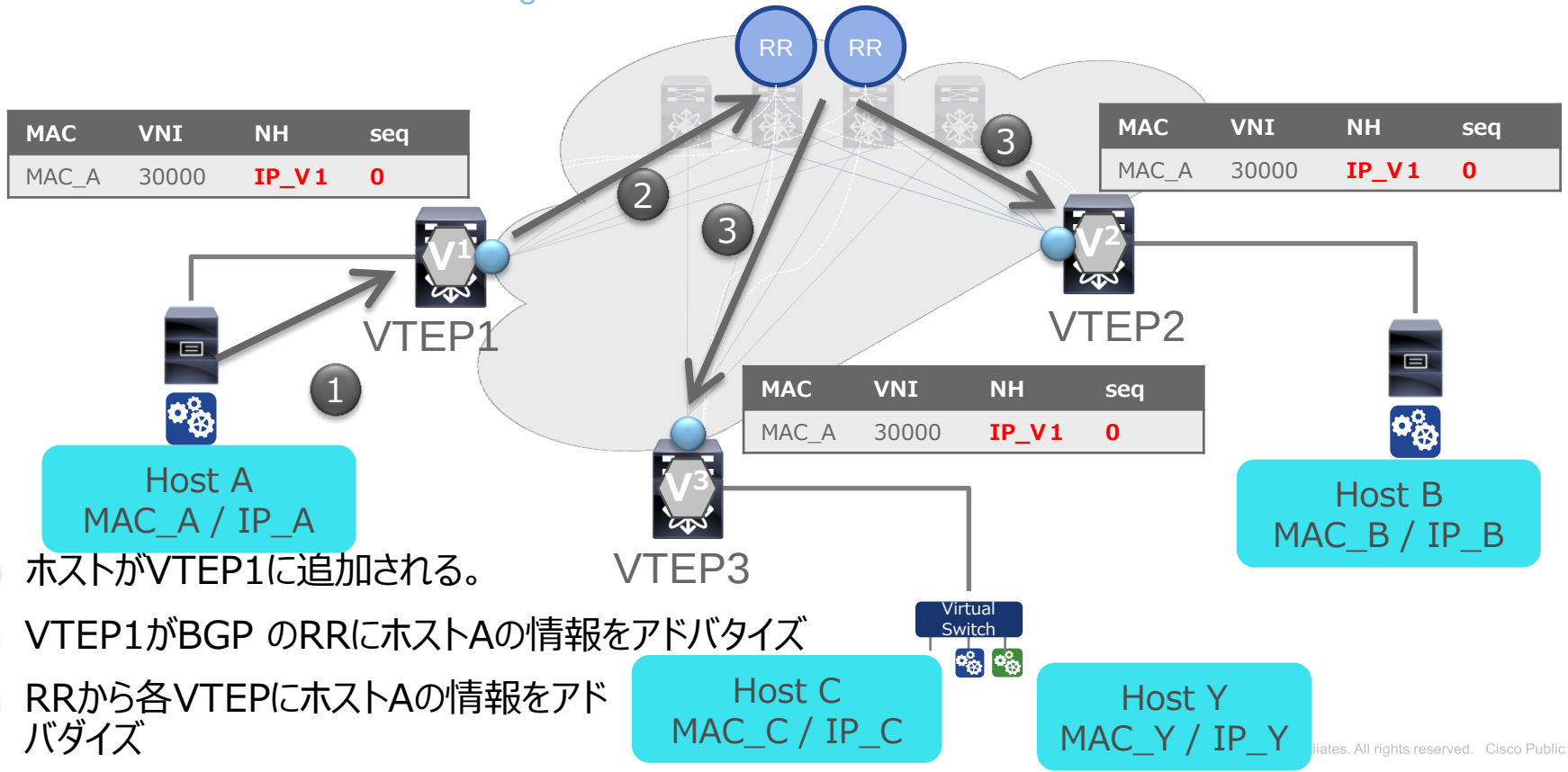


2 BGP RRは各VTEPのVNI情報を集約し、他のVTEPにアドバタイズする

3 各VTEPは他のVTEPのVNI情報を登録する

ホストアドレスの学習方法

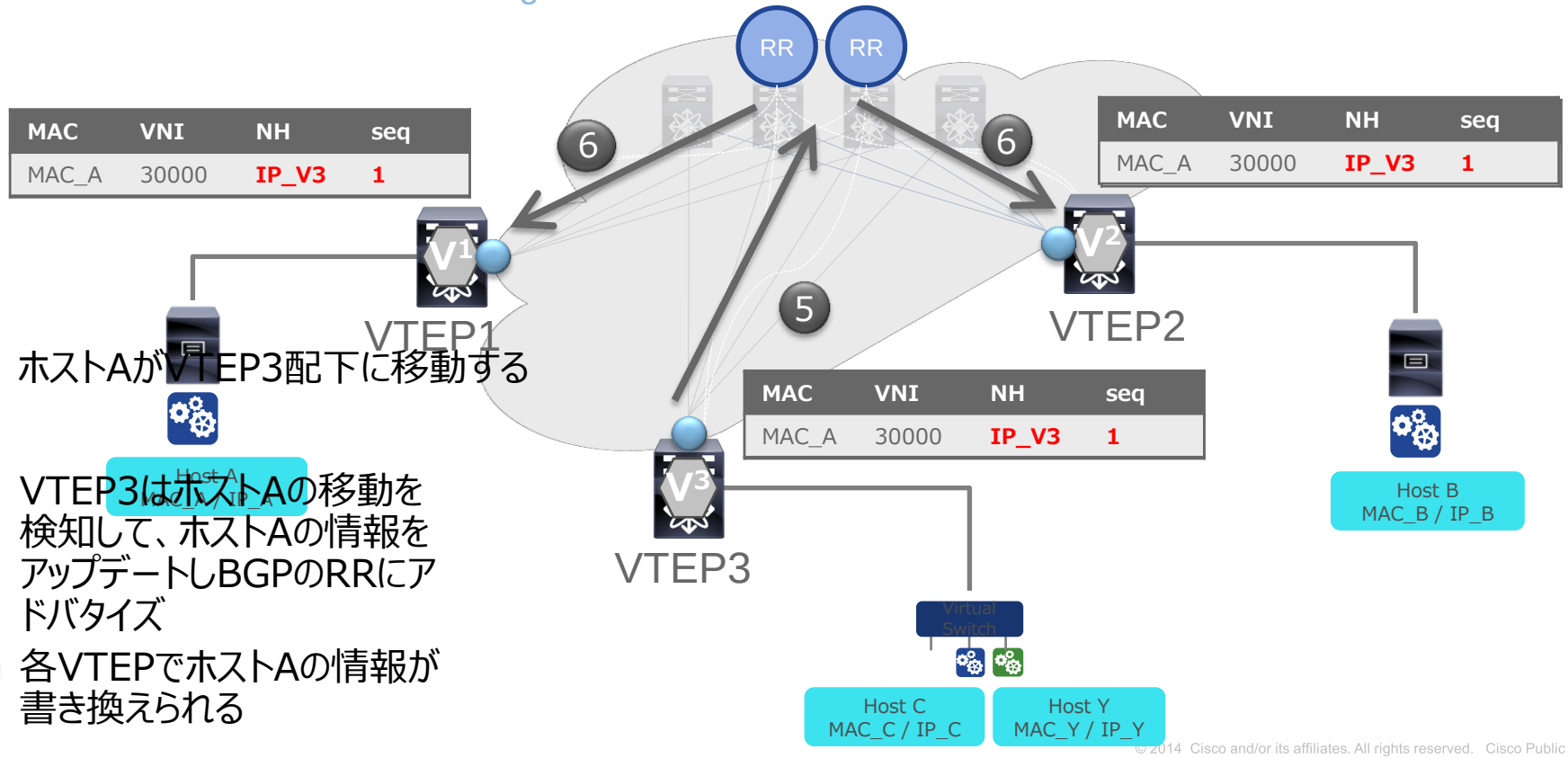
VXLAN EVPN HOST Route leaning



1. ホストがVTEP1に追加される。
2. VTEP1がBGP のRRにホストAの情報をアドバタイズ
3. RRから各VTEPにホストAの情報をアドバタイズ

ホストアドレスの学習方法(つづき)

VXLAN EVPN HOST Route leaning



4 ホストAがVTEP3配下に移動する

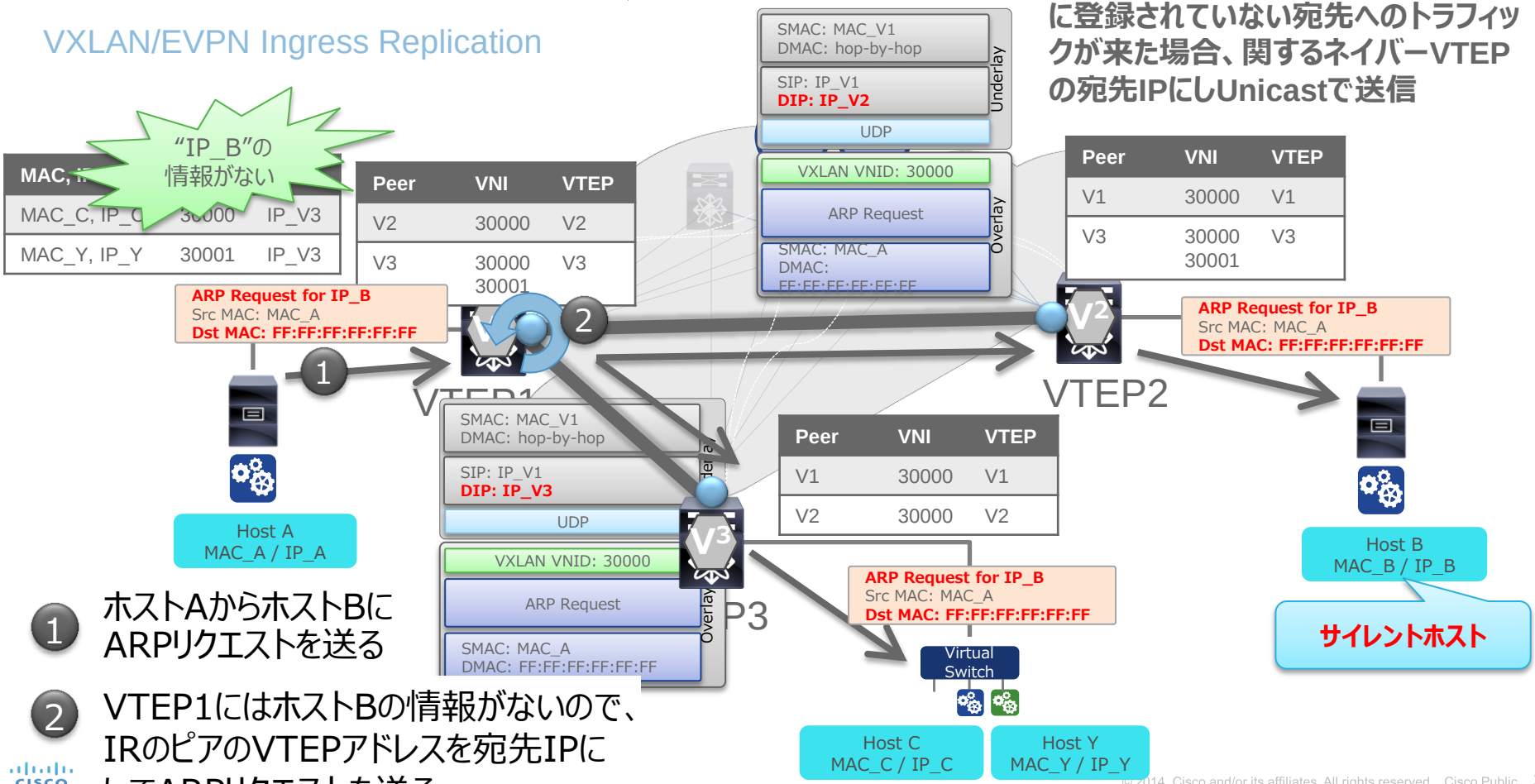
5 VTEP3はホストAの移動を検知して、ホストAの情報をアップデートしBGPのRRにアドバタイズ

6 各VTEPでホストAの情報が書き換えられる

BUMトラフィックの扱い方

VXLAN/EVPN Ingress Replication

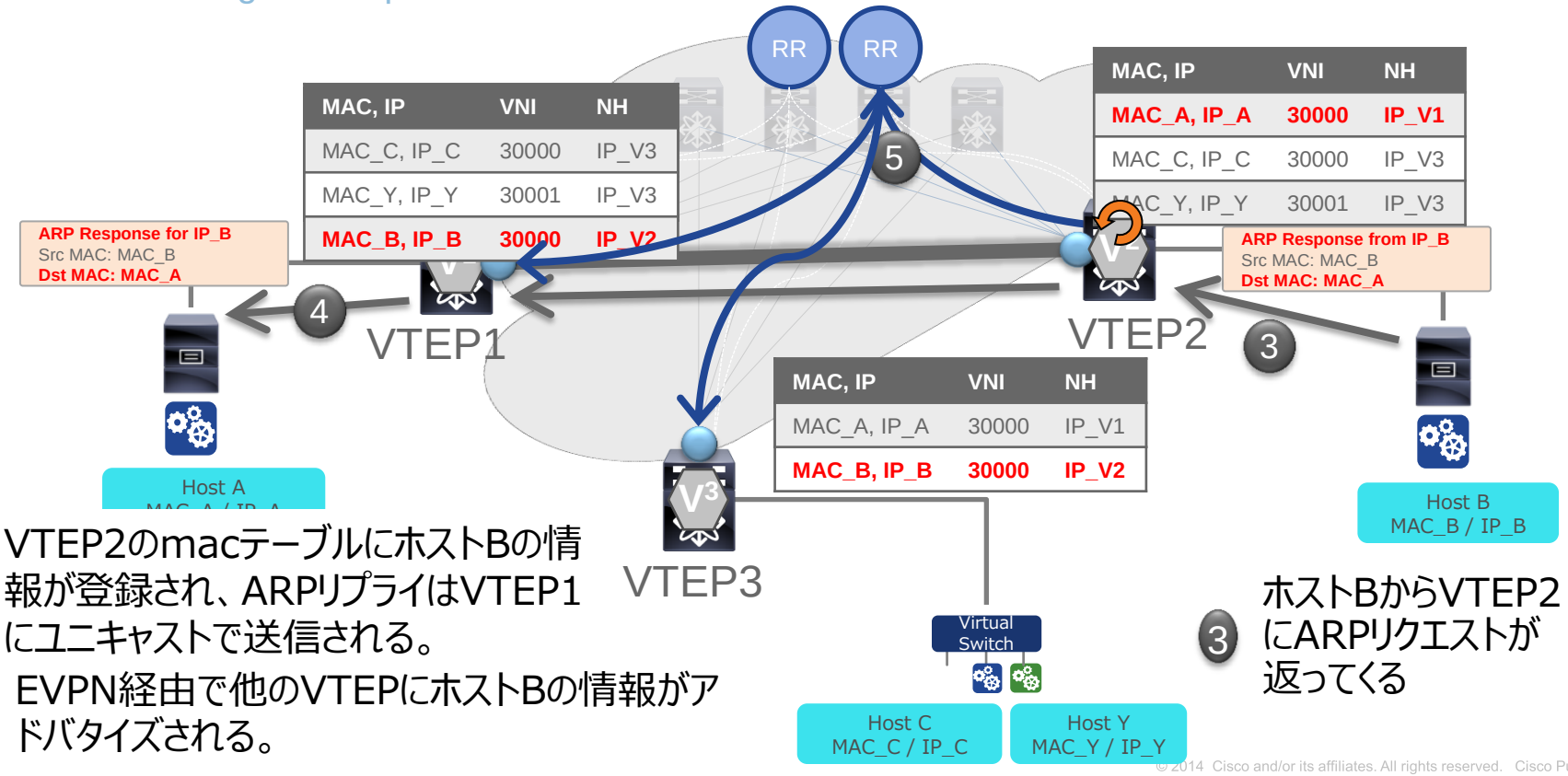
ホスト側のインターフェースからVTEPに登録されていない宛先へのトラフィックが来た場合、関するネイバーVTEPの宛先IPにしUnicastで送信



1. ホストAからホストBにARPリクエストを送る
2. VTEP1にはホストBの情報がないので、IRのピアのVTEPアドレスを宛先IPにしてARPリクエストを送る。

BUMトラフィックの扱い方(つづき)

VXLAN/EVPN Ingress Replication



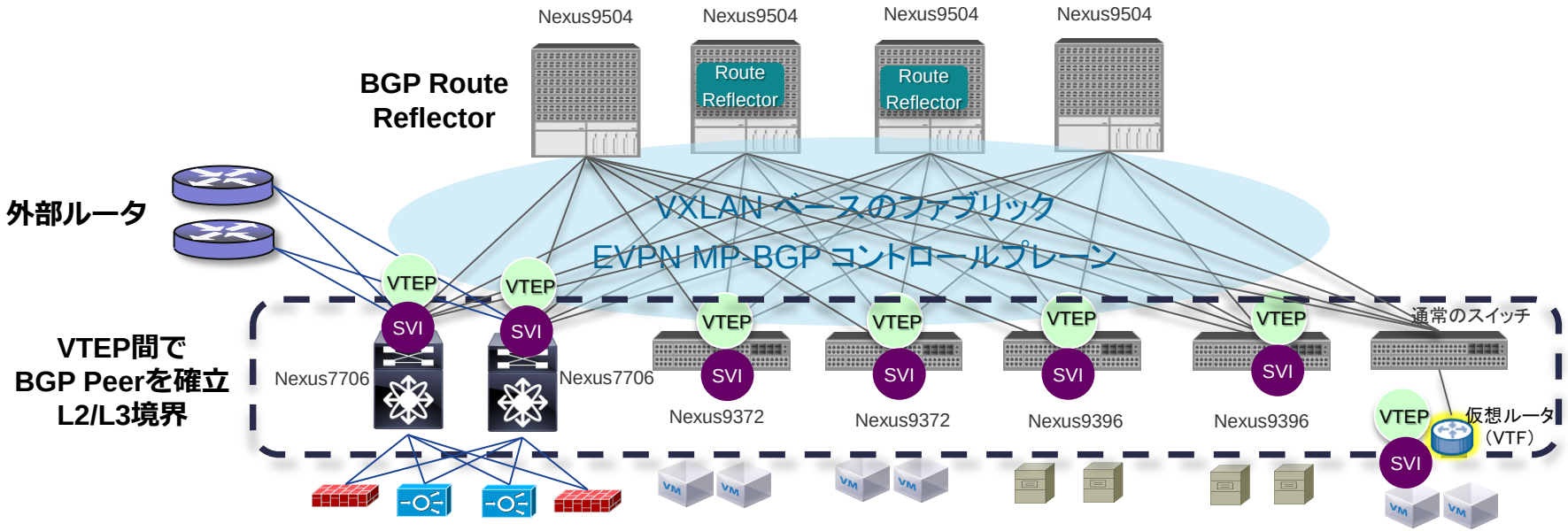
4 VTEP2のmacテーブルにホストBの情報が登録され、ARPリプライはVTEP1にユニキャストで送信される。

5 EVPN経由で他のVTEPにホストBの情報がアドバタイズされる。

3 ホストBからVTEP2にARPリクエストが返ってくる

VXLAN EVPNの活用例

VXLAN EVPNでDCファブリックの構成

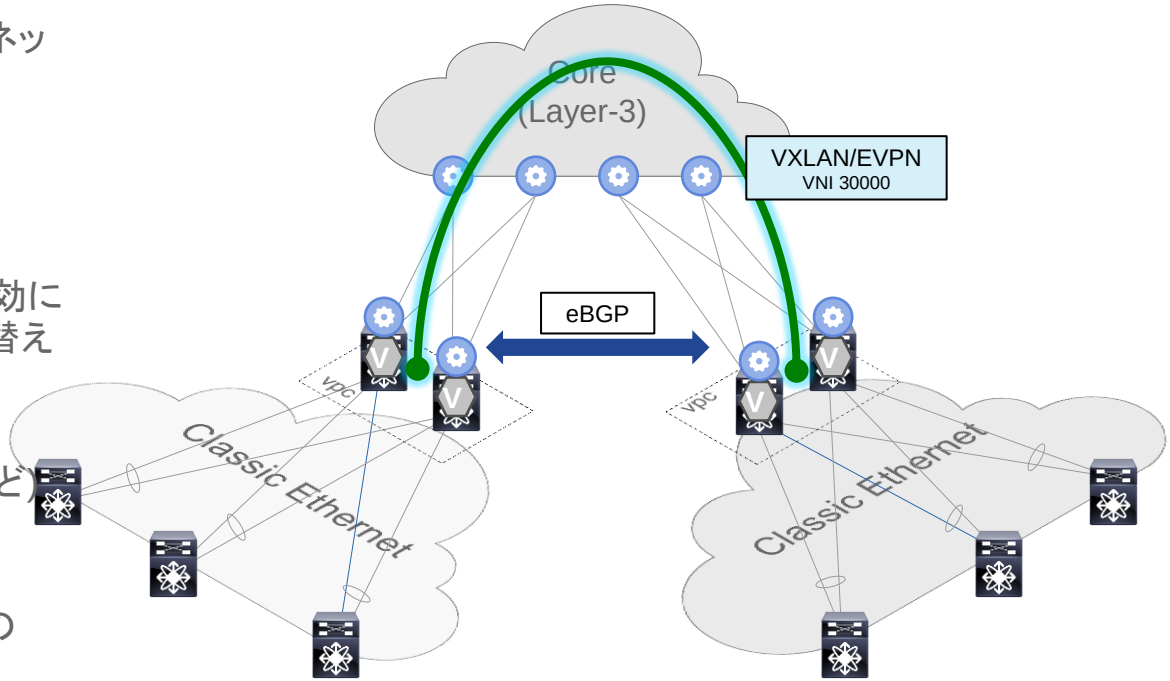


- オープンなプロトコルVXLAN with EVPNによりハイブリッドオーバレイを提供
- BGPによってネイバーVTEPやホストの発見を動的に実施
- SDNコントローラによりオーバレイネットワークの自動化やIP Mobilityを実現

クラシックネットワークの相互接続 (Layer-2)

Pod間の接続

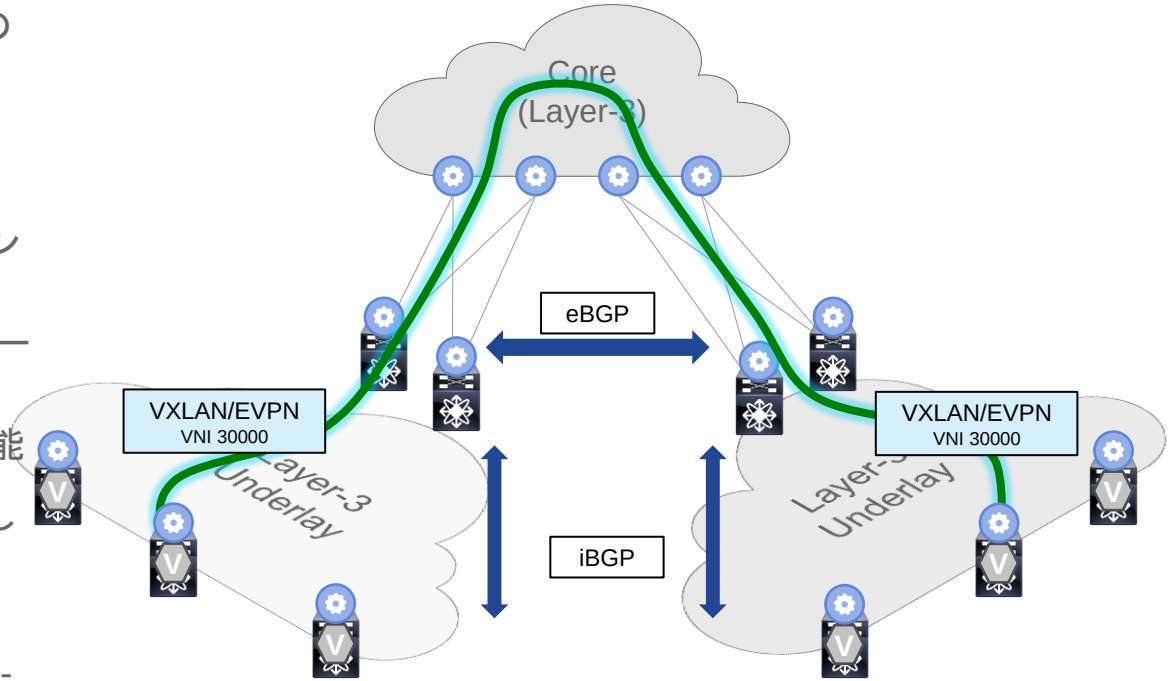
- VXLAN/EVPNを持つクラシックイーサネットPodの相互接続では、
  - EVPN Control-Planeは、ARP suppression機能を提供することでFlood & Learnを防止する
- Coreのアンダーレイにて Multicastを有効にせずともIngress Replication機能で代替え可能
  - 様々なネットワーク要件に依存 (トラフィックの特性やサイトの総数など)
- マルチホーミングにはvPC機能を活用
- 片方のクラシックイーサネットPod上でのループは、他のPodに影響する



VXLANネットワークの相互接続 (Layer-2)

Pod間の接続

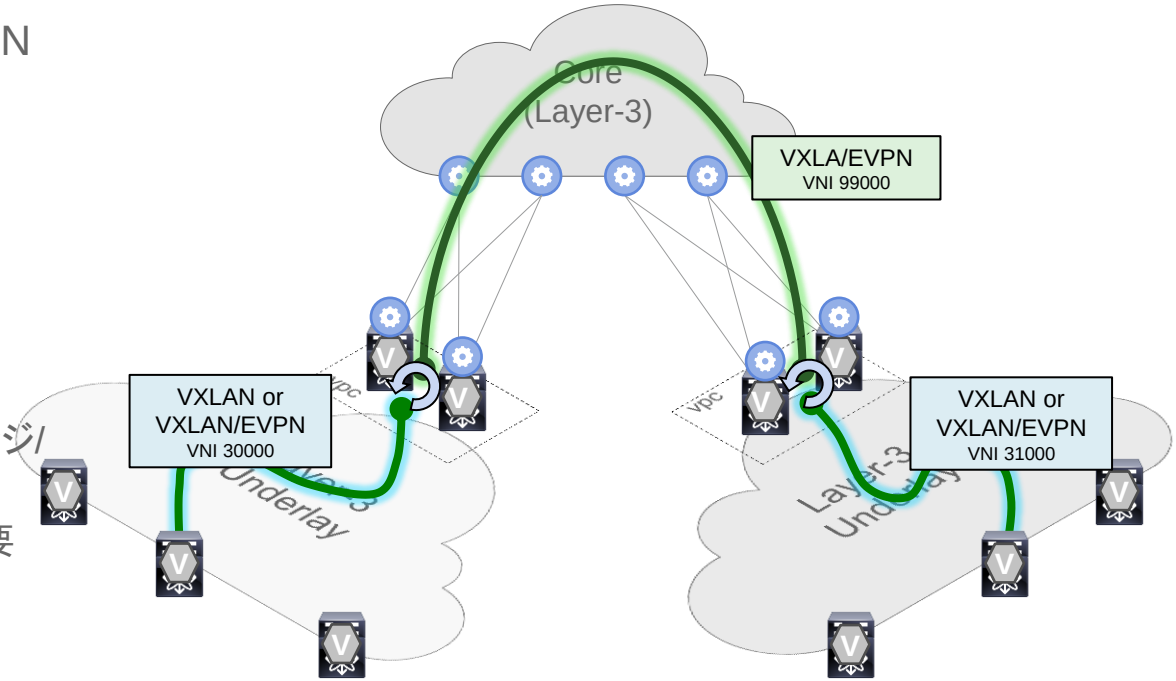
- VXLAN/EVPNを持つVXLAN/EVPN Podの相互接続では、
  - Control-Planeドメイン (EVPN) は、iBGP/eBGPによって分離される
- Data-PlaneのVTEPによるカプセル化はエンド-to-エンドで提供
  - Leaf/ToRスイッチは、2つのデータセンターを跨り、全てのVTEP情報を知っている
  - BUMトラフィックは、Pod跨ぎでも到達可能
- Pod間を接続する際、BUMトラフィックに対しIngress Replication (Unicast) またはMulticastを選択可能
- ルート交換を確実にするのに共通の Route-Targetを使用する



VXLANネットワークの相互接続 (Layer-3)

Pod間の接続

- VXLAN/EVPNを持つVXLAN/EVPN Podの相互接続では、
  - Control-Planeドメイン (EVPN)は、iBGP/eBGPによって分離される
- Layer-3相互接続を用いて, Data-Plane カプセル化は分離される
  - VXLANルーティングの処理はDCエッジ/アグリゲーションスイッチで実施
  - サイト間でTransit VNI(L3 VNI)が必要
- これはLayer-2相互接続ではない



VXLAN EVPNのコンフィグ例

VTEP (VXLAN Tunnel End-Point)を作る

Overlay

Configuration Example

```
# Features & Globals
feature bgp
feature nv overlay
nv overlay evpn

# Spine (S1)

# Leaf (V1)
interface nve1
  source-interface loopback0
  host-reachability protocol bgp
```

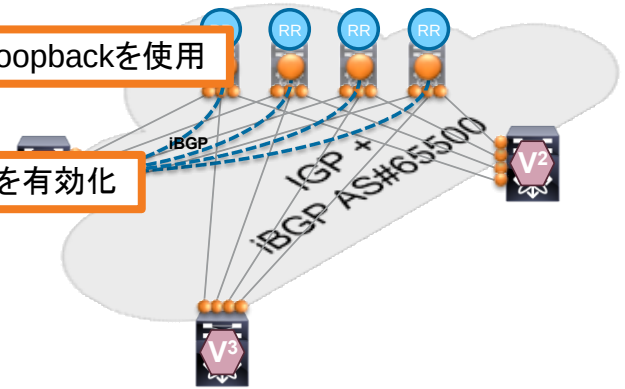
VTEPの有効か (Leaf やBorderスイッチ上で必要)

EVPN Control-Plane in BGPの有効化

VTEPインターフェースの設定

Source InterfaceとしてLoopbackを使用

Host reachabilityのためにBGPを有効化



*Simplified BGP configuration; would have 4 BGP peers (RR)
IGP not shown

Overlayコントロールプレーンを作る

Overlay

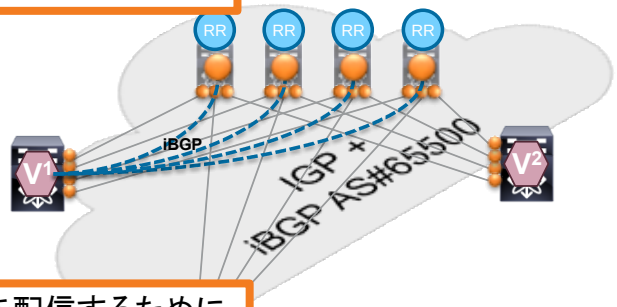
Configuration Example

```
# Features & Globals
feature bgp
feature nv overlay
nv overlay evpn

# Spine (S1)
router bgp 65500
  router-id 10.10.10.S1
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 10.10.10.0/24 remote-as 65500
    update-source loopback0
    address-family l2vpn evpn
      send-community both
       route-reflector-client

# Leaf (V1)
router bgp 65500
  router-id 10.10.10.V1
  address-family ipv4 unicast
  neighbor 10.10.10.S1 remote-as 65500
    update-source loopback0
    address-family l2vpn evpn
      send-community both
```

EVPN Control-Plane in BGPの有効化

各BGPネイバー配下でL2VPN EVPNで有効化

EVPN route attributesを配信するために
Extended BGPコミュニティを送信

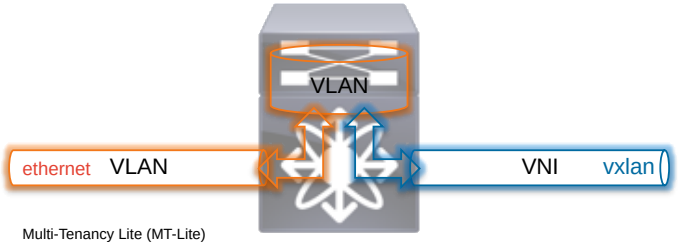


*Simplified BGP configuration; would have 4 BGP peers (RR)
IGP not shown

VLAN to VXLANのマッピングを作る

Overlay

- **MT-Lite コンフィグレーションは、IEEE 802.1Q VLAN ID からVXLAN VNIにマッピングするシンプルな方法です**
  - **VLAN とVNIのマッピング設定は、スイッチ毎に行う**
  - **VLANは、単なる各スイッチ上でのローカル識別子になる**
  - **VNI は、ファブリックワイドの識別子になる**
  - **4k VLANの制限は、各スイッチ単位での制限となる**
  - **4k ネットワークセグメントの制限は解除される**



Multi-Tenancy Lite (MT-Lite)

Configuration Example MT-Lite (N9300/N5600)

```
# Features
feature vn-segment-vlan-based

# VLAN to VNI mapping (MT-Lite)
vlan 43
 vn-segment 30000

# Activate Layer-2 VNI for EVPN
evpn
  vni 30000 l2
    rd auto
    route-target import auto
    route-target export auto

# Activate Layer-2 VNI on VTEP
interface nve1
  source-interface loopback0
  host-reachability protocol bgp
  member vni 30000
    mcast-group 239.239.239.100
    suppress-arp
```

VLAN とLayer-2 VNI マッピング

Layer2サービス向けに EVPN Control-Planeを有効化

ユニキャストモード時は “ingress-replication protocol bgp”を使う

VTEP上でLayer2 VNIを有効化し、オプションで ARP代理応答を有効化

VXLAN 参考情報

Cisco VXLAN EVPN サポートOS情報

ASR9000： IOS-XR 5.2.0以降

Nexus9300/9500： NX-OS 7.0(3)I2(1)以降

Nexus7000/7700： NX-OS 7.2(0)D1(1)以降
（F3 モジュールが必要）

Nexus5600： NX-OS 7.3（2016年Q1リリース予定）

Cisco VXLAN Multicast サポートOS情報

ASR9000： IOS-XR 5.2.0以降

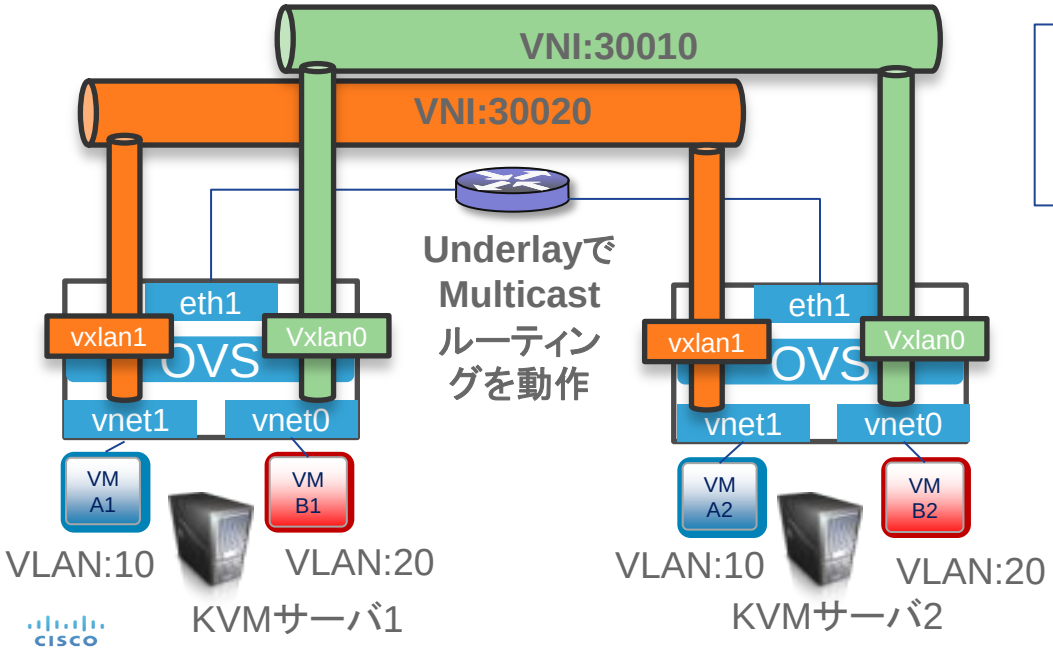
ASR1000： IOS-XE 3.13S以降

Nexus9300/9500： NX-OS 7.0(3)I2(1)以降

Nexus7000/7700： NX-OS 7.2(0)D1(1)以降
（F3 モジュールが必要）
Nexus5600： NX-OS 7.1(0)N1(1a) 以降
Nexus3100：NX-OS 7.0(3)I2(1)以降

Linux 環境におけるVXLANの実装

Linux Kernel 3.7から、VXLAN Multicast モードをサポート
Red Hat はiproute2, Ubuntuはiprouteのパッケージをインストールする




```
<vxlanデバイスを作成例>
ip link add vxlan0 type vxlan id 30010  group 239.1.1.1 dev eth1
ip link add vxlan1 type vxlan id 30020  group 239.1.1.1 dev eth1
ip link set up vxlan0
ip link set up vxlan1
```

※シスコとの接続するためには、VXLAN UDP ポート番号がLinux kernel は、8472のため、IANA標準規格の4789へ変更する

参考情報

- RFC 7432 - BGP MPLS-based Ethernet VPN:
https://tools.ietf.org/html/draft-ietf-l2vpn-evpn-11
- IETF Draft - Network virtualization overlay solution with EVPN:
https://tools.ietf.org/html/draft-ietf-bess-evpn-overlay-00
- IETF Draft - Integrated routing and bridging in EVPN:
https://tools.ietf.org/html/draft-ietf-bess-evpn-inter-subnet-forwarding-00
- IETF Draft - IP prefix advertisement in EVPN:
https://tools.ietf.org/html/draft-rabadan-l2vpn-evpn-prefix-advertisement-02
- RFC 4271 - Border Gateway Protocol 4 (BGP-4):
https://tools.ietf.org/html/rfc4271
- RFC 4760 - Multiprotocol extensions for BGP-4:
https://tools.ietf.org/html/rfc4760
- RFC 4364 - BGP/MPLS IP VPNs:
https://tools.ietf.org/html/rfc4364

TOMORROW starts here.