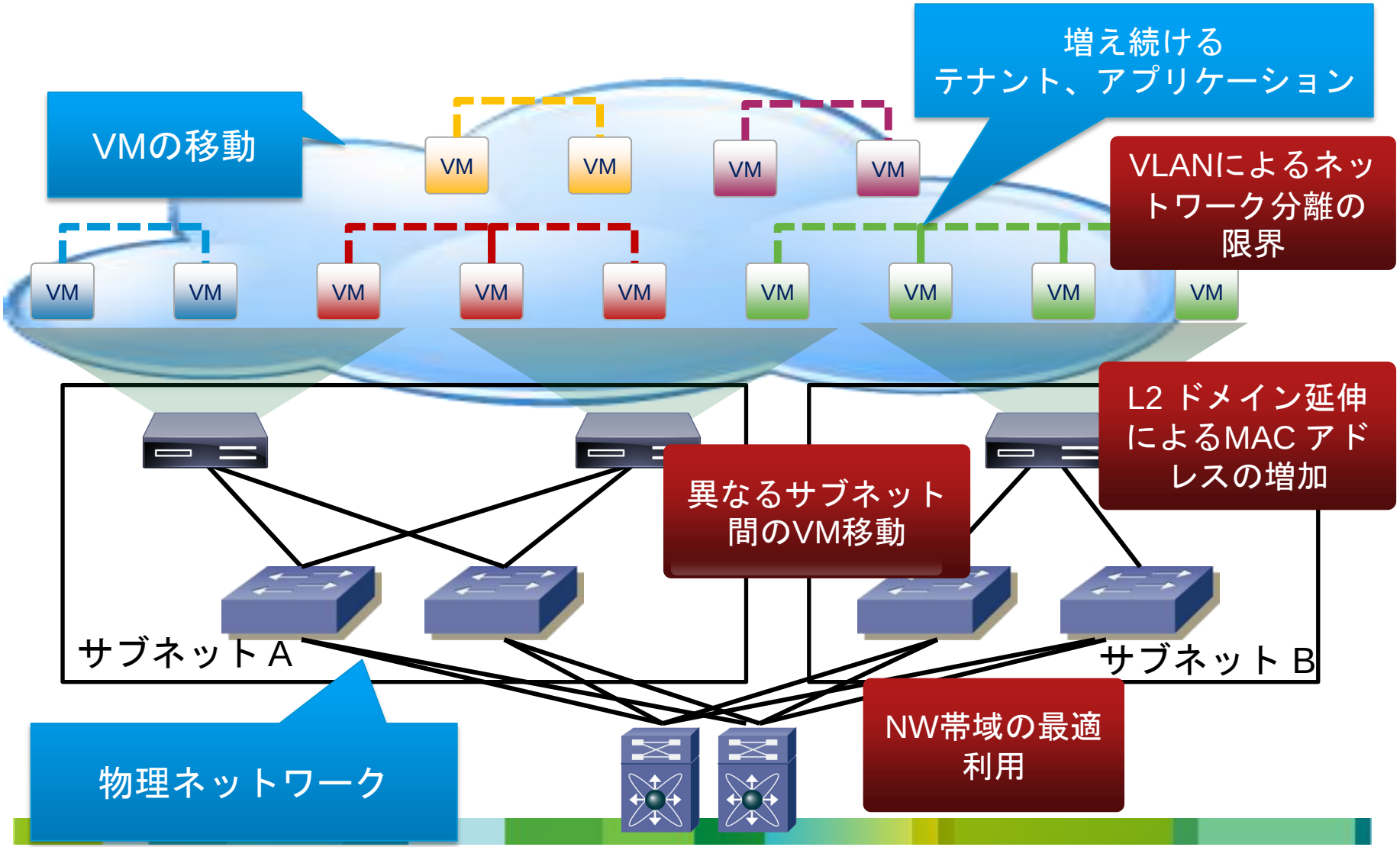
どうする？どうやる？データセンター間ネットワーク（L3オーバーレイ）

シスコシステムズ合同会社
中本滋之

2012/1/20

クラウドにおけるネットワークのチャレンジ

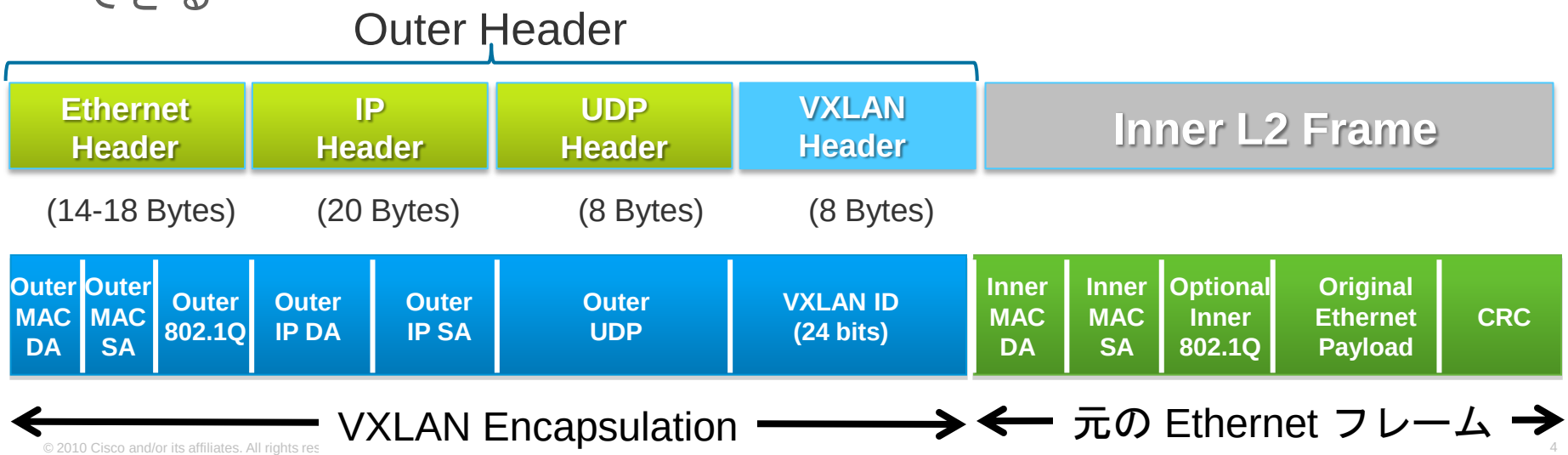


IP ネットワーク上にオーバーレイ

- VXLAN
  - UDP でカプセル化（50 bytes のオーバーヘッド）
  - VMware, Cisco, Arista, Broadcom, Citrix, Red Hat
- NVGRE
  - GRE でカプセル化（42 bytes のオーバーヘッド）
  - Microsoft, Intel, Dell, HP, Broadcom, Arista, Emulex
- 特徴
  - IP ネットワーク上でイーサネットフレームを転送
    - アクセススイッチ(VTEP/ NVGRE Endpoint)でトンネル
  - IP マルチキャストを利用
    - セグメント内のブロードキャストとマルチキャスト
  - 24 ビットのセグメント ID (VNI/TNI)

Virtual Extensible Local Area Network (VXLAN)

- IP オーバーレイネットワーク上のイーサネット
  - UDP 50バイトのオーバーヘッドにより、L2 フレームを完全にカプセル化。
  - 24ビットの VXLAN 識別子
  - 16 百万の論理ネットワーク
- VXLAN は Layer 3 を越えることができる
- VTEP 間でトンネル
  - VM は、VXLAN ID を意識しない
- L2 ブロードキャスト/マルチキャスト/Unknown unicast には、IP マルチキャストを利用
- ソフトでもハードでも実装可能
- IETF にて標準化中



VXLANの構成要素

- VXLAN Segment
  仮想マシンが通信を行う L2 Overlay ネットワーク
- VXLAN Overlay Network
  VXLAN Segment の別名
- VNI (VXLAN Network Identifier)
  VXLAN ID
- Virtual Wire
  同一L2セグメントをつなぐトンネル。VTEPで終端される
- VTEP (VXLAN Tunnel End Point)
  トンネル終端ポイント。IPアドレスを持つ
- VXLAN Gateway
  VXLAN と VLAN(non-VXLAN) の接続機能を提供する

VXLAN の基礎

- L2 ブリッジの通信同様フラッド&ラーニングによる転送メカニズム

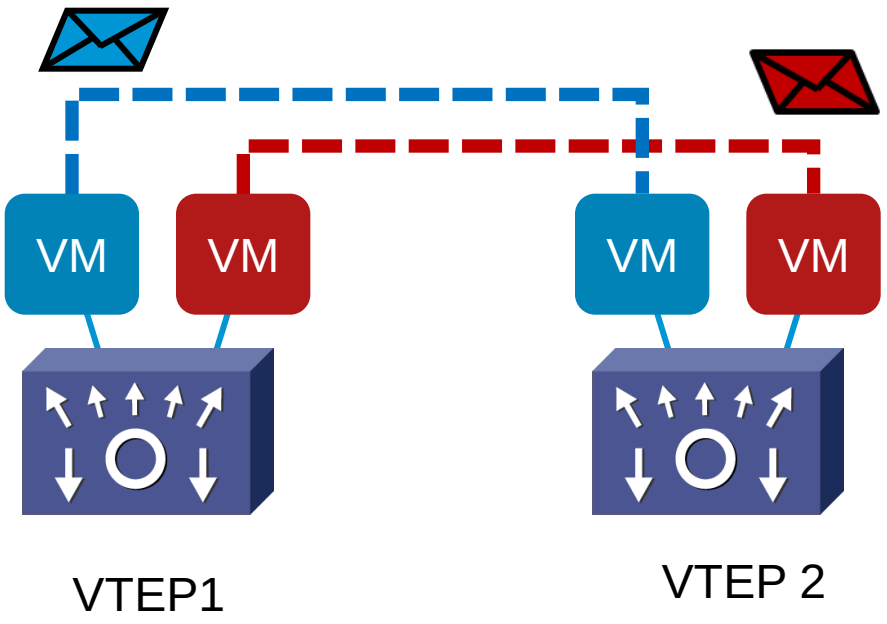
  VTEPが VM's Source (MAC, Host VXLAN IP) tupleを学習

- ブロードキャスト、マルチキャスト、Unknown ユニキャスト

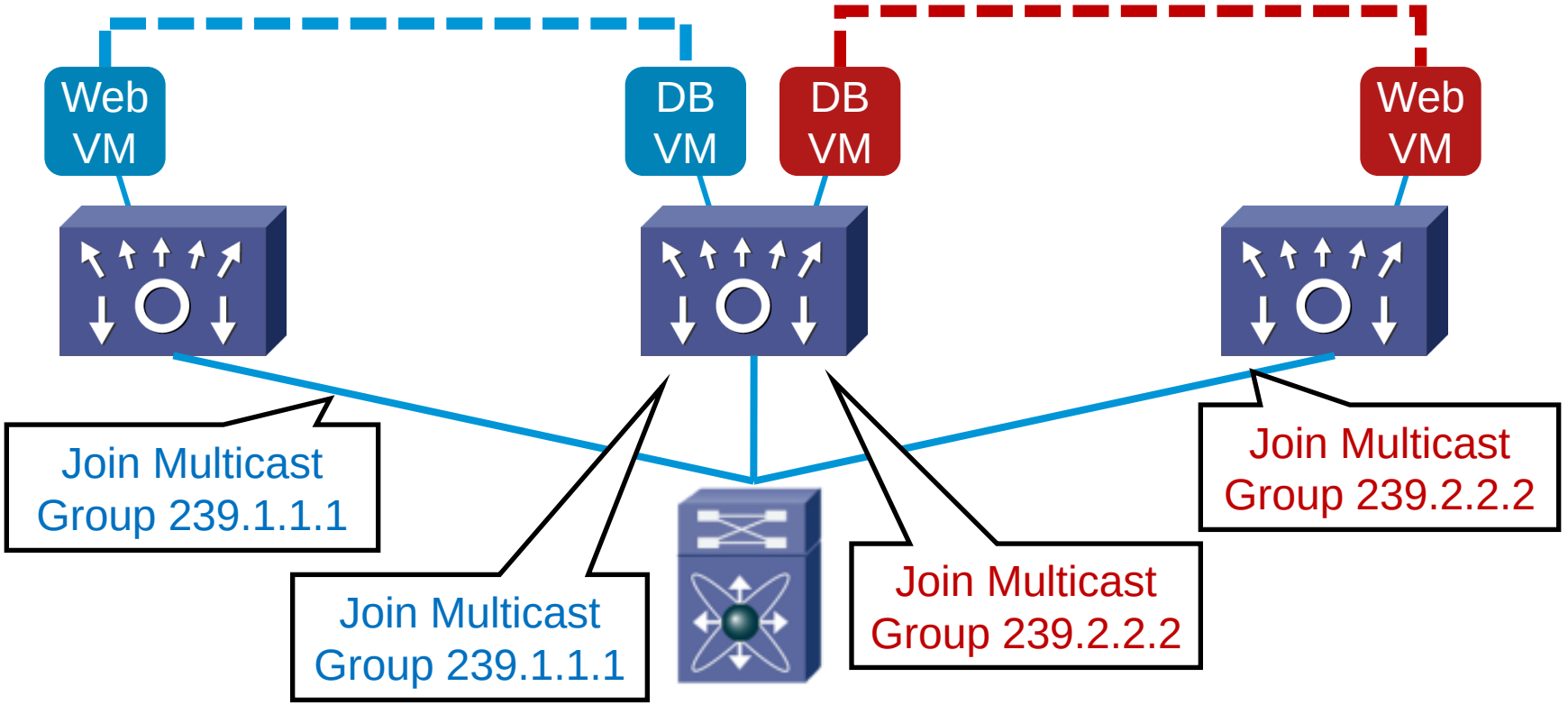
  これらは、マルチキャストとして送信される。

- ユニキャスト

  ユニキャストパケットは直接 VXLAN IP（送信先 VTEP）に転送



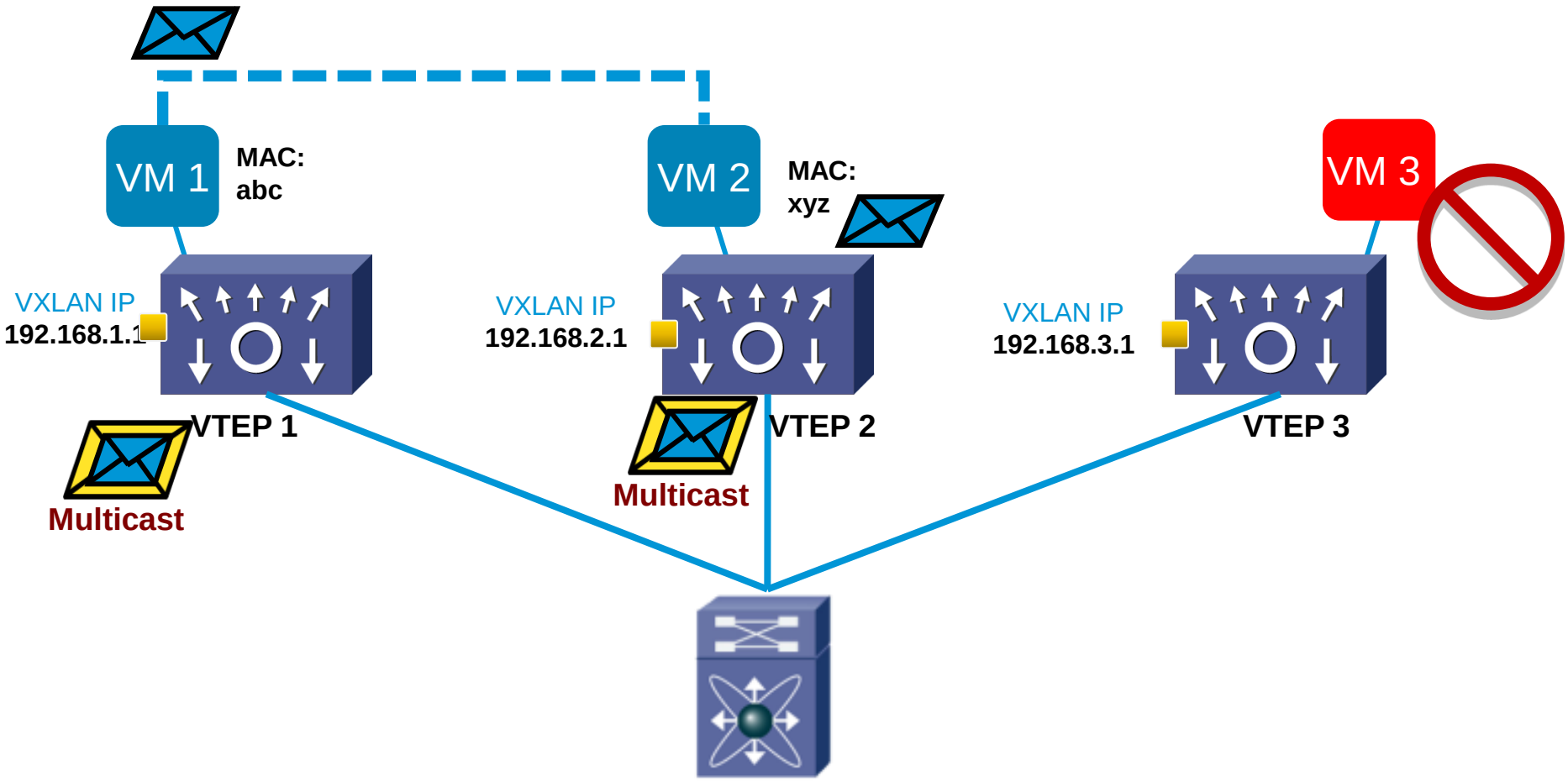
VXLAN コントロールプレーン



- **VXLAN セグメント毎にマルチキャストグループを構成**
  - 仮想マシンとVNI、マルチキャストグループへの関連付けを実施
- **効率的な転送**
  - 仮想マシンが配下にいなければ、該当VNIのマルチキャストグループから脱退

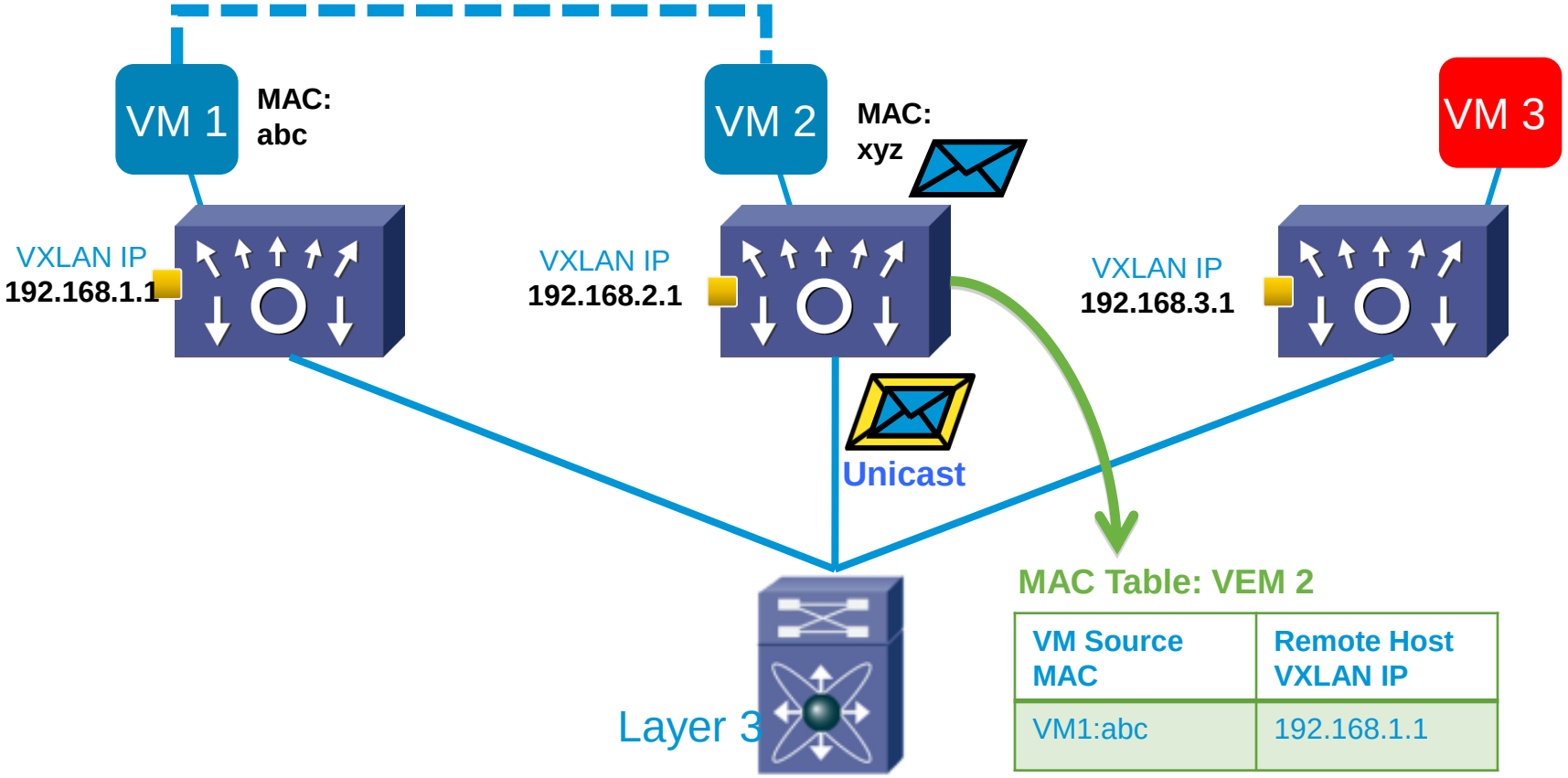
VXLAN の通信

VM1 は VM2 と VXLAN 内で通信



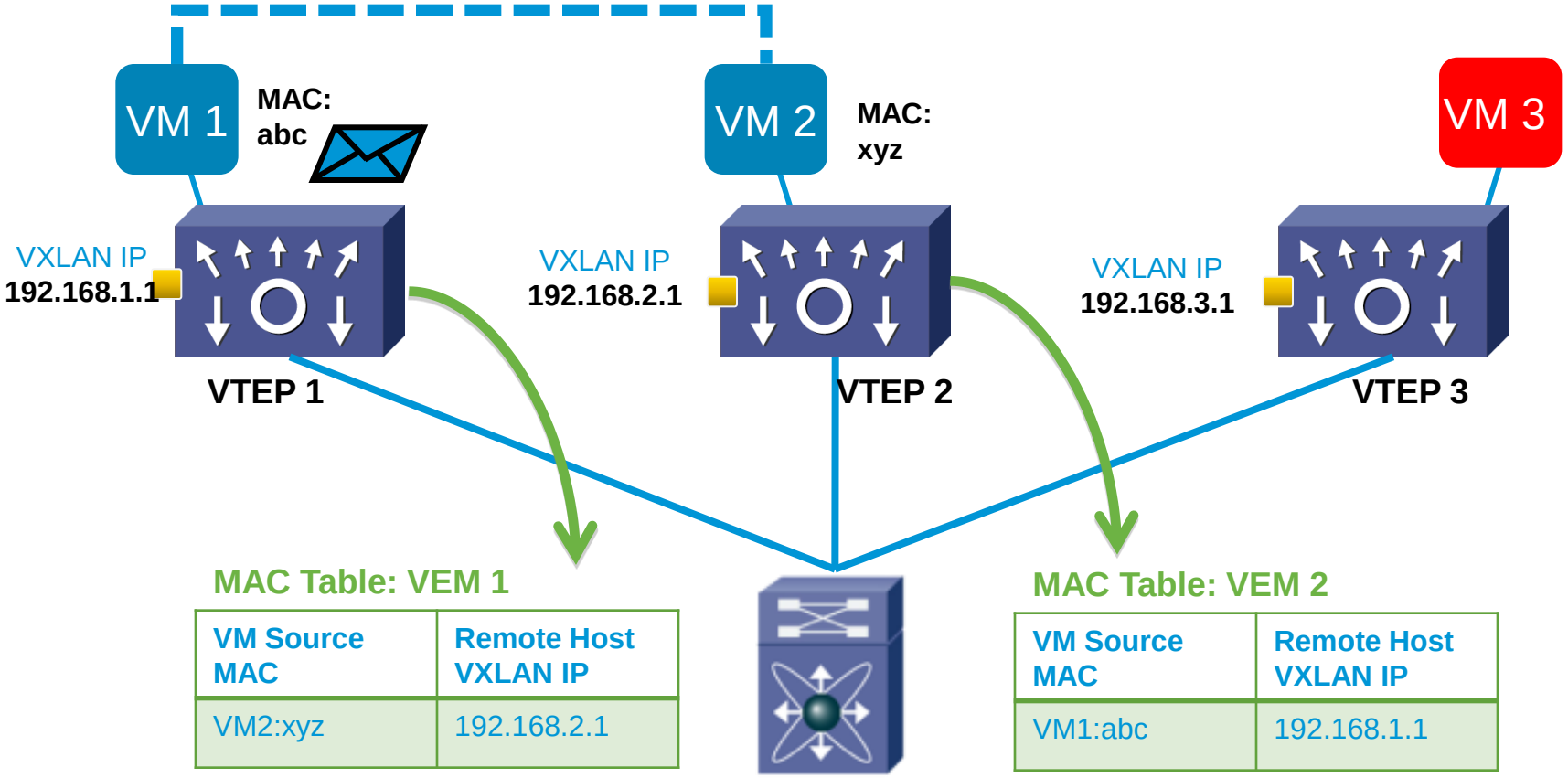
VXLAN の通信

VM1 は VM2 と VXLAN 内で通信



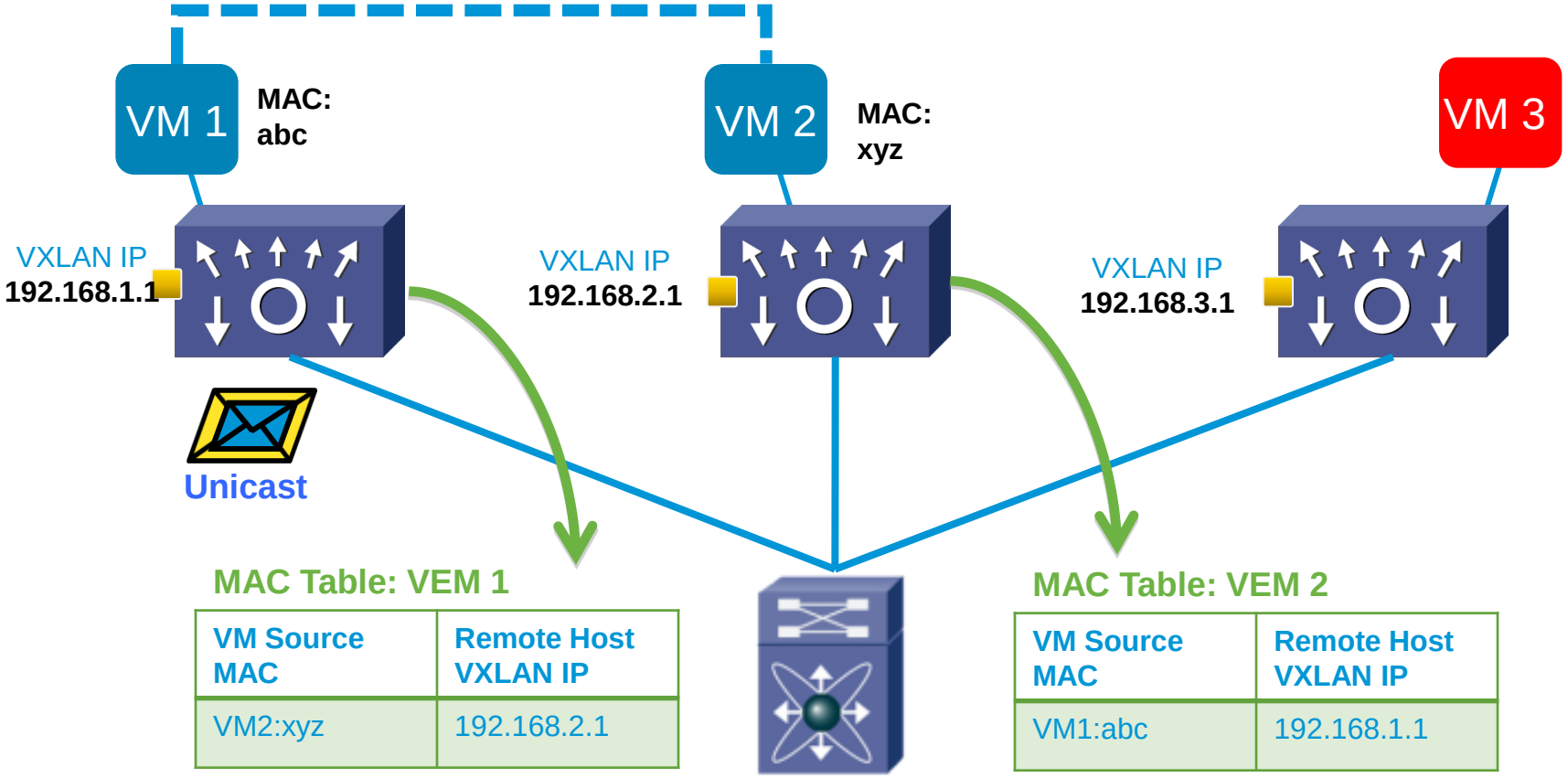
VXLAN の通信

VM1 は VM2 と VXLAN 内で通信



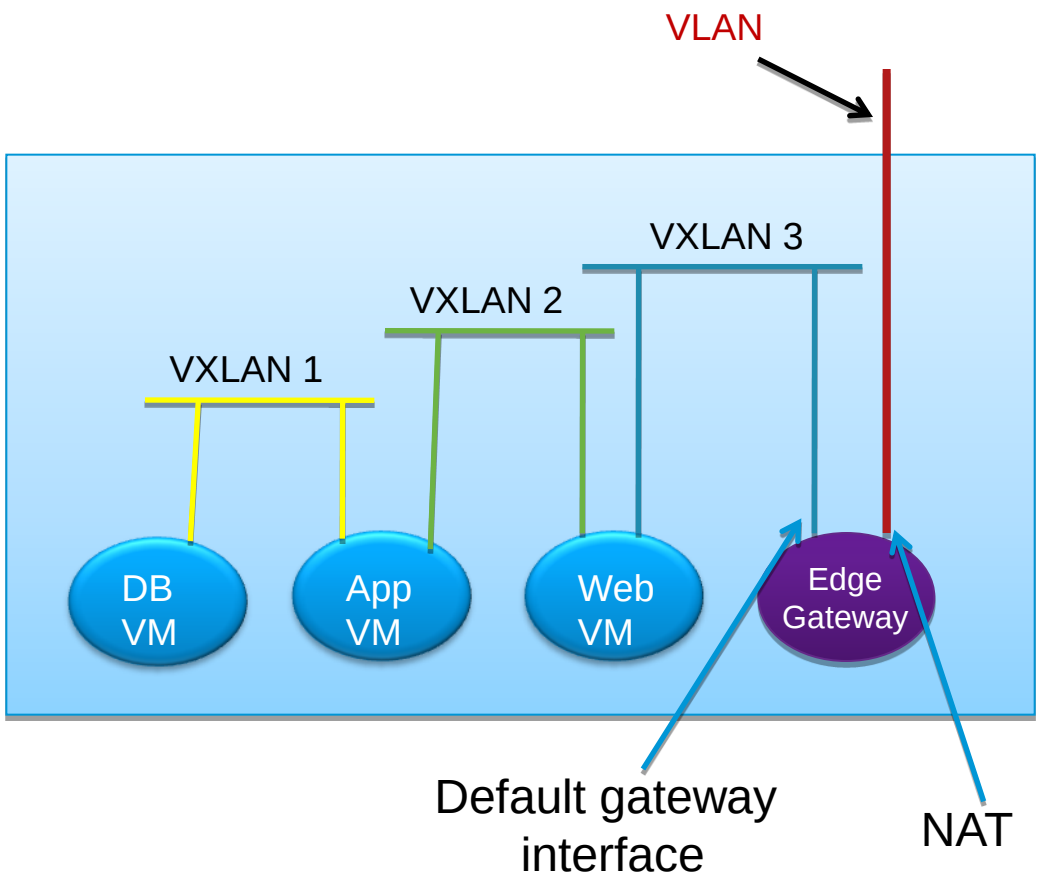
VXLAN の通信

VM1 は VM2 と VXLAN 内で通信



Layer 3 Connectivity

- VXLAN と VLAN の接続 (VXLAN to VLAN)は、ゲートウェイが必要。
- 必要に応じて、VPN または NAT を使用して外部ネットワークのゲートウェイへ接続
- 多くのセグメントは、L3 接続の必要はない。

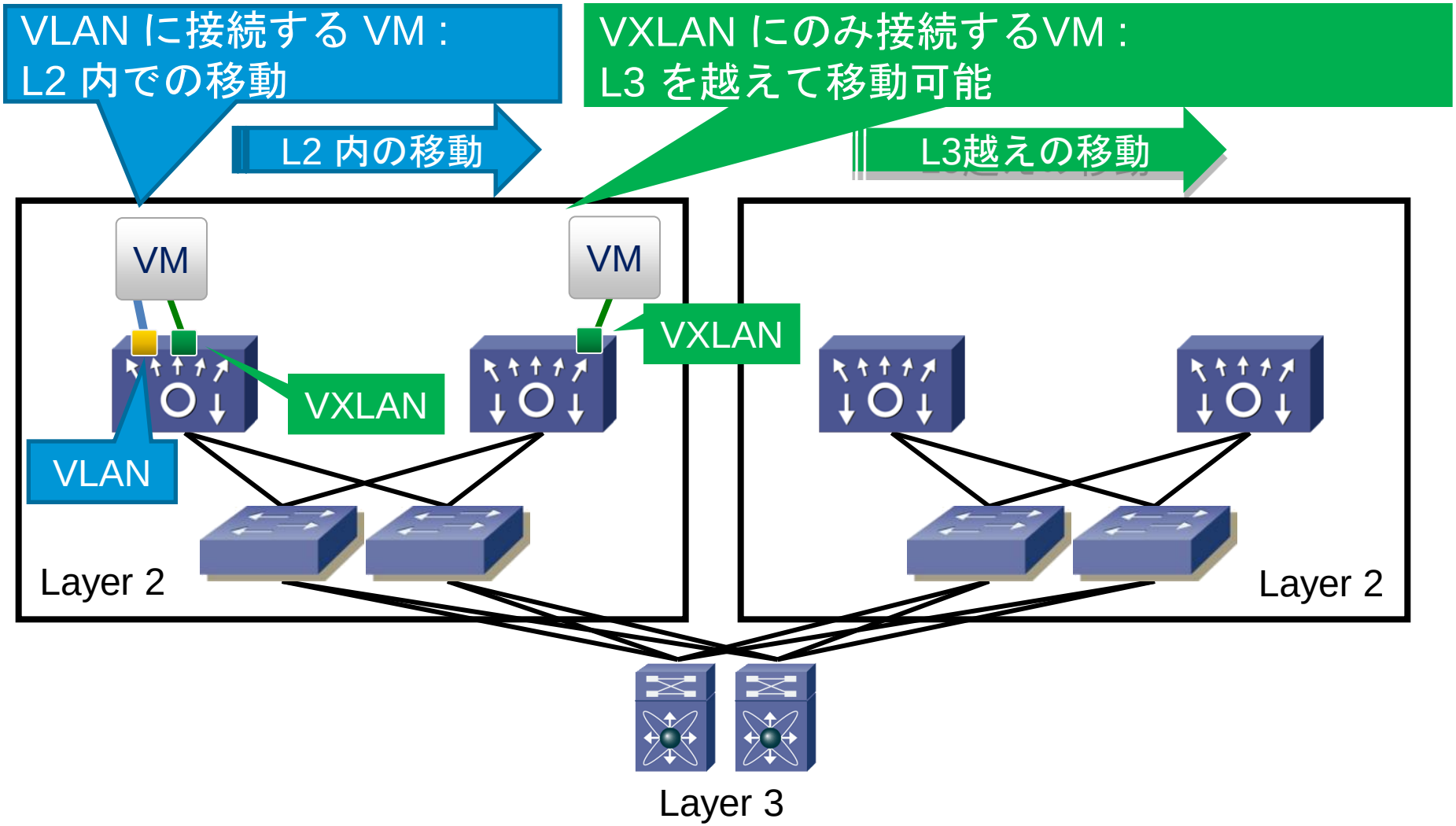


VXLAN インフラに求められること

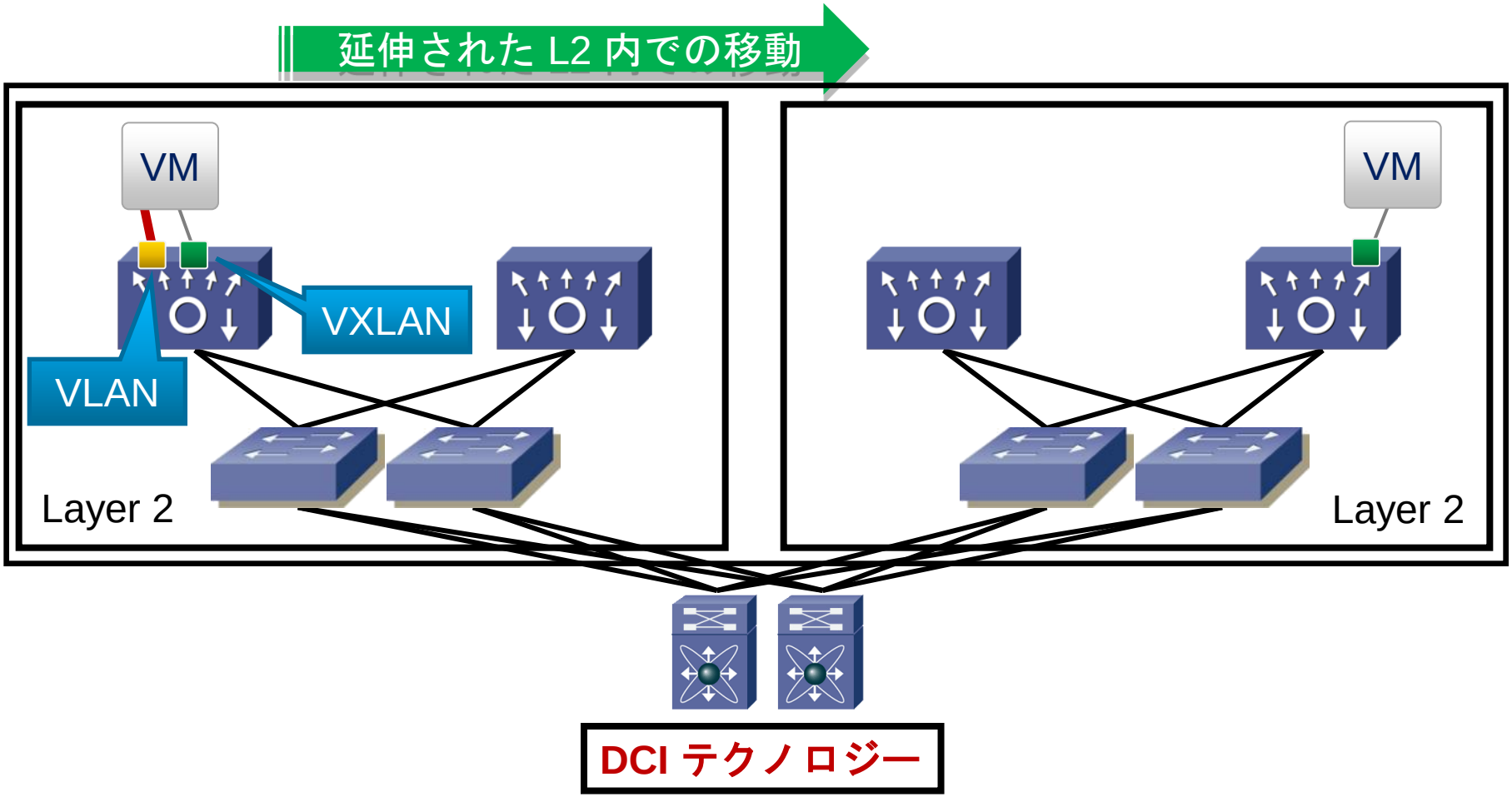
- IP マルチキャスト が必須
  - マルチキャストグループが多いほど良い。
  - 複数のセグメントを一つのマルチキャストグループにマップ可能
  - L2ネットワークでは、IGMP クエリーを有効にする。
  - VXLAN がルータを越える場合は、マルチキャストルーティングを有効にする。

- VXLAN のオーバーヘッドを考慮
  - VM の VNIC MTU サイズより、50byte のオーバーヘッドが付与されるため、物理インフラの対応が必要。
  - e.g. 1500 MTU on VNIC -> スイッチとルーターは、1550 MTU

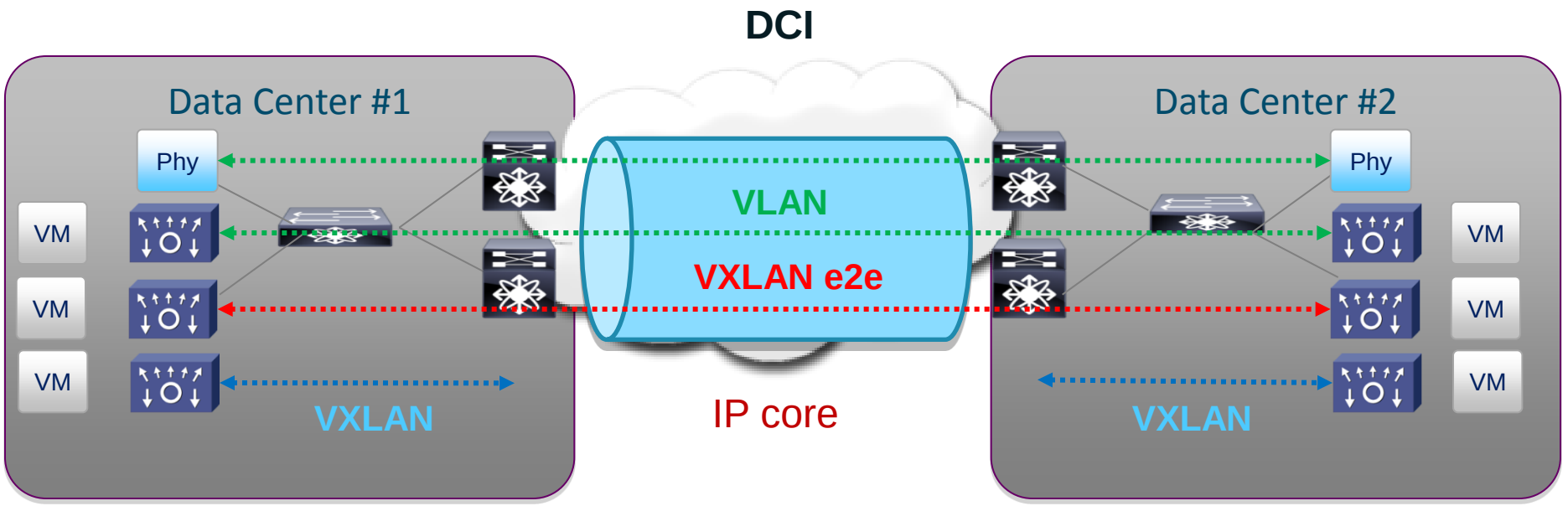
VLAN と VXLAN の移動範囲



DCI を利用した場合の移動範囲



VXLAN と DCI



VXLAN とその他 DCI 技術は共存する？

VXLAN が解決する課題

- 広範囲で安全な L2 ネットワークの構築
  STPの問題、リンクの有効利用
- マルチテナント
  VLAN 数
- MACアドレステーブルサイズ
  サーバー仮想化およびL2 ドメインの拡張により TORスイッチ学習すべき MAC アドレスの劇増

Thank you.