Interop2012

新たなネットワーク仮想化技術: L3 NWへのオーバーレイ方式

さくらインターネット(株)
研究所 大久保 修一
ohkubo@sakura.ad.jp

自己紹介

- 所属：さくらインターネット(株) 研究所
- 氏名：大久保 修一
- 数年後のビジネスのネタになりそうな技術の評価、検証など
  - クラウド技術
    ネットワーク仮想化、分散ストレージ
  - IPv4アドレス枯渇対策
    IPv6移行技術
    トランスレータ、トンネル技術(6rd、4rd他)
- 「さくらのクラウド」ネットワーク担当

Agenda

- データセンターネットワークの要件
- 従来の実装方法と問題点
- オーバレイ方式による解決策
- オーバレイ方式実装における課題と解決可能性の展望
- まとめ

データセンターとは？

- 本日のセッションでは、以下のようなデータセンターを想定します。
  - 多数（数万台規模）のサーバを提供、管理
  - インターネット接続、VPN接続を提供
  - コロケーション、物理サーバ、仮想サーバ、レンタルサーバ

弊社（さくらインターネット）の例

ハウジングサービス

ハウジング

顧客が所有するサーバなどの機器類を設置するスペースと回線、電源などを貸与するサービス

専用サーバサービス

さくらの専用サーバ sakura dedicated server

専用サーバ Platform St

専用サーバ Platform Ad

1台　～　複数台

クラウドサービス

さくらのクラウド sakura cloud

高性能サーバと拡張性の高いネットワークを圧倒的なコストパフォーマンスで実現

仮想サーバサービス

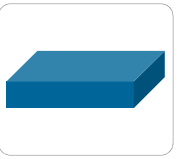
さくらのVPS

仮想化技術を用いて、1台の物理サーバ上に複数の仮想サーバを構築し、仮想専用サーバとして利用するサービス

レンタルサーバサービス

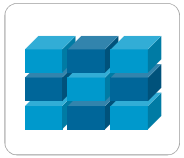
さくらのマネージドサーバ

1台を専有



さくらのレンタルサーバ

1台を共有



・サービスの主な利用用途

エンタープライズ

SNS、Webアプリケーション、SaaS、ASP

会員制サイト、キャンペーンサイト

ネットビジネス、電子商取引、動画・音楽配信

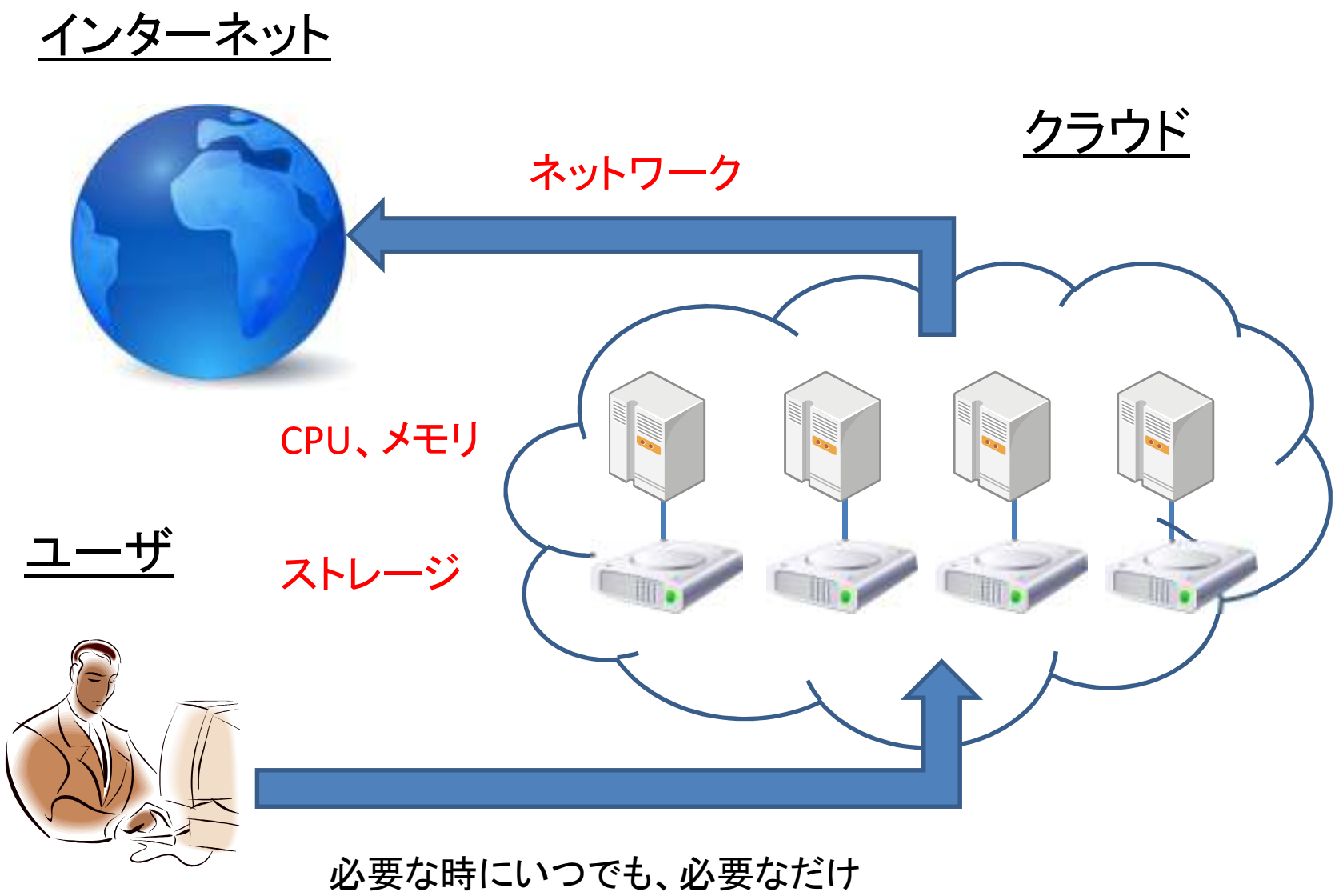
インターネットメール、Webサイト運営

DCネットワークへの要求の高まり

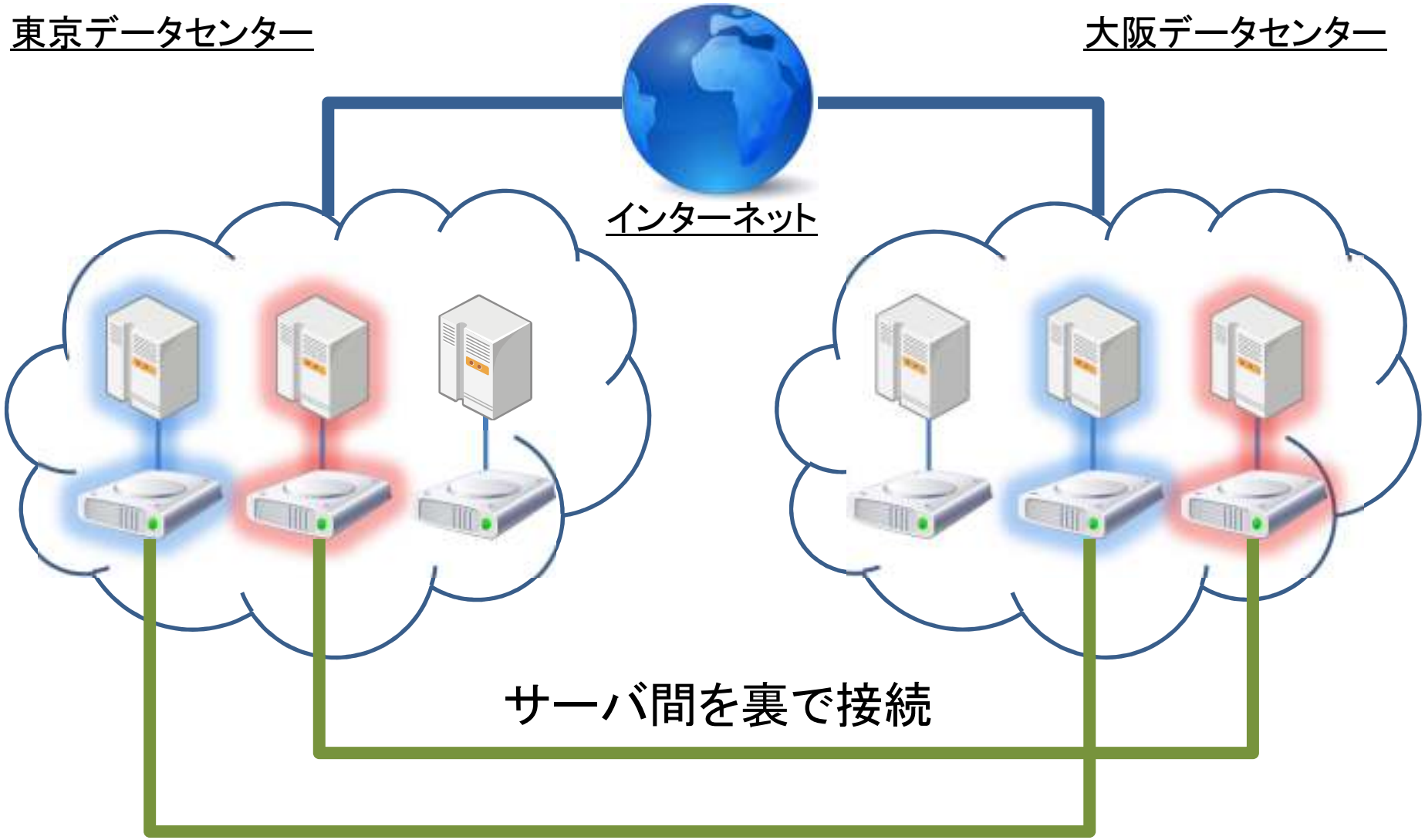
以前に比べ、柔軟性が求められるようになった

1. コンピューティングリソースのプール化
   – クラウドサービス、マルチテナント
2. プラットフォーム化
   – サービス間ローカル接続
3. ディズアスタリカバリ
   – 拠点間ローカル接続
4. モビリティ
   – ワークロードの分散、VMのライブマイグレーション等
5. オンデマンド、セルフサービス
   – 即時提供、設定の自動化

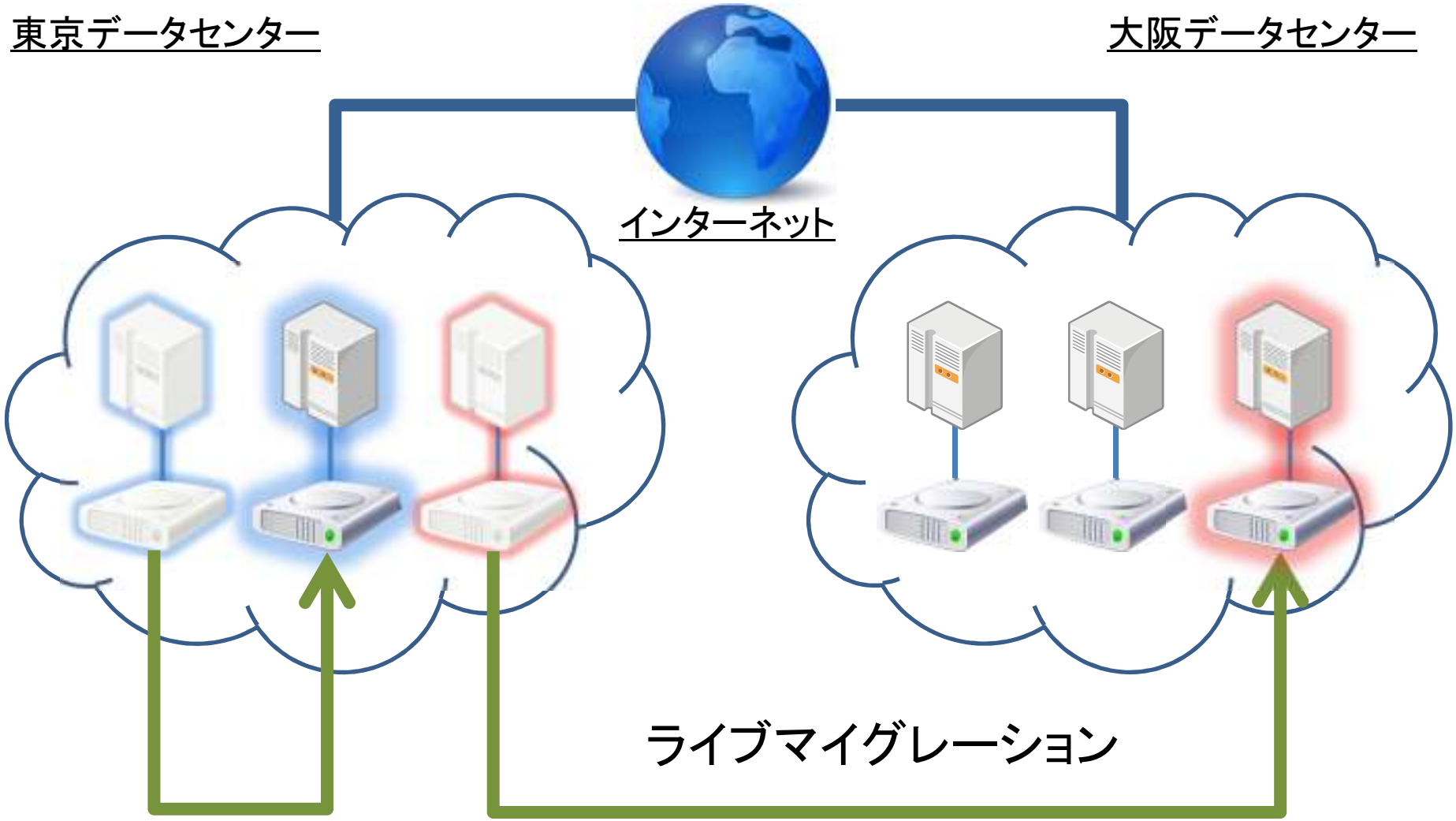
クラウドサービス



ローカル接続



モビリティ



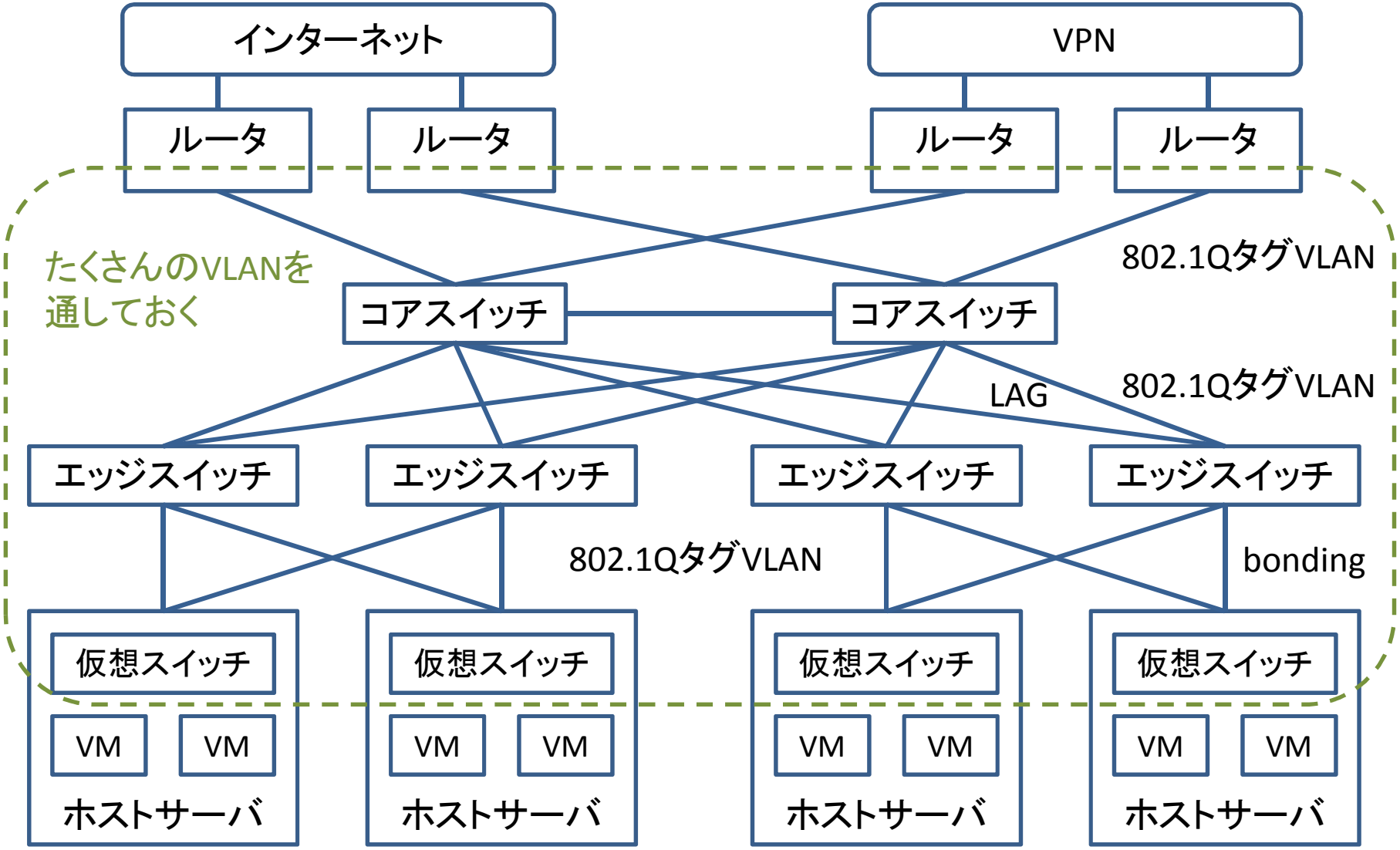
従来の実装方法

- これらの要件を満たすために従来取られてきた手法は？

- 広帯域のL2網を広域に伸ばしておく。
- 高密度のL2スイッチを多数配置。
- スイッチ間は10Gのトランク接続。
- VLANを用いたマルチテナント。

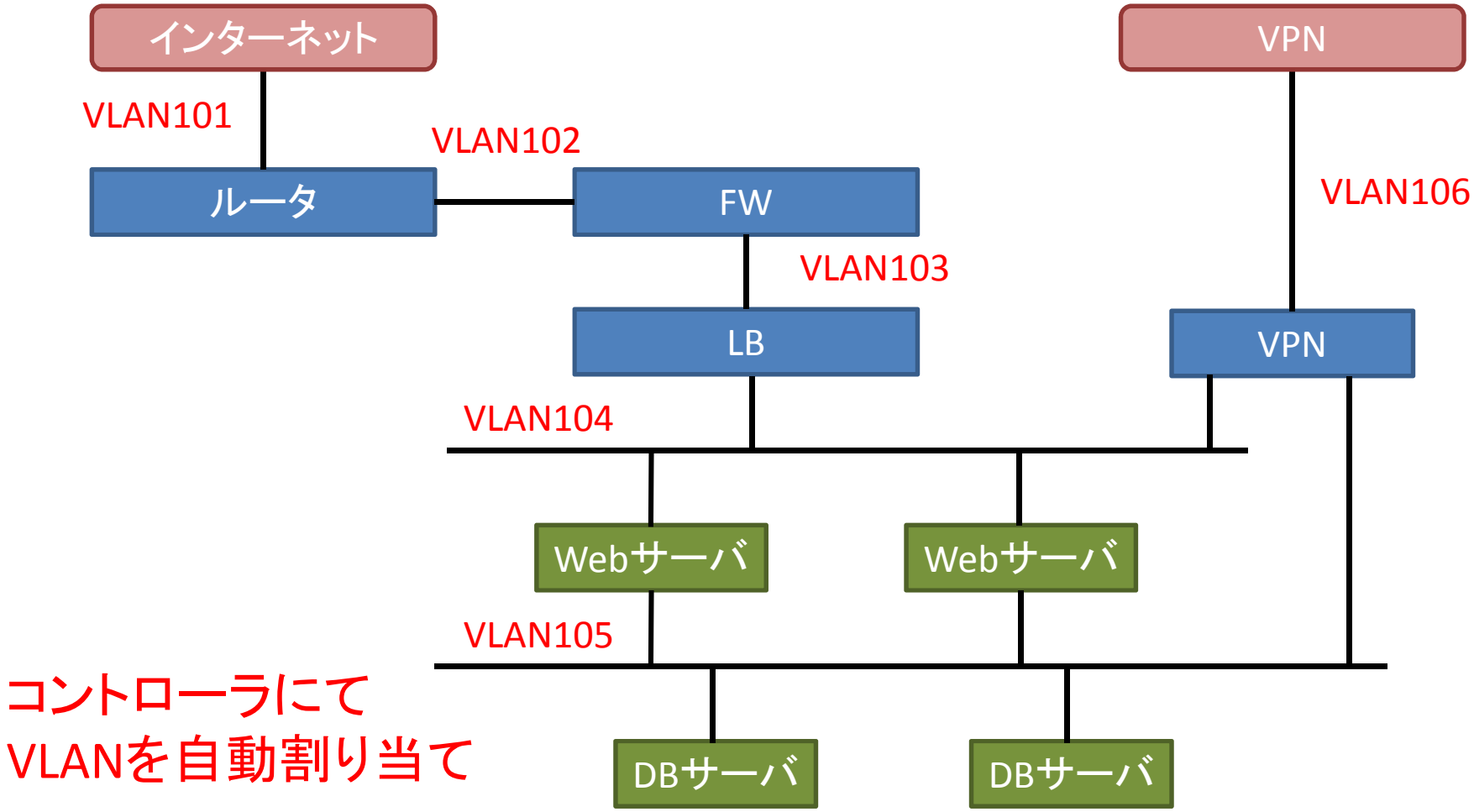
- L2網に接続すれば、どのポイント間でも自由に結ぶことが可能。

IaaSの物理ネットワーク構成例



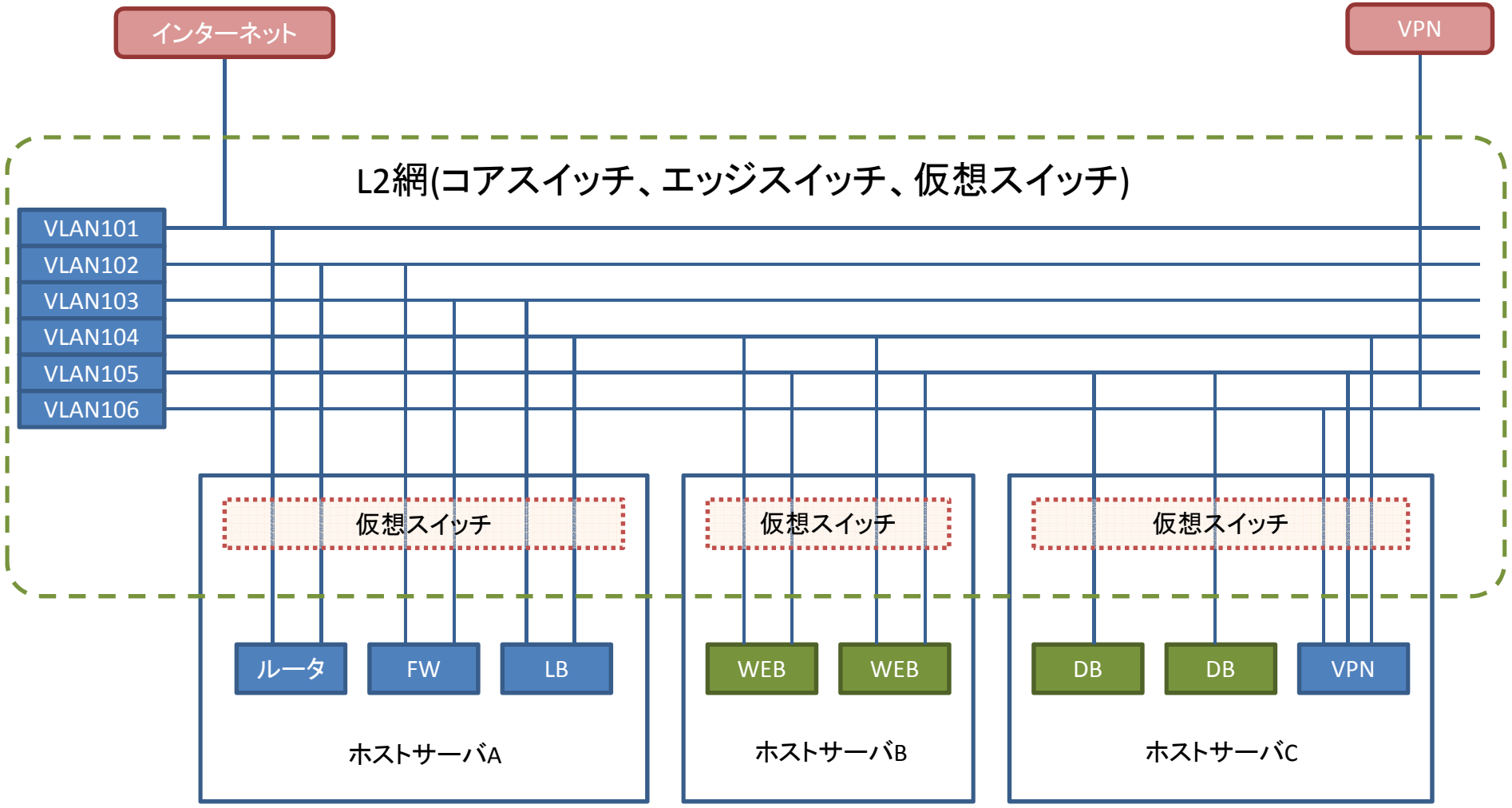
仮想システムプロビジョニング例

このようなシステムを、IaaSインフラ上に展開してみる



コントローラにて
VLANを自動割り当て

クラウド上に配置したシステム

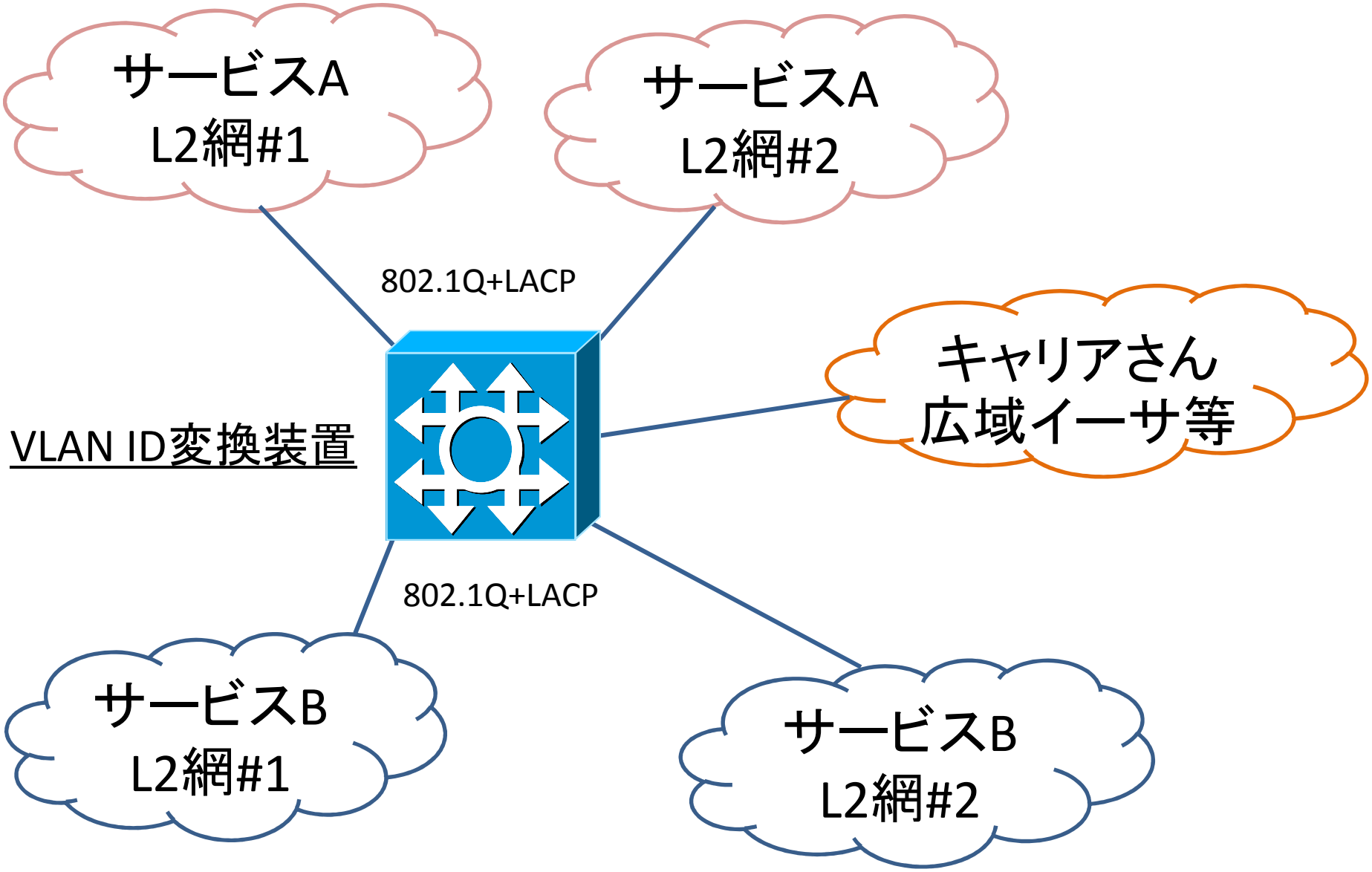


VMはどのホストサーバ上に配置してもよい

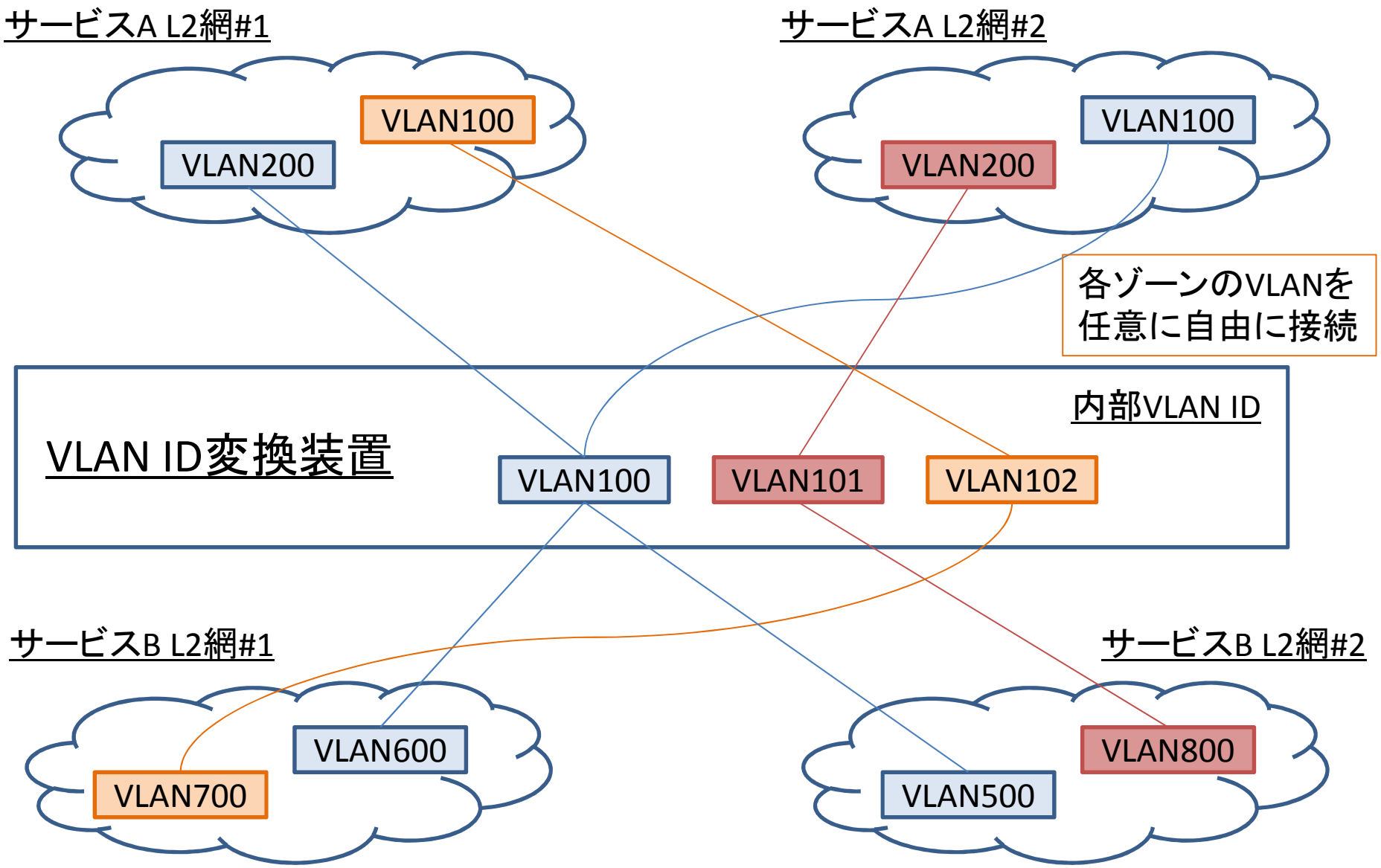
サービス間、拠点間の接続

- 各サービス、拠点ごとにL2網を独立に構築。
- VLAN ID数の制約（最大4,096）により、網を分割して複数構築する場合もあり。
- サービス間、拠点間の接続を行うためには、それらを相互接続する必要がある。
- VLAN IDのマッピング（VLAN ID変換）で実現してきた。

サービス間、拠点間の接続



VLANの相互接続

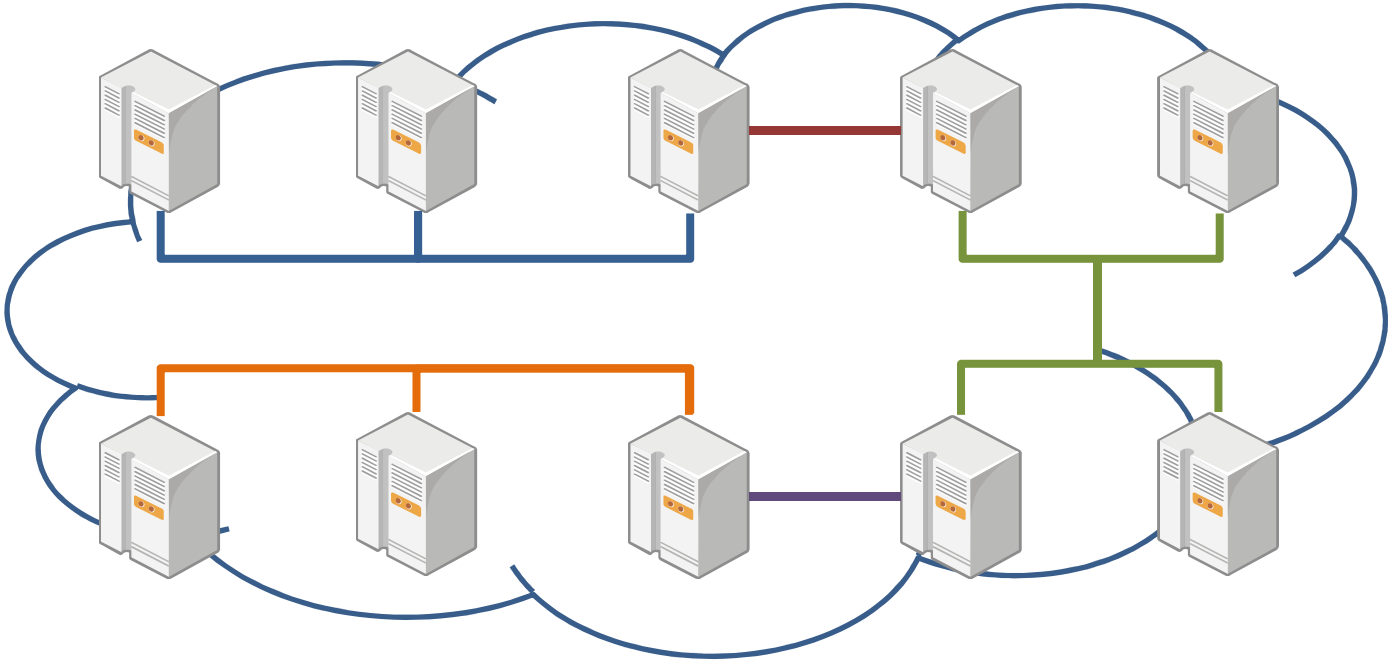

サービスA L2網#1
VLAN100
VLAN200
サービスA L2網#2
VLAN100
VLAN200
各ゾーンのVLANを
任意に自由に接続
VLAN ID変換装置
内部VLAN ID
VLAN100
VLAN101
VLAN102
サービスB L2網#1
VLAN600
VLAN700
サービスB L2網#2
VLAN800
VLAN500

L2スイッチ、VLANを用いた問題点

- VLAN ID数の不足
- MACアドレス数の増大
- フラッディングパケットの増大
- L2セグメントのスケーラビリティ
- リンク帯域の無駄
- コンバージェンス
- トポロジの制約

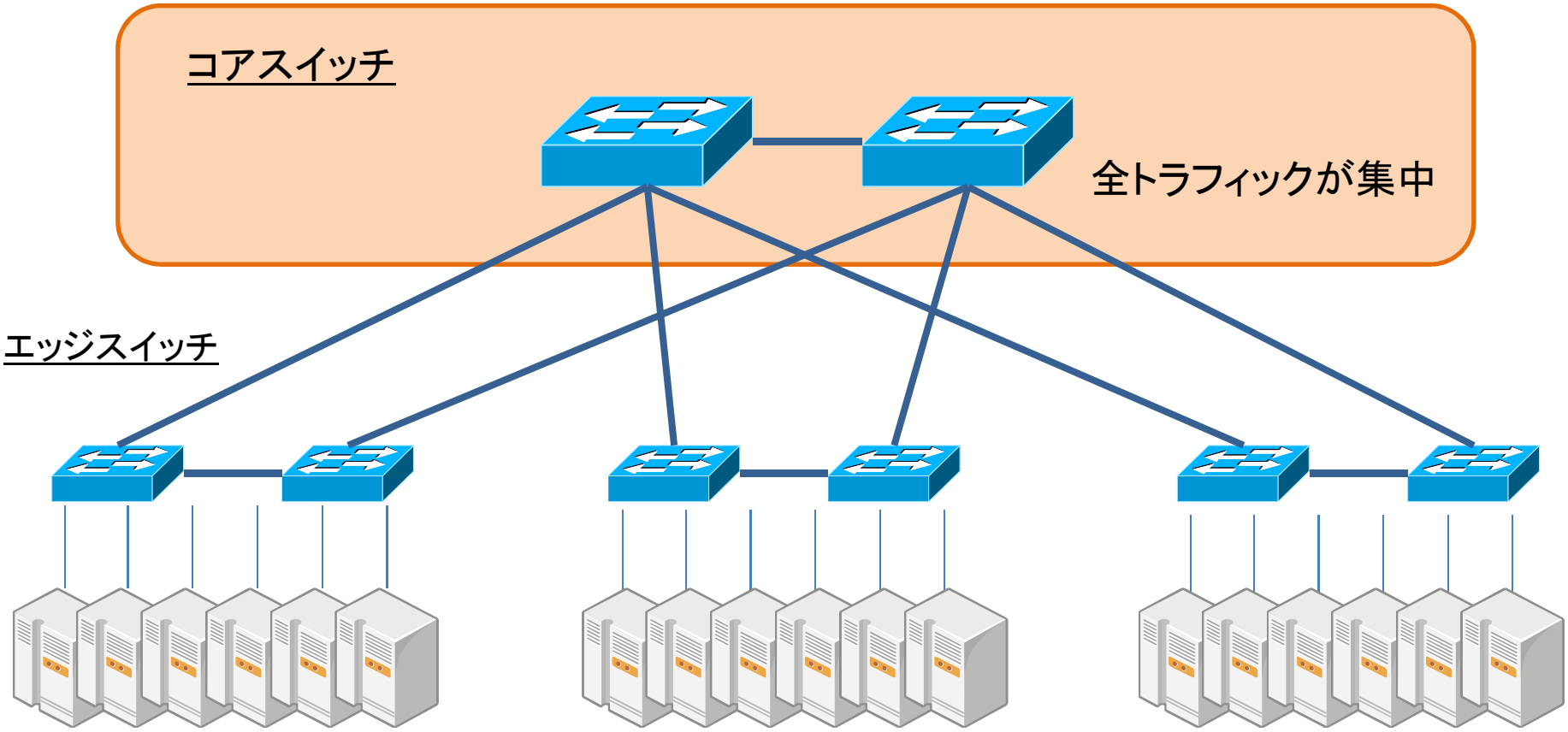
VLAN ID数の不足

- 単一のL2セグメントで、最大4,096VLANまで。
- 10VLAN/ユーザだと400ユーザしか収容できない。
- VLAN IDマッピングは複雑。



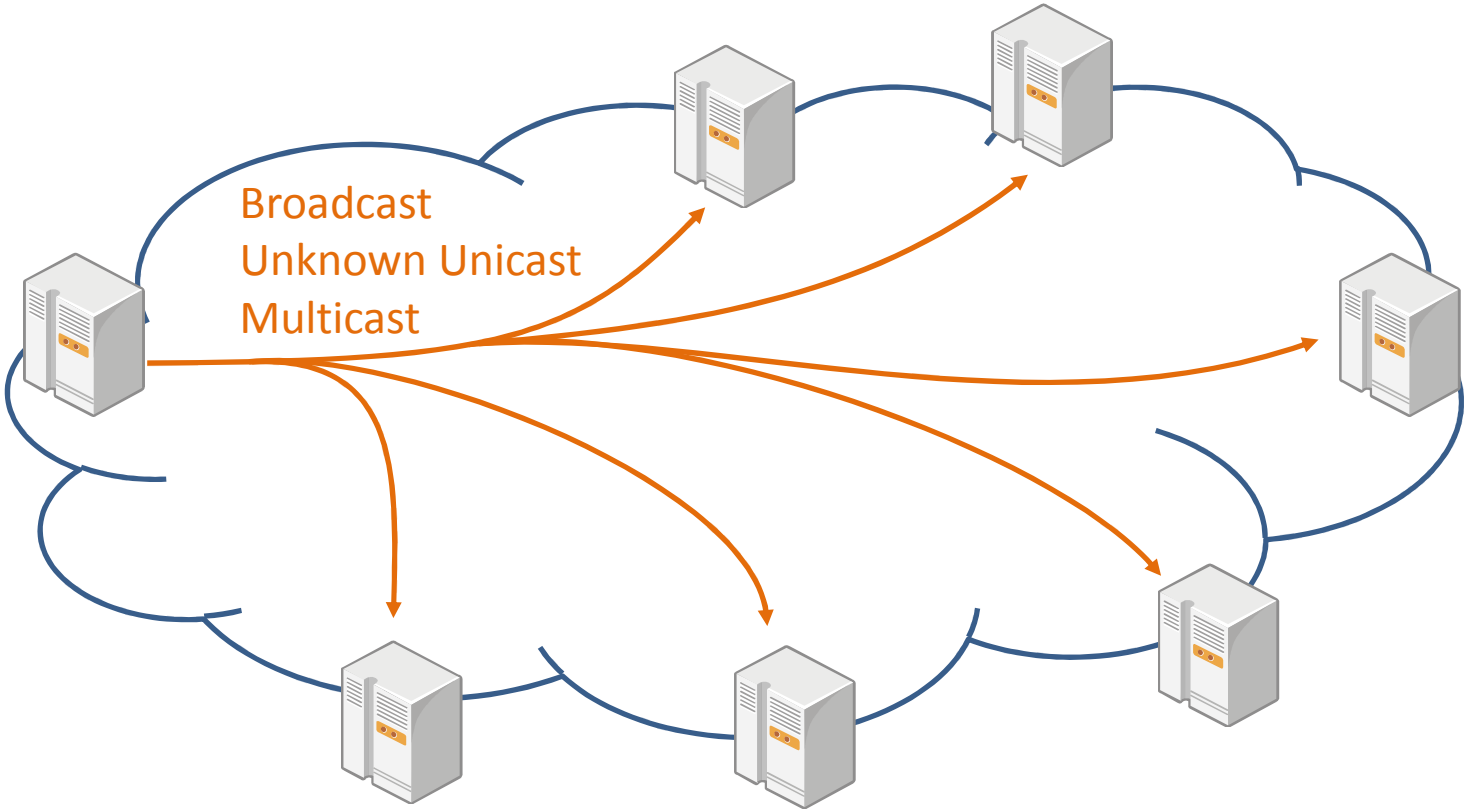
MACアドレス数、スケーラビリティ

コアスイッチでは、網内全てのMACアドレスを学習する必要がある。無限に広げることはできない。



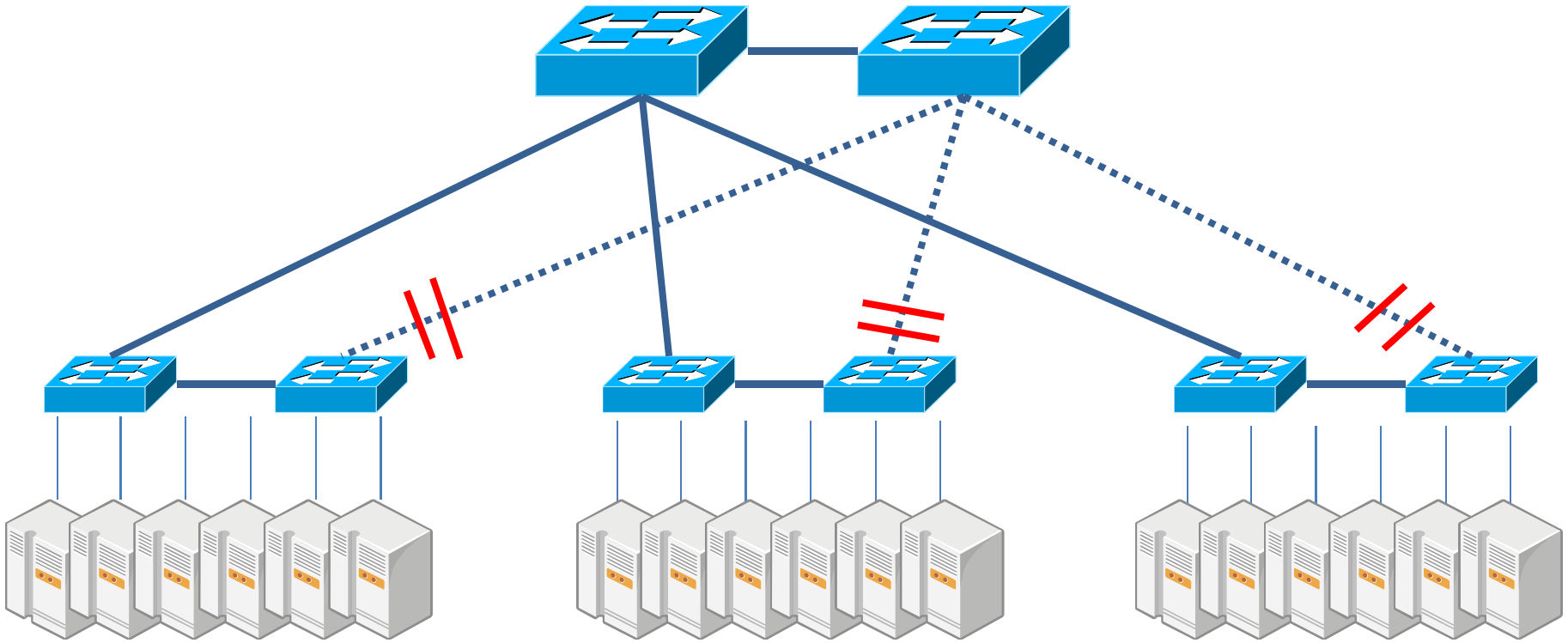
フラッディングパケットの増大

- Broadcast, Unknown Unicast, Multicastパケットは、網全体に配送する必要がある。
- 網側でレートリミットやプロトコルの制限が必要。



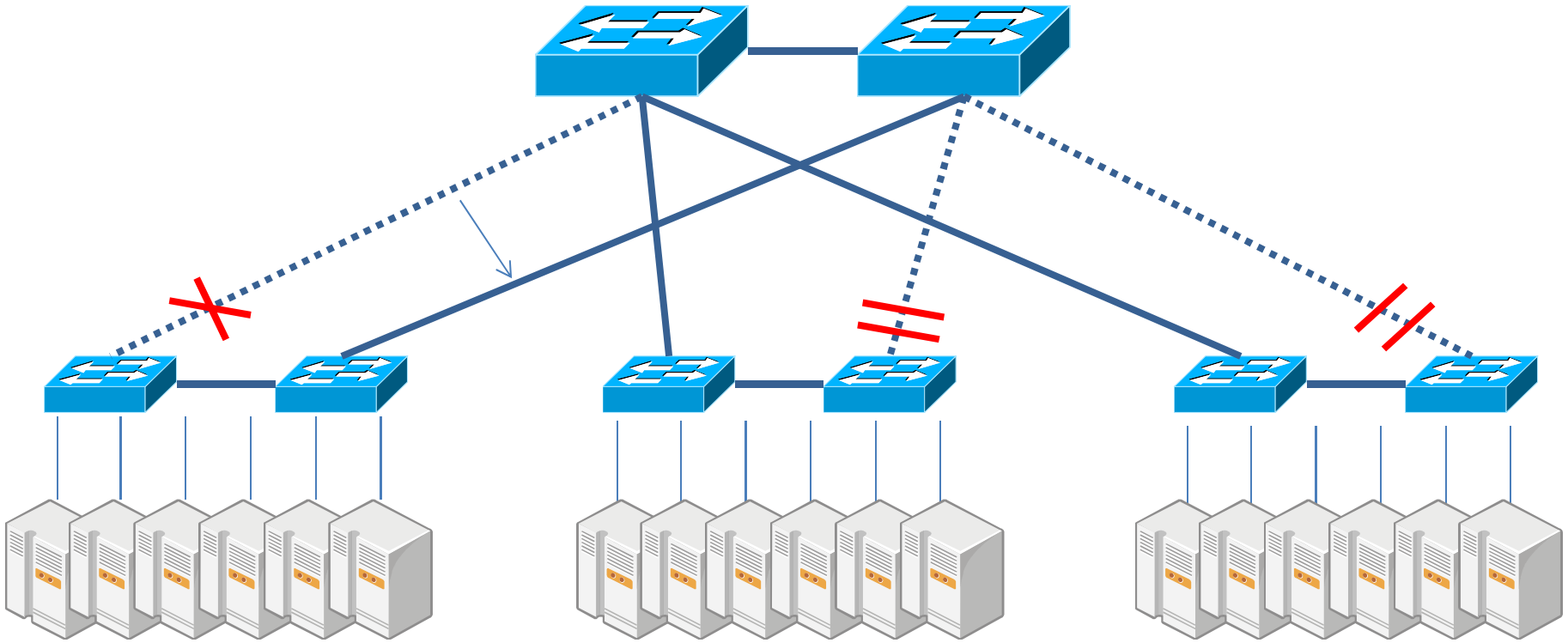
リンク帯域、トポロジの制限

STPを使っている場合、ブロックポートの回線は
利用できない。



コンバージェンス

障害発生時の切替えに時間がかかる。

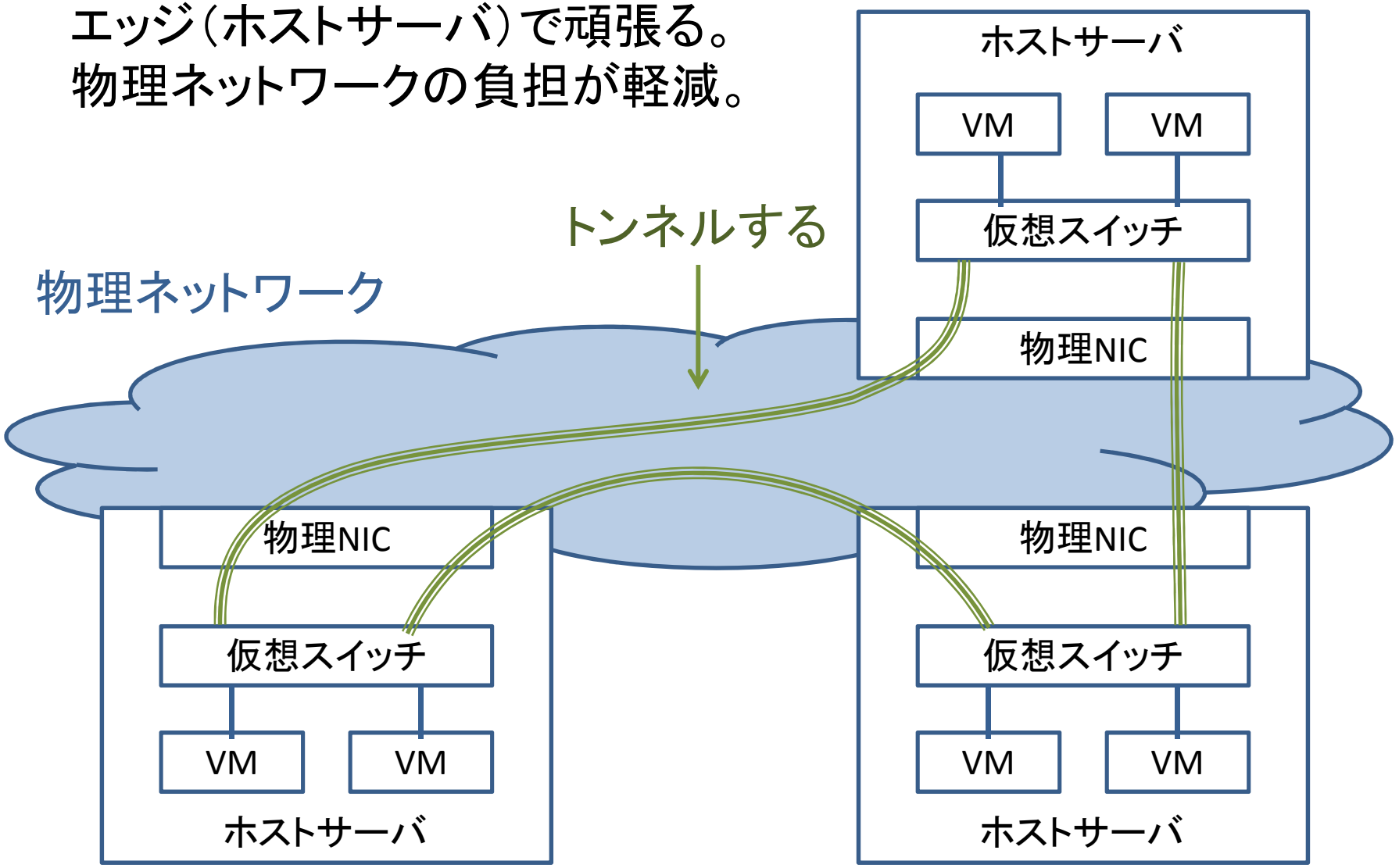


解決策

- オーバレイ方式（トンネル方式）
  - エンドホストにてIPパケットでトンネリング
  - VXLAN、NVGRE、STTなど
  - IETFのワーキンググループが設立
  - Network Virtualization over L3(NVO3)
  - https://datatracker.ietf.org/wg/nvo3/charter/
- L2拡張
  - 既存のL2技術を拡張する
  - Trill, SPB, PBB-EVPN等
- OpenFlow(SDN)

オーバレイ方式のアイディア

エッジ（ホストサーバ）で頑張る。
物理ネットワークの負担が軽減。



解決できそうな課題

- VLAN ID数の不足
- MACアドレス数の増大
- フラッディングパケットの増大
- L2セグメントのスケーラビリティ
- リンク帯域の無駄
- コンバージェンス
- トポロジの制約

各オーバレイ方式の詳細

中本さん、よろしくお願いします。

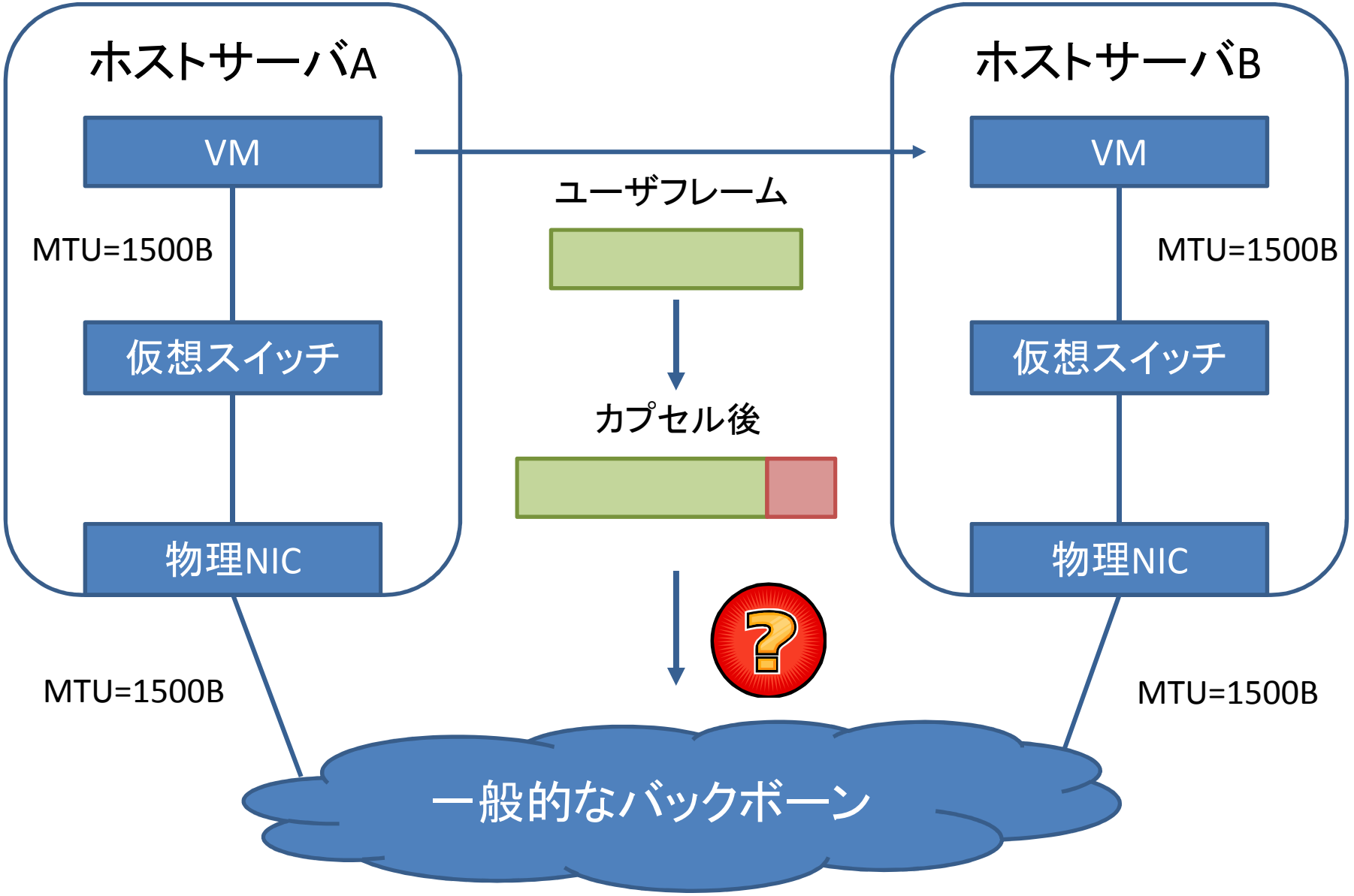
後半資料

オーバレイ方式実装の課題

オーバレイ方式で解決できる問題はあるが、
実装するための課題もある。

- MTUの問題
- ホストサーバの負荷
- バックボーンでのロードバランス
- フラッディング機構
- ゲートウェイ

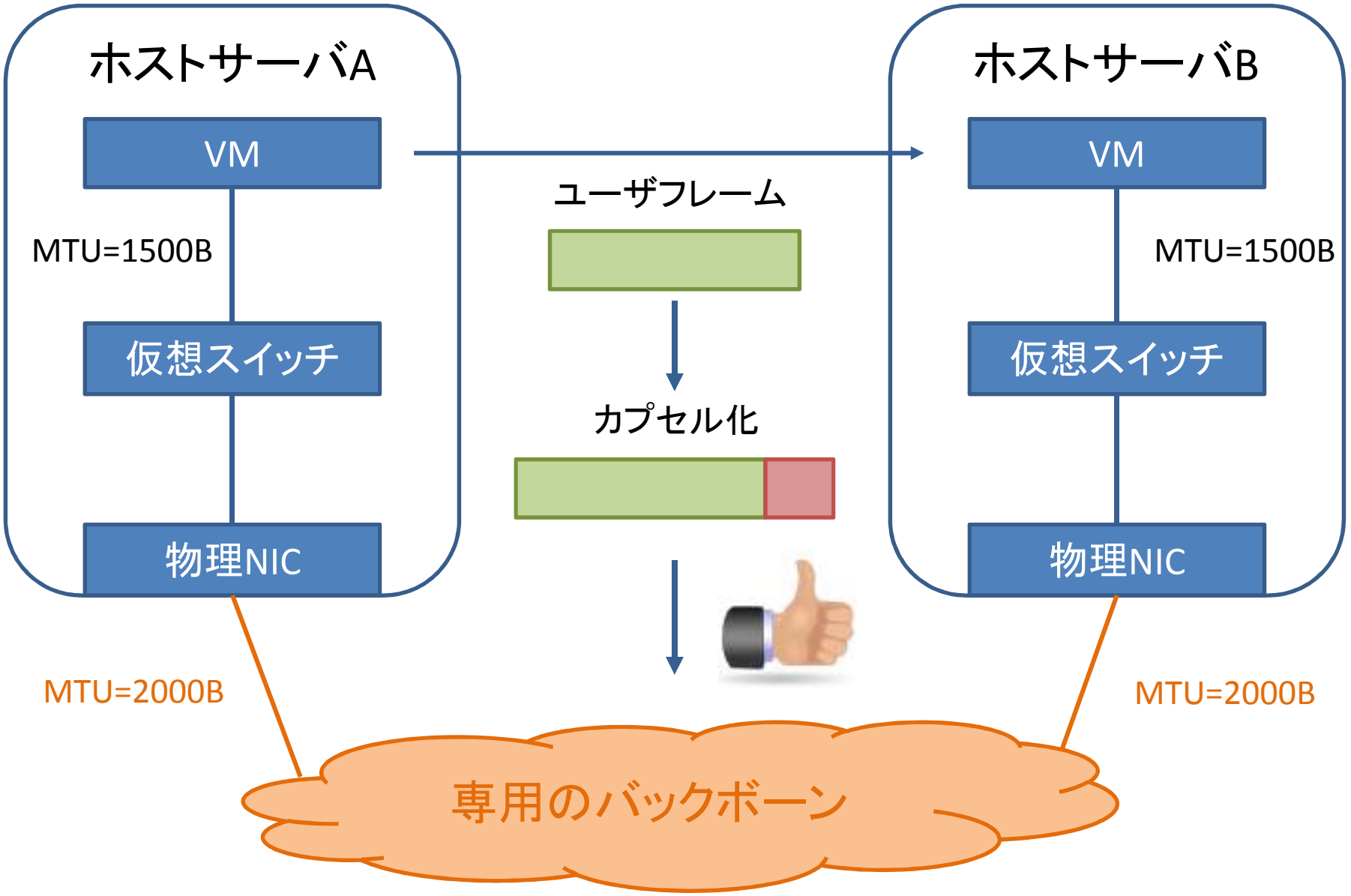
MTUの問題、負荷の問題



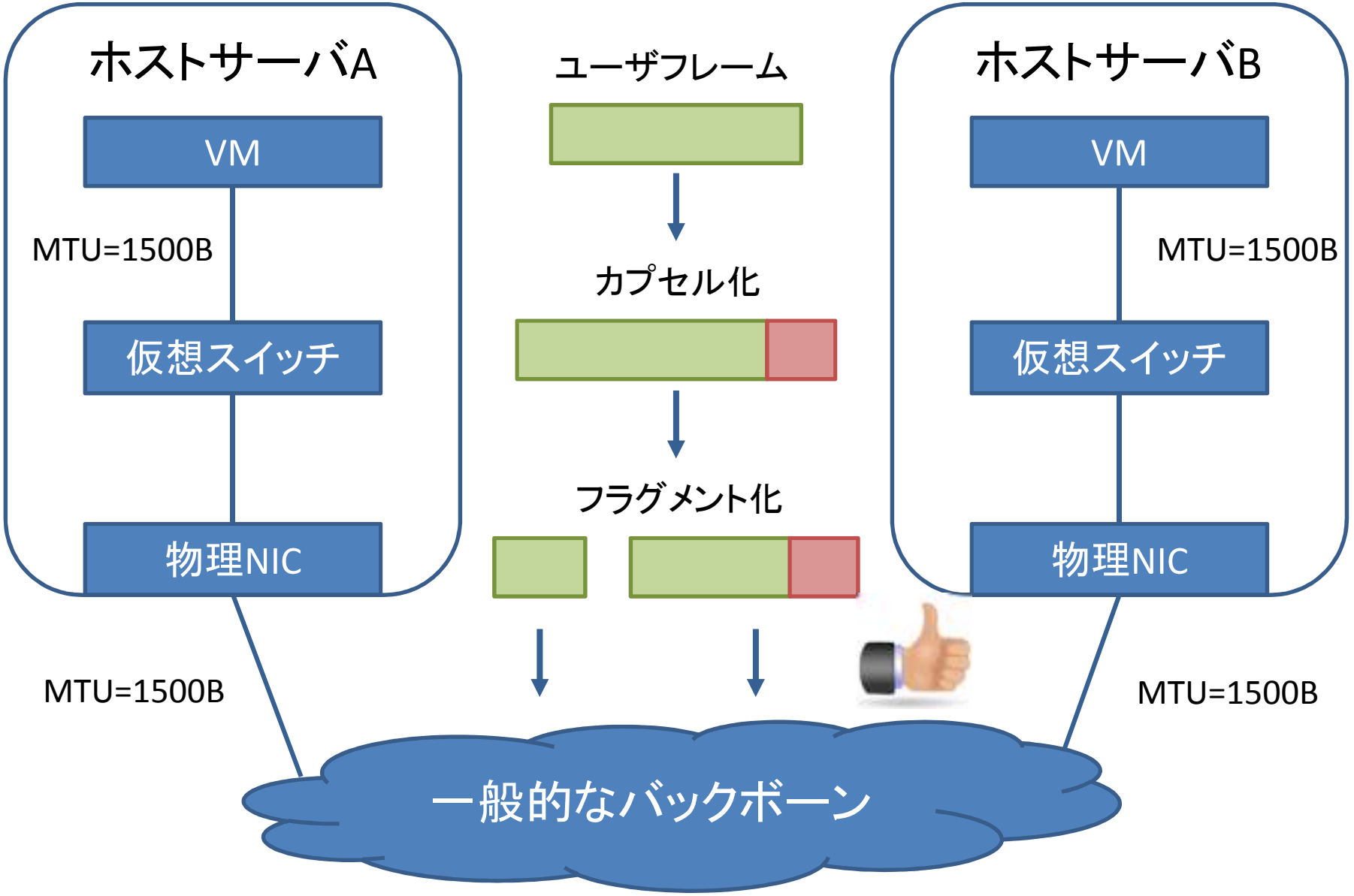
MTUの問題

- 問題点
  - VMに提供しているIP MTUが1500Bに対して、一般的なバックボーンのIP MTUは1500Bしかない。
  - カプセリングを行うとパケット長が増え、バックボーンを通過できない。
- 解決策
  - バックボーンでジャンボフレームを通す
    （MTUを1500B以上に増やす）
  - ホストサーバでフラグメントする
    Path MTUが異なる場合のケア必要

ジャンボフレームによる解決

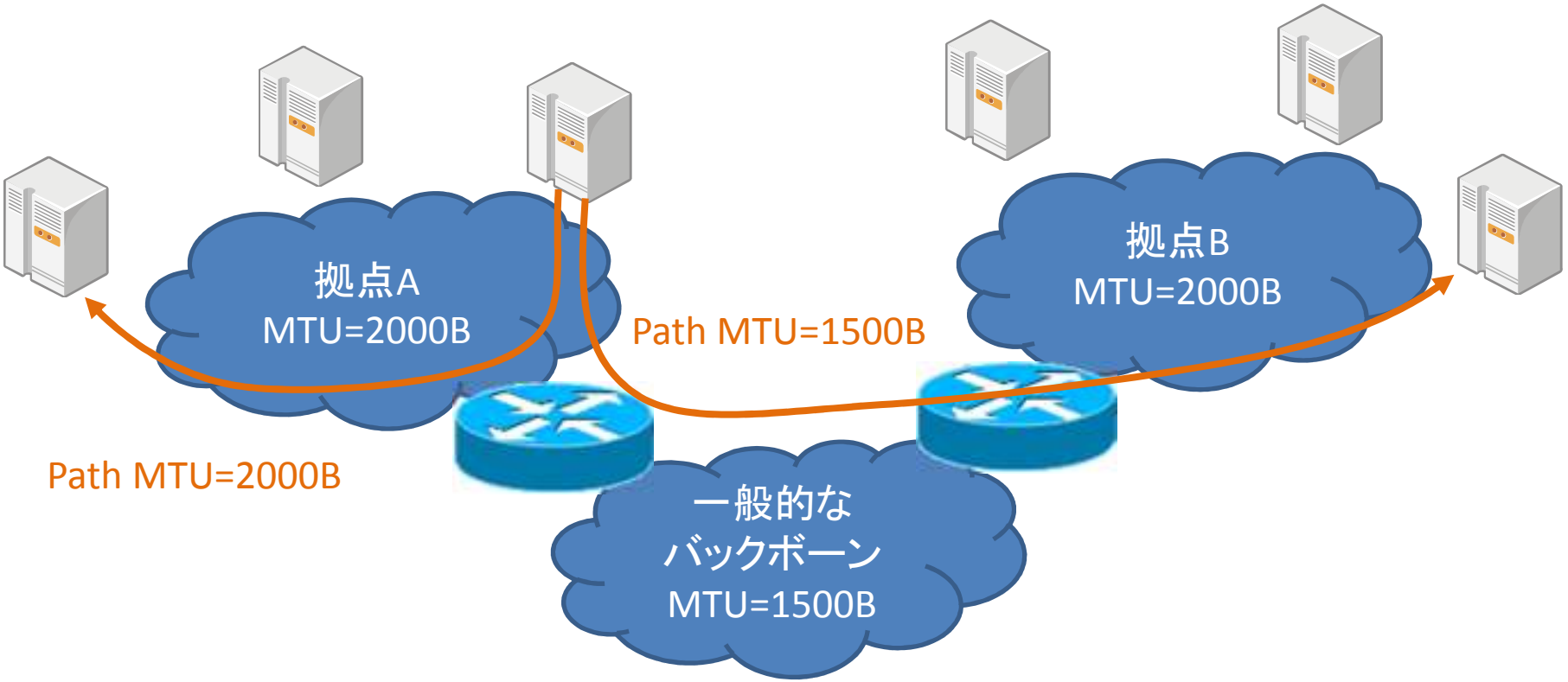


フラグメントによる解決



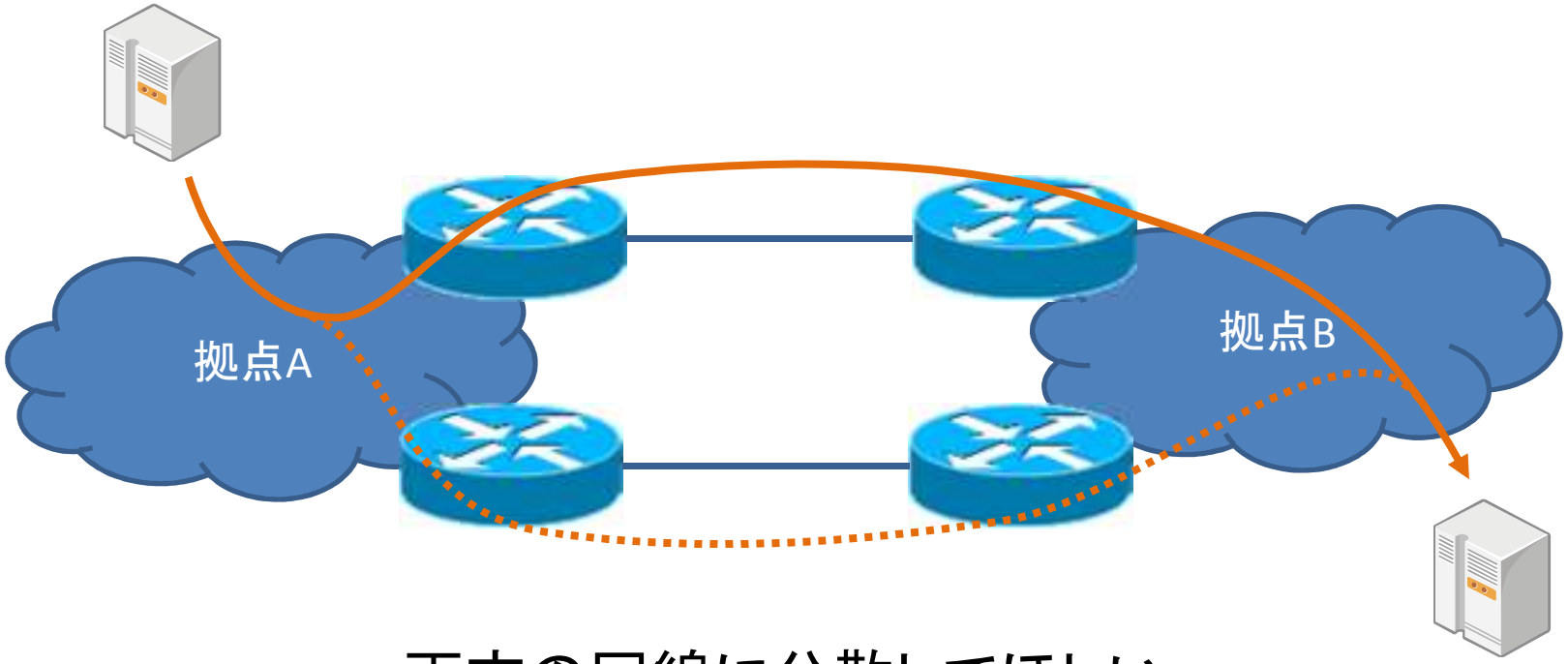
PathMTUが異なる状況

フラグメントで対応する場合、
Path MTU Discoveryが必要な状況もある。



ロードバランスの問題

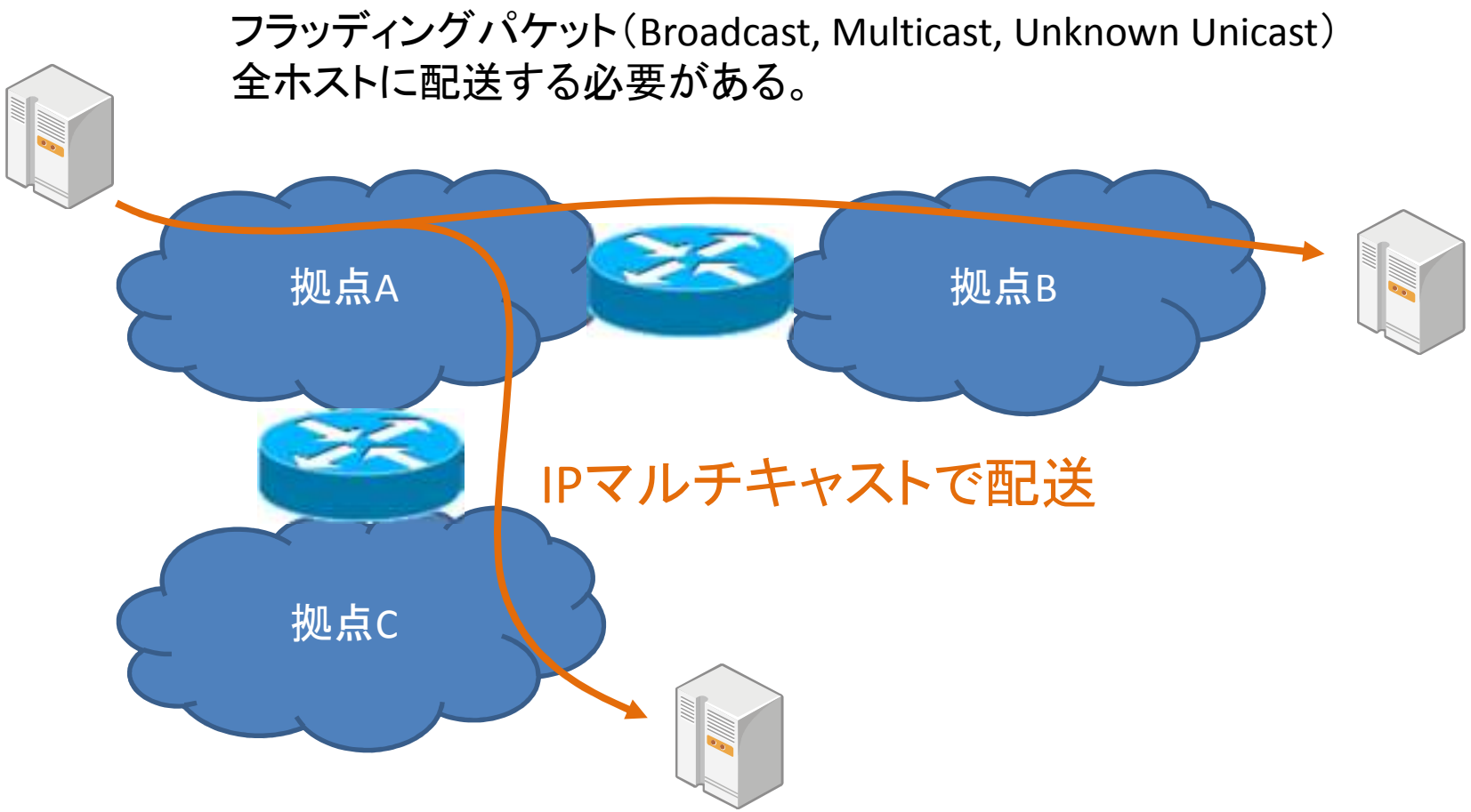
- カプセリングされることにより、内部フローが隠ぺいされる。
- バックボーンでロードバランスできない。



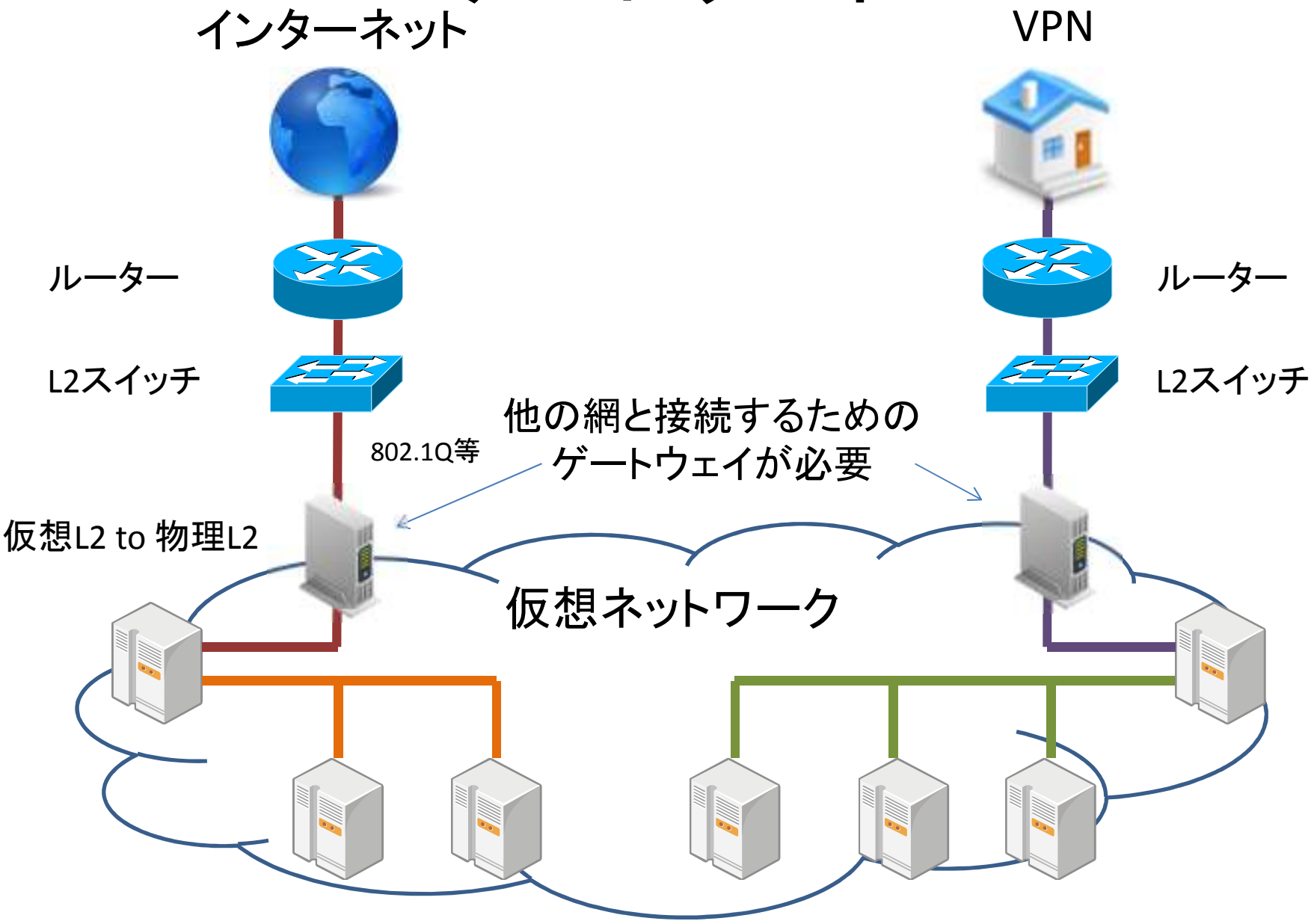
両方の回線に分散してほしい

フラッディング機構の問題

- IPマルチキャストの実装が必要



ゲートウェイ


インターネット
VPN
ルーター
ルーター
L2スイッチ
L2スイッチ
802.1Q等
他の網と接続するための
ゲートウェイが必要
仮想L2 to 物理L2
仮想ネットワーク

参考資料

IPoIBへのNVO3適用案

物理ネットワークはInfiniBand(QDR 40Gbps)で構成。
お客様(VM)からは、変わらずEthernetで見える。