Internet Week 2012

オーバーレイネットワーク
現状と今後への期待

2012-11-20

ヴイエムウェア株式会社
小松康二

※本資料の内容は発表者の個人的な見解に基づくものであり、
ヴイエムウェア株式会社の見解を示すものではありません。

vmware

アジェンダ

今なぜオーバーレイなのか？

オーバーレイの実装例 – VXLAN

オーバーレイ実装方式の違い

現在の課題と取り組み例 – VXLAN

2011～2012年に相次いでドラフト化

VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks

- http://tools.ietf.org/html/draft-mahalingam-dutt-dcops-vxlan-02
- バージョン 02 (2011年8月に最初のドラフトを公開)

NVGRE: Network Virtualization using Generic Routing Encapsulation

- http://tools.ietf.org/html/draft-sridharan-virtualization-nvgre-01
- バージョン 01 (2011年9月に最初のドラフトを公開)

STT: A Stateless Transport Tunneling Protocol for Network Virtualization

- http://tools.ietf.org/html/draft-davie-stt-02
- バージョン 02 (2012年2月に最初のドラフトを公開)

オーバーレイネットワークの一般的なメリット

論理的に構成可能なネットワーク

- 物理ネットワークの構成作業に依存しない
- オンデマンドかつフレキシブルに定義し利用可能

スケーラブルなレイヤ 2 ネットワーク

- 1000万をゆうに超える論理ネットワークを生成可能
- ネットワークインフラの共有と隔離性の両立
- MAC アドレス、IP アドレスの重複の許容

物理レイヤ 3 境界を越えたレイヤ 2 ネットワーク

- ゲストOSのネットワーク設定の変更なしで、仮想マシンの可搬性を最大化
- レイヤ3構成とカプセル化で、ToRスイッチのMAC アドレステーブル肥大化を回避
- レイヤ3構成により、L2 ネットワークにおける STP のリンク非効率性から解放

オーバーレイの特長 1/2

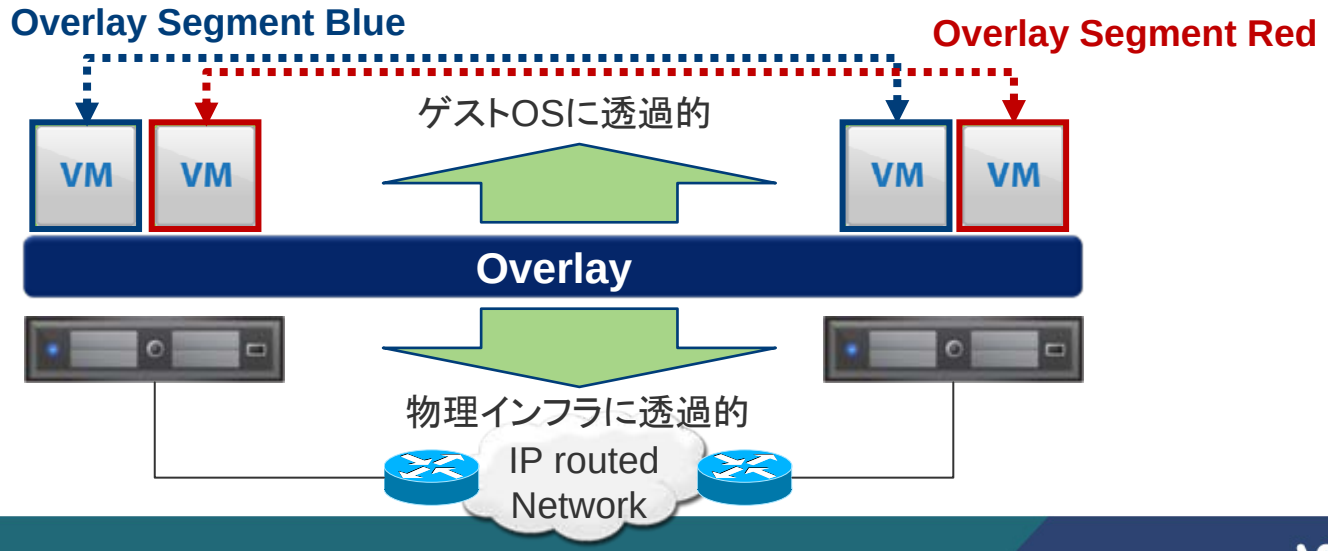
既存資産を生かし、従来の技術・ノウハウの延長で構築・運用できる

ゲストOS／アプリケーションの観点

- 従来のVLANに接続している場合の動作と違いがない (区別できない)
- MACアドレスやIPアドレスは、各VXLANセグメントの中で一意であればよい

物理ネットワークの観点

- 物理インフラストラクチャの構成変更は最低限に抑えられる
- 物理ネットワークではOverlayに接続しているノード(MAC)を意識する必要がない

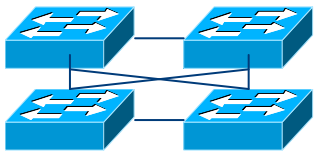


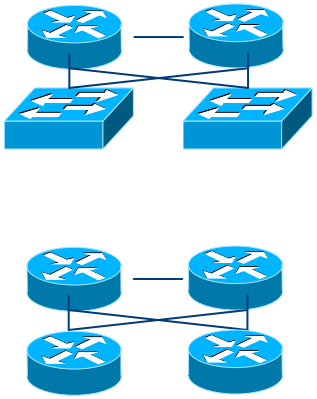
オーバーレイの特長 1/2 – 続き

802.1q の普及
Per-VLAN RSTP
Link Aggregation

L3 Switch の低価格化
L3実装のメリットの明確化
•等コスト経路
•信頼性
•収束時間

仮想化／クラウド化への対応の必要性
•仮想マシンワークロードの柔軟性の要求
•ToRスイッチのMACテーブル肥大化
•仮想マシントラフィックの可視性の低下





L2 Switch 全盛期

L3 Switch 全盛期

オーバーレイによって物理と論理を分離

オーバーレイ = L2の利便性とL3の堅牢性の両方のメリットを享受

オーバーレイの特長 2/2

ネットワークの管理者と利用者に異なるビューを提供

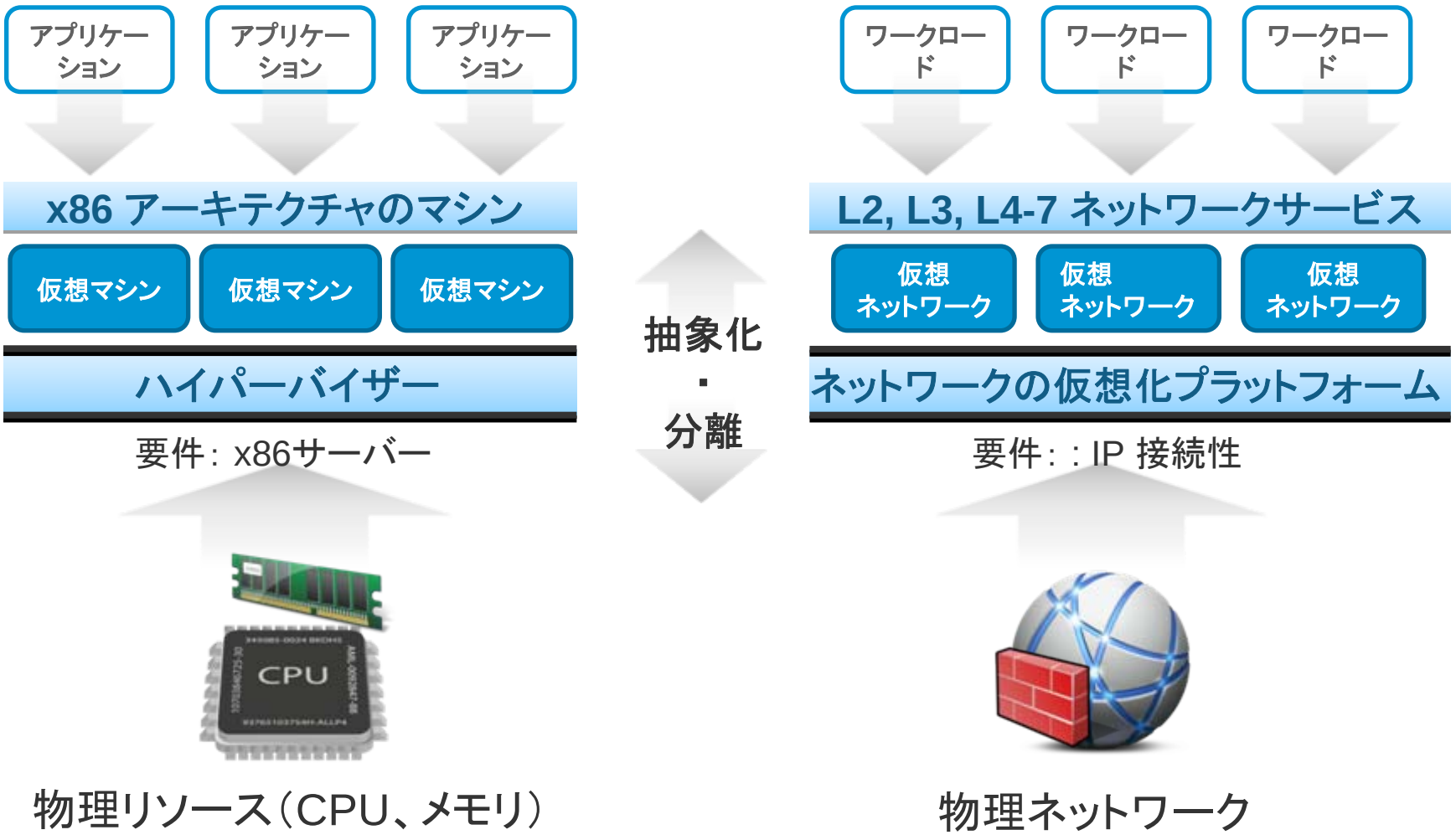
ネットワーク管理者

- VLAN ID の枯渇を心配する必要がない
- 仮想マシン数の増大に伴う、ToRスイッチのキャパシティ不足を回避
  (MAC学習上限や、ポートあたりの VLAN 設定上限を気にする必要がない)
- Live Migration などの仮想化テクノロジーに影響されない、シンプルで堅牢な L3 ネットワーク設計が可能

サーバ管理者

- ネットワークチームの手をわずらわせること無く、必要な時に必要なだけ即座に論理ネットワークを払い出すことができる
- 物理ネットワークの構成を気にすること無く、仮想マシンを自由に Live Migration できる

オーバーレイによるネットワークの仮想化とは？



最近のネットワーク技術との関係 (私見)

OpenFlow との関係

Ethernet Fabric (Trill/SPB等々) との関係

EVB/VEPAやSR-IOVとの関係

アジェンダ

今なぜオーバーレイなのか？

オーバーレイの実装例 – VXLAN

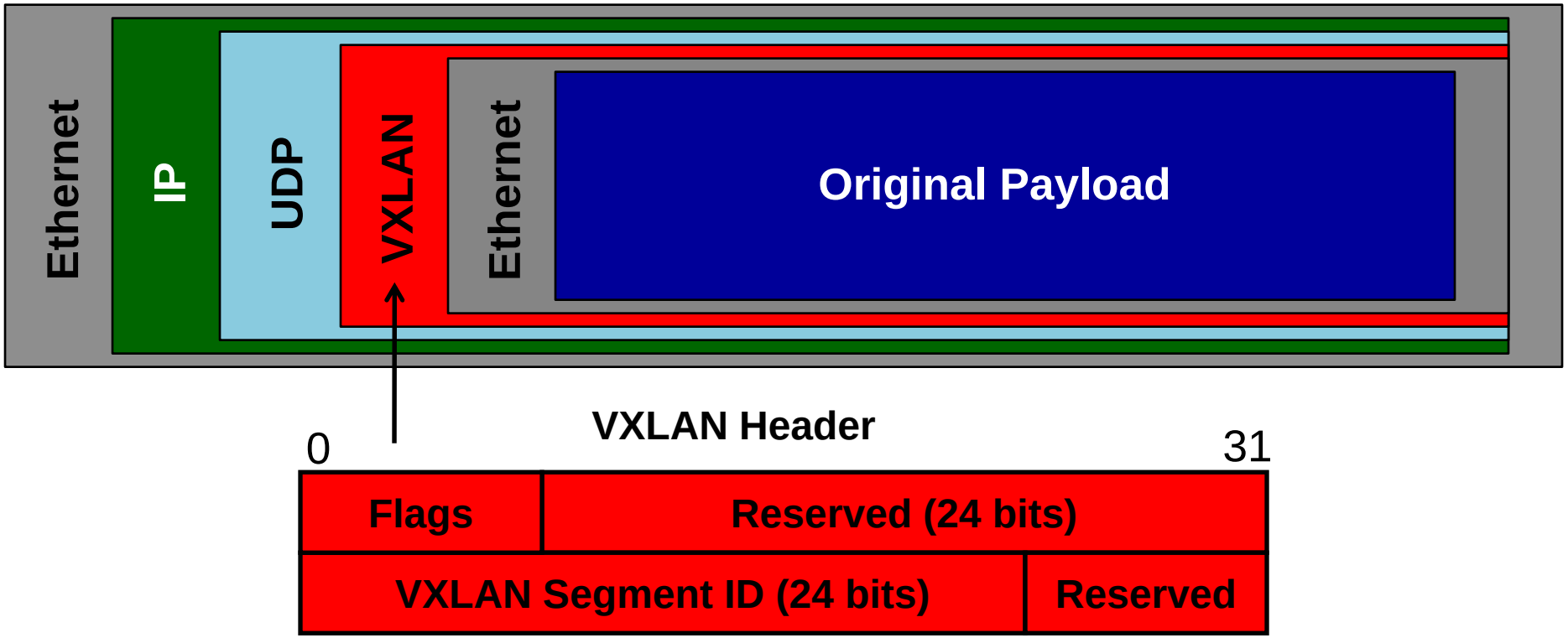
オーバーレイ実装方式の違い

現在の課題と取り組み例 – VXLAN

なぜスケーラブルか？

オリジナルのEthernetフレームをVXLANヘッダでカプセル化

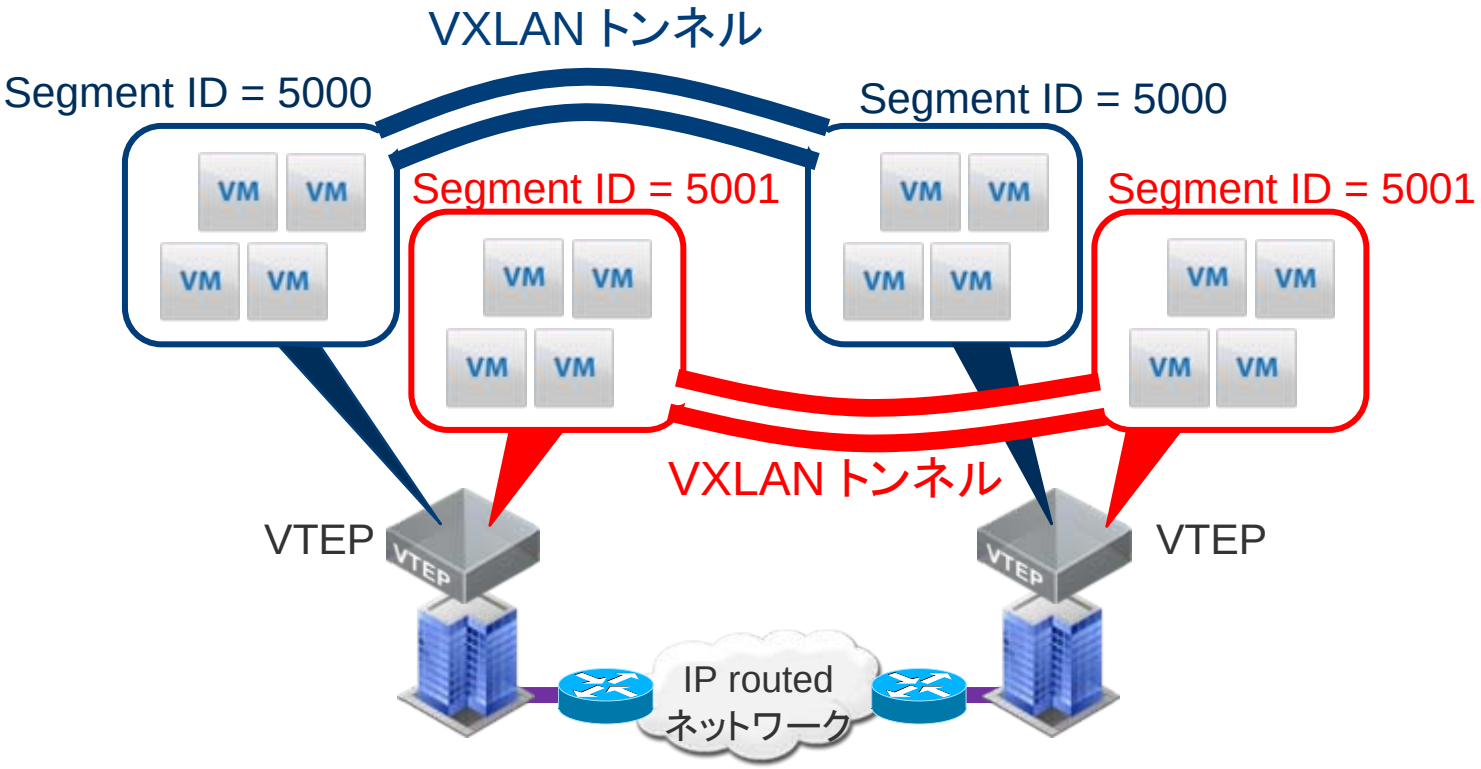
- VLAN ID (12ビット) に対して、VXLAN Segment ID は24ビットと範囲が広い
- 物理ネットワークはOuterヘッダのMACアドレス、IPアドレスのみを参照するため、学習するMACエントリ数が削減される



なぜレイヤ3を超えられるのか？

VTEP (VXLAN Tunnel Endpoint) 間でトンネリング

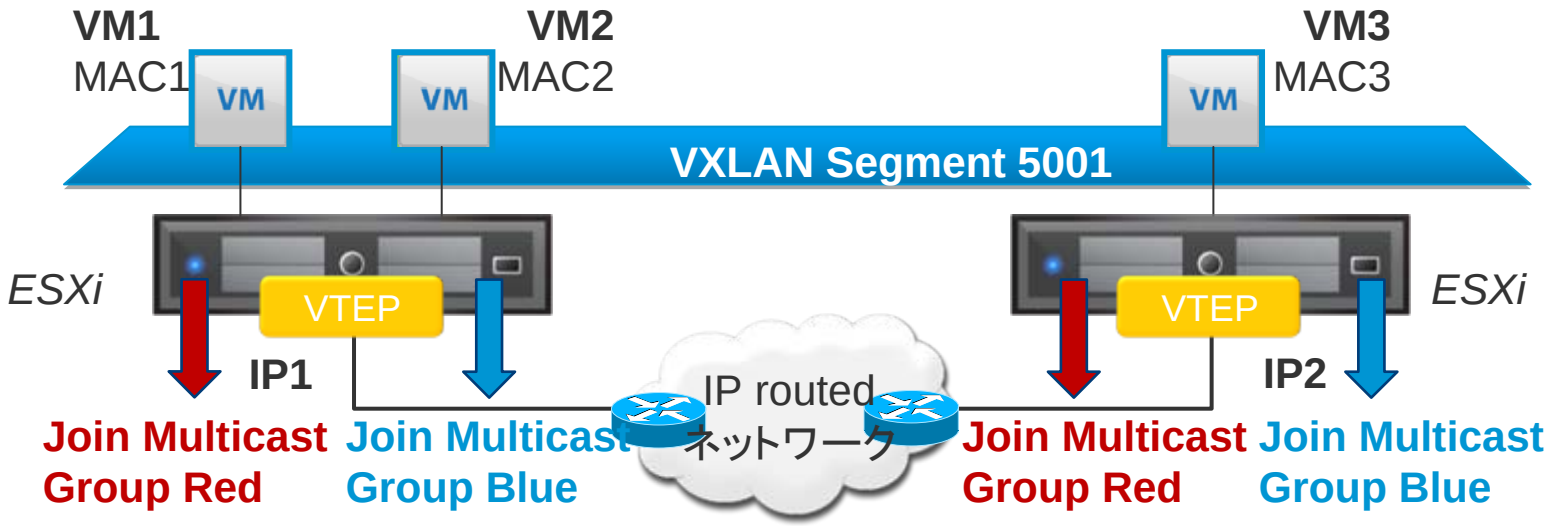
- MAC-in-IP (UDP) によるカプセル化でIPネットワークを伝搬させる
- VTEP が VXLAN トンネルの起点・終点となり、VXLAN Segment ID を追加・解釈する



VXLAN のコントロールプレーン

ノード(MAC) とVTEPのマッピングを送信元情報から自律的に学習

- 学習した内容にしたがって、VXLANカプセル化を行いフレームを転送
- 未学習のノード宛てやmulticast/broadcast はマルチキャストを使って送受信



Segment ID	Inner MAC	Outer MAC/IP
5001	MAC1	Local
5001	MAC2	Local
5001	MAC3	**VTEP2**

Segment ID	Inner MAC	Outer MAC/IP
5001	MAC1	**VTEP1**
5001	MAC2	**VTEP1**
5001	MAC3	Local

VXLAN コントロールプレーン (続き)

参考スライド

ホスト上で構成済みのVXLAN セグメントに仮想マシンが接続されると、VTEPは対応するマルチキャストグループにJoinする

- パワーオンされた仮想マシンがローカルにいなければJoinしなくてよい
- 全ての仮想マシンがパワーオフやMigrationされるとLeaveしてよい

VTEP がReceiverとしてJoin (PIM sparse / Shared Tree)

```
(*, 239.255.28.2), 00:51:15/stopped, RP 192.168.10.254, flags: SJC
  Incoming interface: Null, RPF nbr 0.0.0.0
  Outgoing interface list:
    Vlan20, Forward/Sparse, 00:51:10/00:02:31
    Vlan10, Forward/Sparse, 00:51:15/00:02:26
```

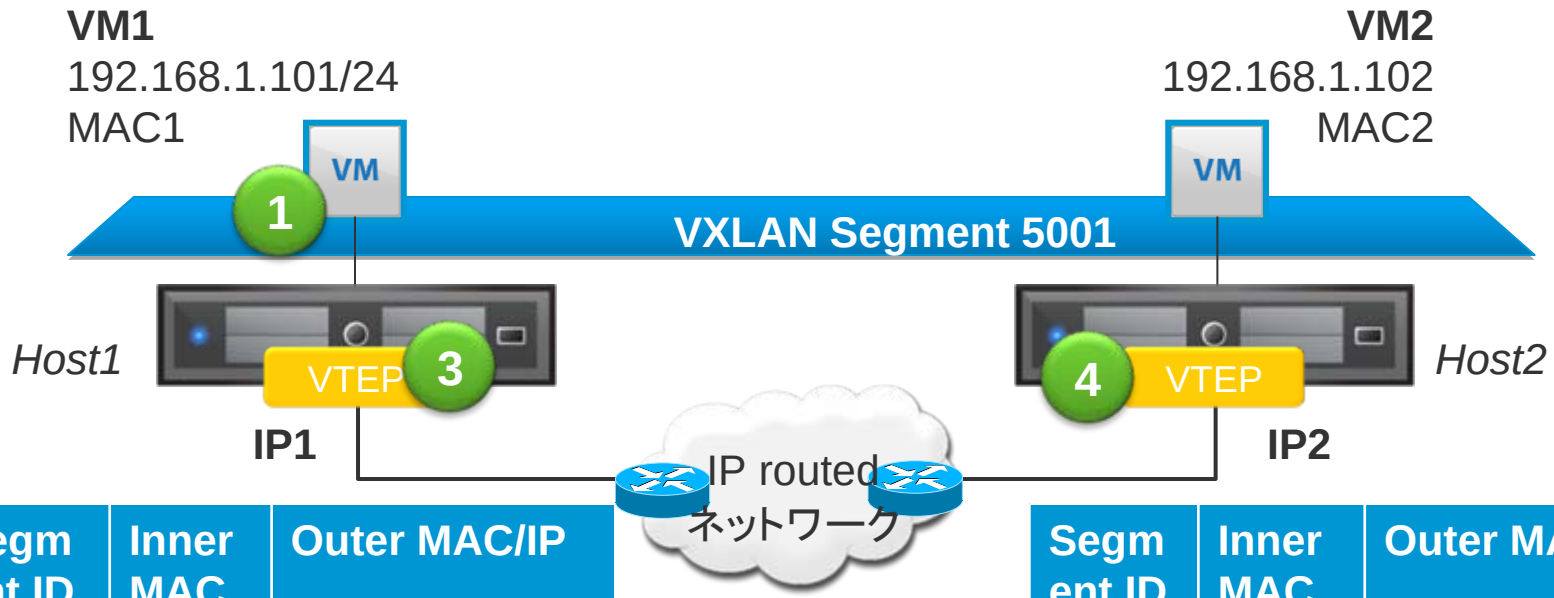
VTEP (192.168.10.1) がSenderとしてRegister (PIM sparse / Shortest Path Tree)

```
(192.168.10.1, 239.255.28.2), 00:00:09/00:02:54, flags: T
  Incoming interface: Vlan10, RPF nbr 0.0.0.0
  Outgoing interface list:
    Vlan20, Forward/Sparse, 00:00:09/00:02:50
```

VXLANの動作例 (既知の宛先へのユニキャスト)

参考スライド

1. VM1がVM2のMACアドレスを宛先とするEthernetフレームを送出
2. Host1 VTEPはVM1が属しているSegment ID5001のエントリの中からMAC2を検索し、VTEP2の先に接続されたノードであることを認識する
3. Host1 VTEP はVXLANカプセル化し、VTEP2が宛先のVXLANフレームを送出
4. Host2 はユニキャストを受信し、VXLANヘッダを取り除いてVM2へ転送



Segment ID	Inner MAC	Outer MAC/IP
5001	MAC1	Local
5001	MAC2	**VTEP2**

2

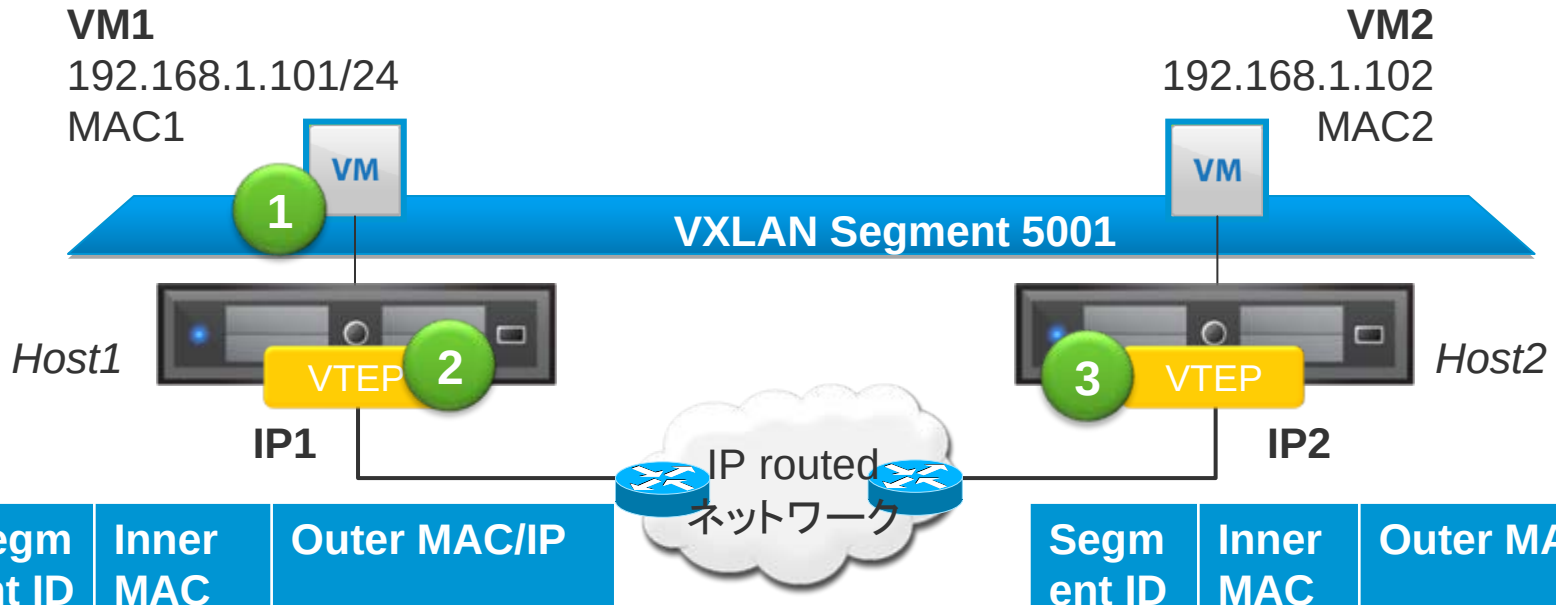
Segment ID	Inner MAC	Outer MAC/IP
5001	MAC2	Local
5001	MAC1	**VTEP1**

vmware

VXLANの動作例 (未知の宛先へのユニキャスト)

参考スライド

1. VM1はVM2のMACアドレス宛てのEthernetフレームを送出
2. Segment ID5001にはMAC2の情報が無いため、Host1 VTEP はVXLANカプセル化を行い、Segment ID5001のマルチキャストグループへVXLANフレームを送出
3. Host2 VTEPはマルチキャストを受信し、VXLANヘッダを除いてVM2へ転
4. 同時に、Host2 VTEPはVM1のMACがVTEP1の先にいることを学習



Segment ID	Inner MAC	Outer MAC/IP
5000	MAC2	VTEPX
5001	MAC1	Local

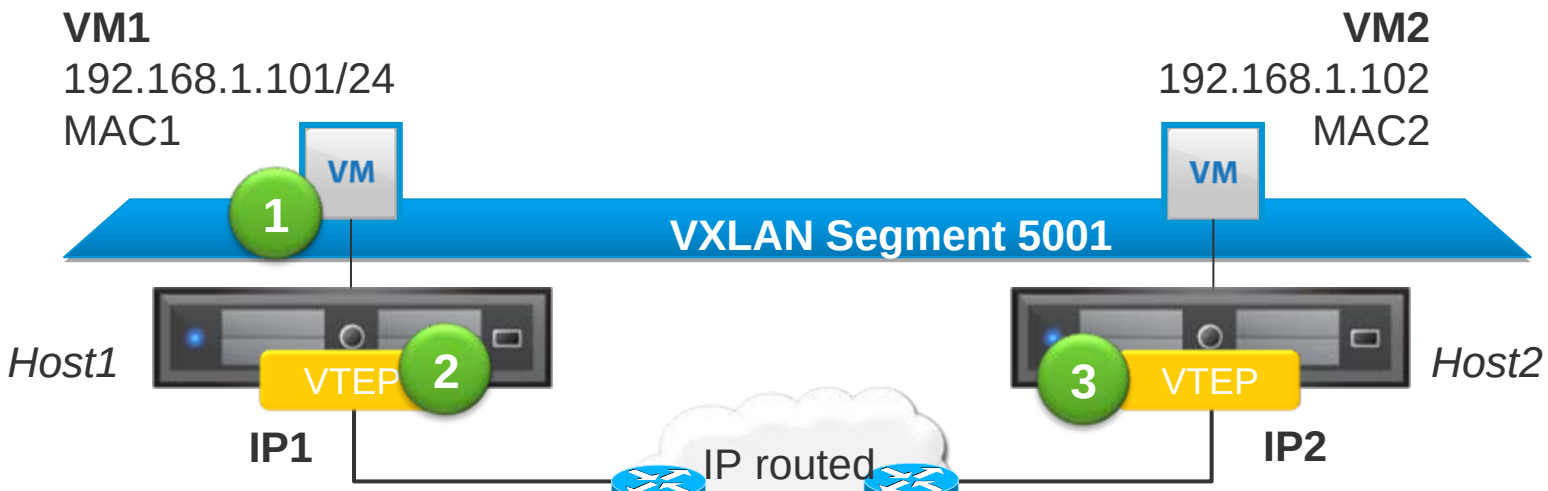
Segment ID	Inner MAC	Outer MAC/IP
5001	MAC2	Local
4 5001	MAC1	VTEP1

vmware

VXLANの動作例 (ブロードキャスト／マルチキャス

参考スライド

1. VM1 は宛先がブロードキャストMACのEthernetフレームを送出
2. Host1 VTEP はVXLANカプセル化を行い、当該Segment IDのマルチキャストグループへVXLANフレームを送信
3. Host2 はマルチキャストを受信、VXLANヘッダを除いて、Localの同一VXLANセグメント内にEthernetフレームをフラッディング
4. 同時に、Host2 VTEPはVM1のMACがVTEP1の先にいることを学習



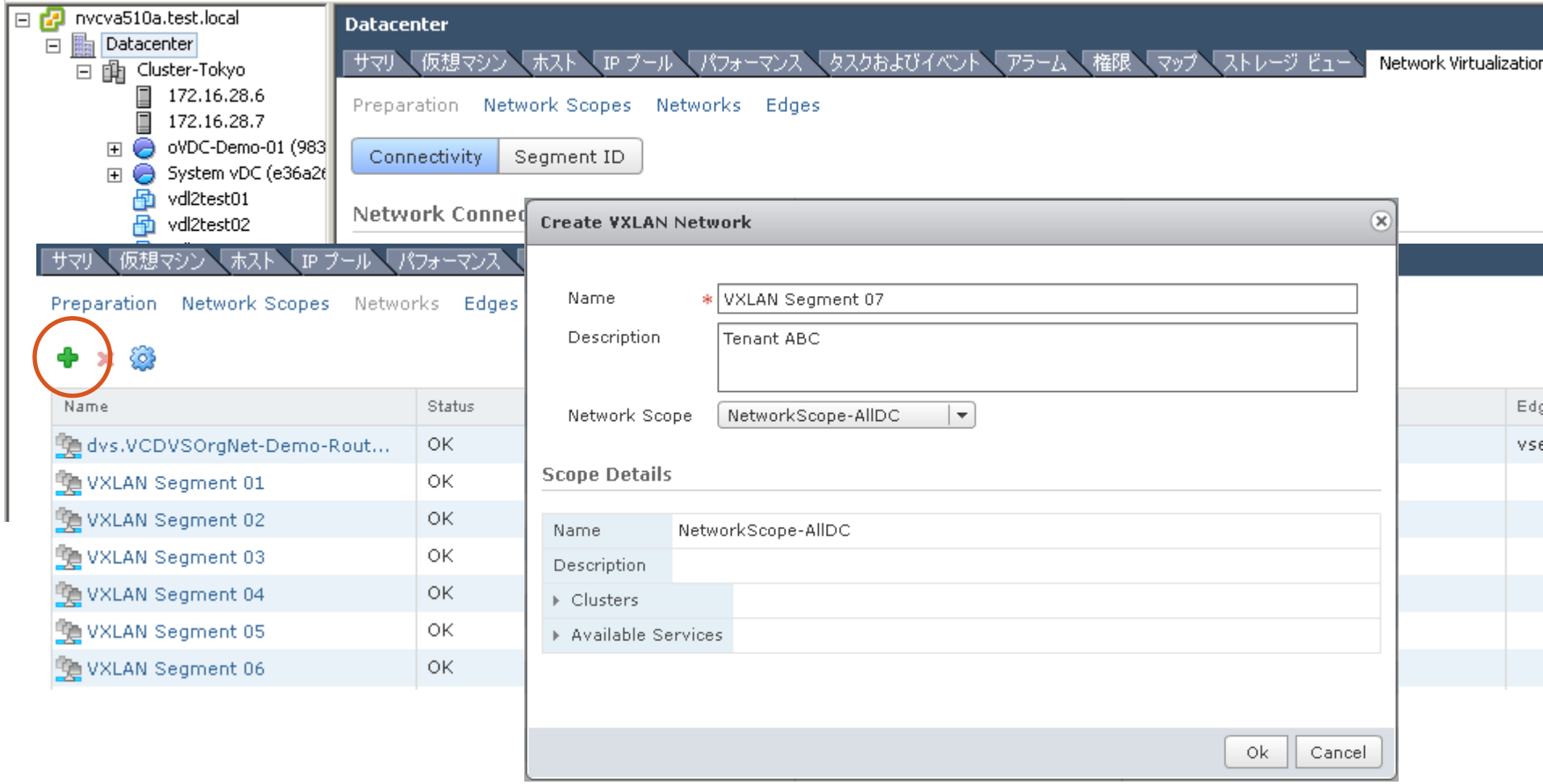
Segment ID	Inner MAC	Outer MAC/IP
5001	MAC1	Local

Segment ID	Inner MAC	Outer MAC/IP
5001	MAC2	Local
4 5001	MAC1	VTEP1

VXLANの構成方法の例 (VMware実装)

vShield Manager から行う (REST API も提供)

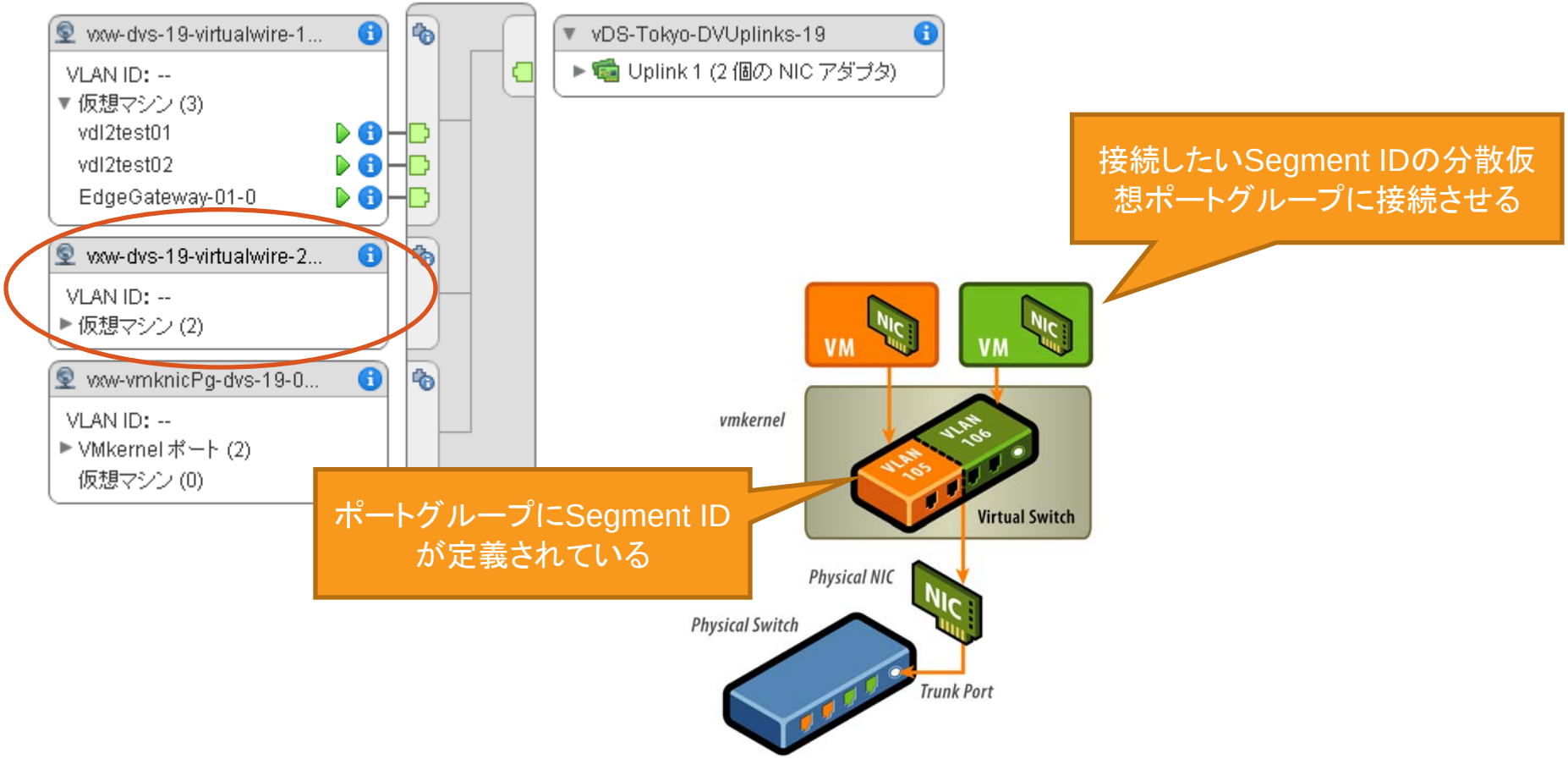
vSphere Client のプラグインとして一元的な管理が可能



仮想マシンからVXLANに接続する例 (VMware実装)

VXLANセグメントは分散仮想ポートグループとして構成される

VTEPの役割はVMkernelモジュールが担う



VXLAN パフォーマンス例 (VMware実装)

詳細は下記をご参照ください
http://www.vmware.com/resources/techresources/10304

アジェンダ

今なぜオーバーレイなのか？

オーバーレイの実装例 – VXLAN

オーバーレイ実装方式の違い

現在の課題と取り組み例 – VXLAN

カプセル化の方式とネットワークコントローラとの関係

カプセル化のプロトコルだけを比較することにはそれほど意味が無い

- ネットワークコントローラーを含めて考えないと、ソリューションとしてのオーバーレイが見えてこない
- カプセル化のプロトコルをどう使うか、が重要
- 念頭に置いたコントローラーの実装のちがいによって、プロトコルの違いが生まれている

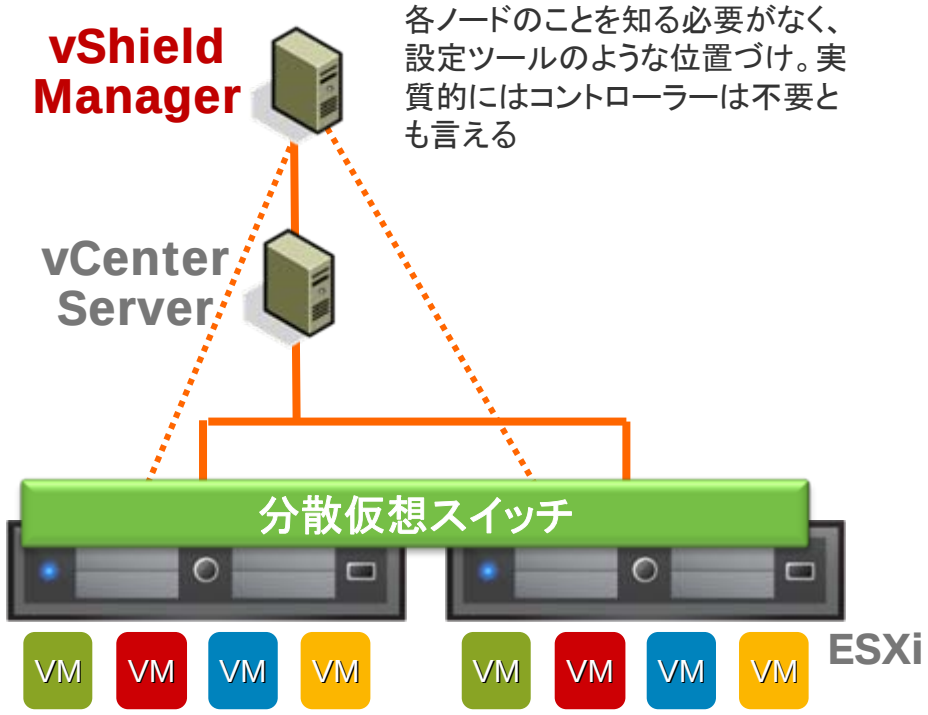
コントローラーと標準化されたプロトコルとの役割分担にも違いがある

- コントローラー部分は仕様が公開・標準化される方向にはないので、プロトコルに依存する部分が大きい方式がInteroperabilityを確保しやすいと言える
- ただし、プロトコルに依存する部分が少ないほど、コントローラーの自由度が高く、より洗練されたアーキテクチャを選択できる余地が大きい

ネットワークコントローラーの例 (VXLAN)

VMware vCNS (vShield Manager) + VXLAN

- 分散自律型のシンプルな構成
- マルチキャストに依存



Cisco Nexus 1000V + VXLAN

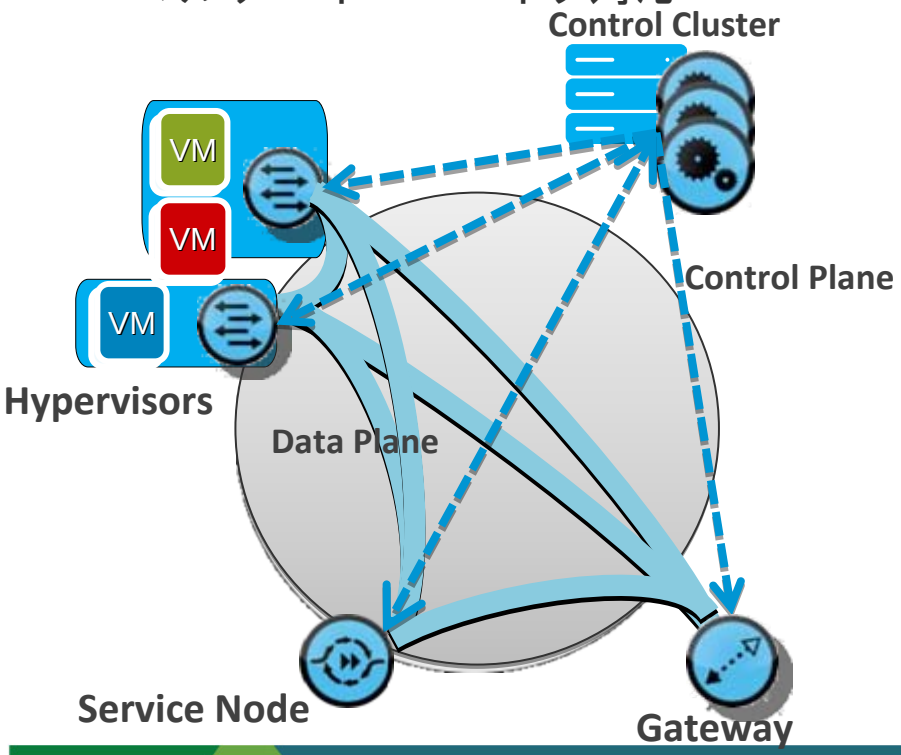
当日のみ

VXLAN ではコントローラーも含めて、エコシステムが広がりやすい仕組み

ネットワークコントローラーの例 (STT / NVGRE)

Nicira NVP + STT

- 洗練された集中管理型アーキテクチャで物理NWへの依存性が極めて小さい
- STT の仕様は公開だが、実質 NVP ありきのアーキテクチャ
- マルチハイパーバイザ対応



MS SCVMM + NVGRE

当日のみ

カプセル化方式そのものがもたらす違い

MTU 変更の必要性

- VXLAN (50byte)、NVGRE (42byte) のオーバーヘッドが発生
- Outer の DF ビットについては実装次第
- STTはパケットドリブンでのカプセル化ではないためMTU変更不要

物理ネットワーク側での ECMP 対応の容易さ

- 物理ネットワーク機器がECMPのために利用できるフィールドが、カプセル化により少なくなる
- VXLAN は Outer UDP src port、STT は Outer TCP src port でフロー識別が可能
- NVGRE は独自のフィールドを認識できる対応機器が必要

カプセル化処理のオーバーヘッドの削減 (物理 NIC オフロード)

- 従来の物理NICオフロードは、カプセル化により基本的に利用できなくなる
- VXLAN は、物理NICベンダとの協業でオフロードを実装中
- STTはTCP を模して作られているためTSO等のオフロードをそのまま利用可能

アジェンダ

今なぜオーバーレイなのか？

オーバーレイの実装例 – VXLAN

オーバーレイ実装方式の違い

現在の課題と取り組み例 – VXLAN

オーバーレイネットワークの今後の課題の例

物理ネットワーク構成(マルチキャスト)への依存性

- NVP/STTはほぼ問題を解消済み。VXLANではドラフトに言及あり。NVGREは？

カプセル化処理のオーバーヘッドの排除

- STTはほぼ問題を解消済み。VXLANはエコシステムで取組中。NVGREは？

既存ネットワークとの相互接続

- VXLAN/NVGREでは短期的にはルーテッドとなる。既存VLANとのブリッジングや、遠隔ネットワークとのL2接続は今後のテーマ
- NVPはGatewayにて、すでに既存ネットワークとのL2ブリッジ可能

Interoperabilityの更なる強化

- 今のところ、ネットワークコントローラーも含めて多様なソリューションが提供される可能性があるVXLANで特に有効。

ネットワークサービスの提供

- 論理ネットワークの接続性・分離性だけでは SDN と言えない
- ネットワークサービス (Security / LB / DHCP, DNS ...) をどのように提供するかも大きなテーマ

セッション当日は、前スライドの各項目について
業界での現在の取り組みの具体的な内容を
ご紹介しましたが、
現時点では不確定な要素も多いため、
配布資料からは割愛させていただきます。

VXLAN Ecosystem

VXLAN Gateway

ARISTA

BROCADE

f5

AVAYA

BROADCOM

VXLAN Server Offload

intel

BROADCOM

EMULEX
We network storage

VXLAN Visibility / Efficiency

APRESIA

BROCADE

riverbed

VXLAN vDS

CISCO

物理環境・仮想環境を問わず、スケーラブルに VXLAN を利用可能に

vmware

まとめ

オーバーレイネットワークはSDNの流れの中でも、現実の問題を、現実的な手法で解決しようとする「現実的なアプローチ」

すでに使えるソリューションとして市場にリリースされており、実際に利用できるフェーズに来ている

一方、課題も少なからず存在するため、どの規模まで適用するか事前に検討した上で導入し、発展に合わせて適用規模を拡大していくことがのぞましい

今後の、ネットワークコントローラーを含めたソリューションの発展、業界全体でのオーバーレイへの取り組みに大いに期待

VMware はネットワーク仮想化に今後ますます注力していきます

The Global Leader in Business Infrastructure Virtualization