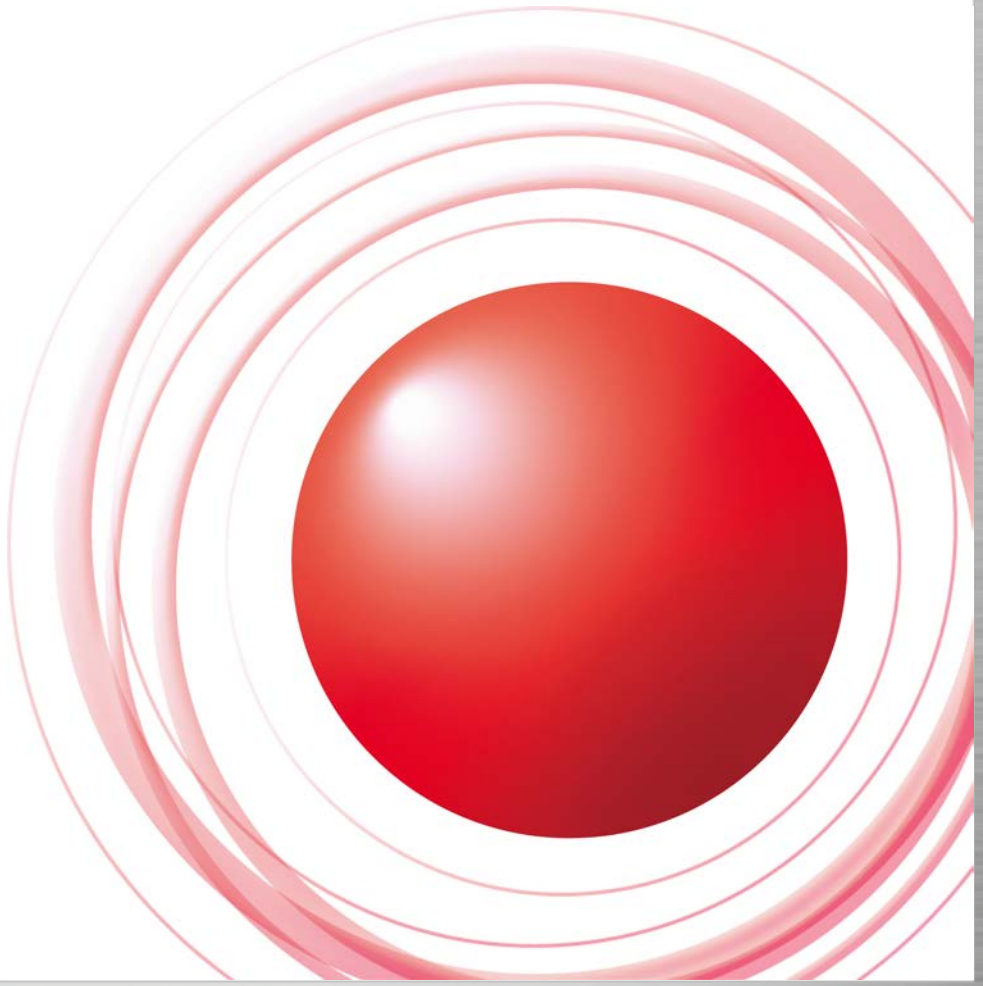
IIJ Technical WEEK 2015
刷新されたIIJ GIOクラウド基盤におけるSDNの実践

IIJ
Internet Initiative Japan



2015/11/11

Ongoing Innovation

IIJ GIOインフラストラクチャ P2

IIJオリジナルのクラウド基盤を全面的にリニューアル
パブリッククラウドとプライベートクラウドの融合
刷新されたネットワーク基盤のお話しします

2015/11/30 Release

P2のパブリッククラウドとプライベートクラウド

性格の異なる二つのリソースプールを併せ持つクラウドサービス

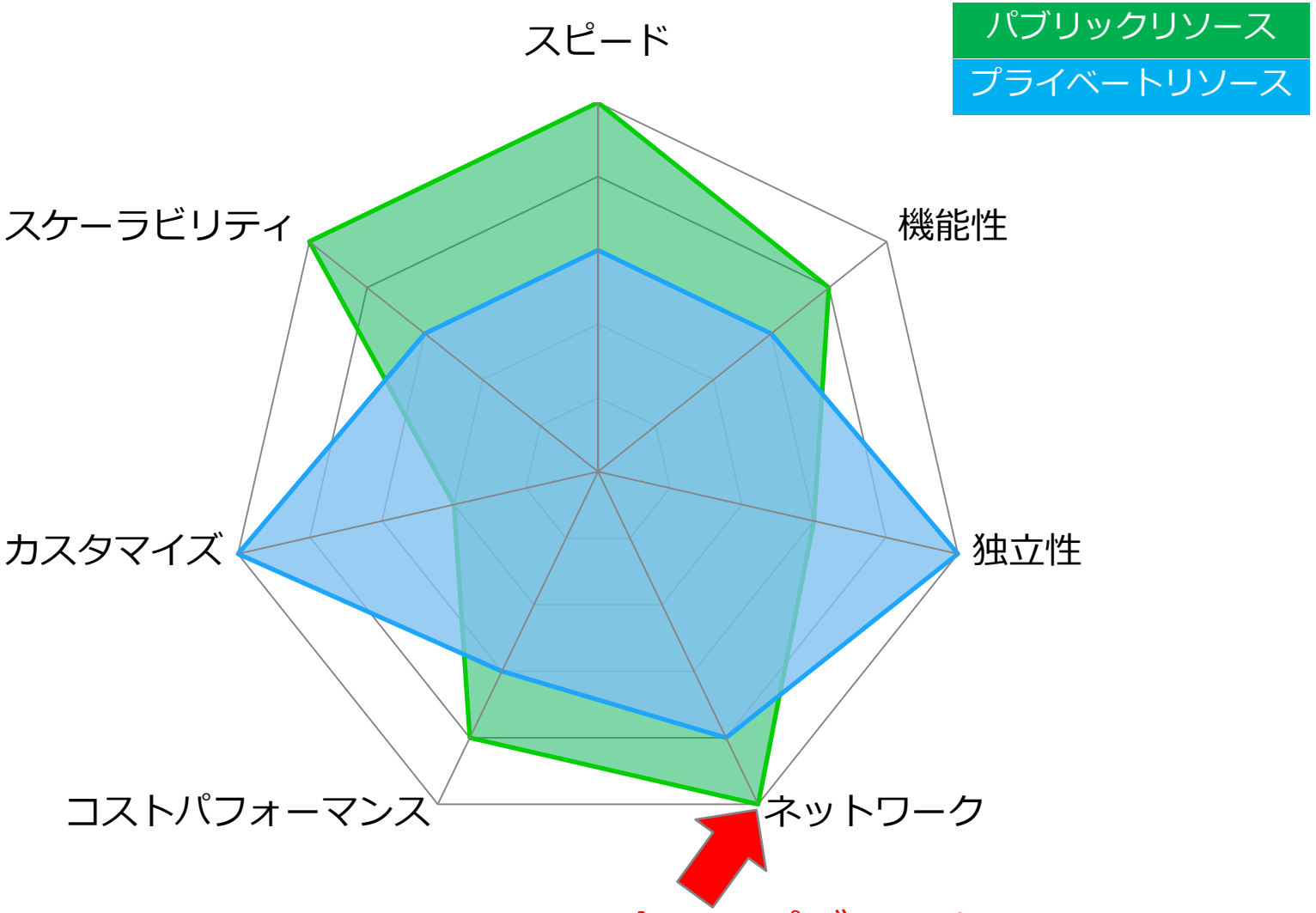
パブリックリソース

均質、高密度
無人、効率最大化
仮想サーバ
仮想アプライアンス
マルチテナント

プライベートリソース

カスタマイズ、個別機器
システム＋人、柔軟性
物理サーバ＋VMware
物理アプライアンス
専有

システムコンセプトの違い



今日はパブリックリソースの
ネットワークについてお話しします

何を重視してサービスをデザインしたかを示したグラフ

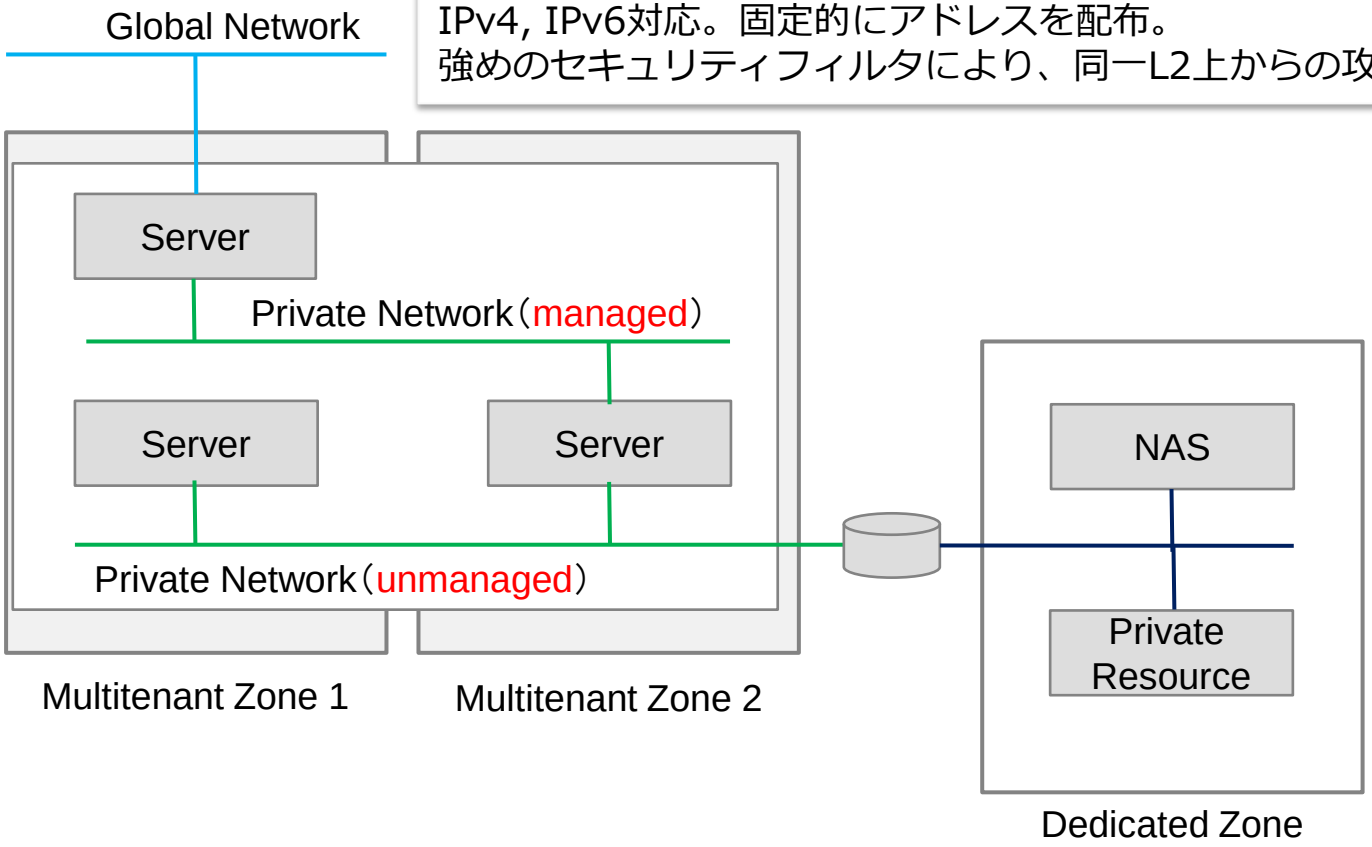
P2 Stack アーキテクチャ

Console
P2 API
Image Archive
Control Panel
API Platform Sakagura
Billing Back Office
IaaS
Image Installer
Orchestrator
Boot Manager
Facility Logic
Metrics Collector
Hypervisor Controller
Storage Controller
Network Controller
Log Aggregator
HyperVisor
Storage API
SSP
watch
Server
Storage
Open vSwitch
HaaS

仮想ネットワーク
管理コンポーネント群

ネットワークタイプ

グローバルネットワークは仮想化はされていない
IPv4, IPv6対応。固定的にアドレスを配布。
強めのセキュリティフィルタにより、同一L2上からの攻撃を遮断



プライベートネットワークは仮想化されている。
IPv4のみ対応
アドレスが配布されるマネージドネットワークと自由に設定できるアンマネージドネットワークを提供。前者には弱いセキュリティフィルタあり

ネットワークを仮想化する切実な理由

- **同一テナントのリソースは近所に配置したい**
  1. トラフィックを最適化し、ネットワークパフォーマンスを引き出すには、同一テナントのサーバ群を同一L2上に配置したい
  2. フローティングアドレス付け替えでフェイルオーバーするには正系、副系が同一L2上に必要
  3. 障害対応や基盤ソフトウェアのアップデート、負荷分散などを目的にライブマイグレーションを実施するには、srcとdestが同一L2上に必要
- **その一方で複数のゾーンに分散もさせたい**
  1. データセンター全体で在庫が十分でも、特定テナントを収容するゾーン（L2）の在庫が不足すれば意味が無い
  2. 特定ゾーン（L2）だけの負荷が上昇しないように、高負荷リソースを分散配置して負荷を平準化したい
- **同一テナントのリソースを同一L2に収容したくもあり、分散させたくもある**

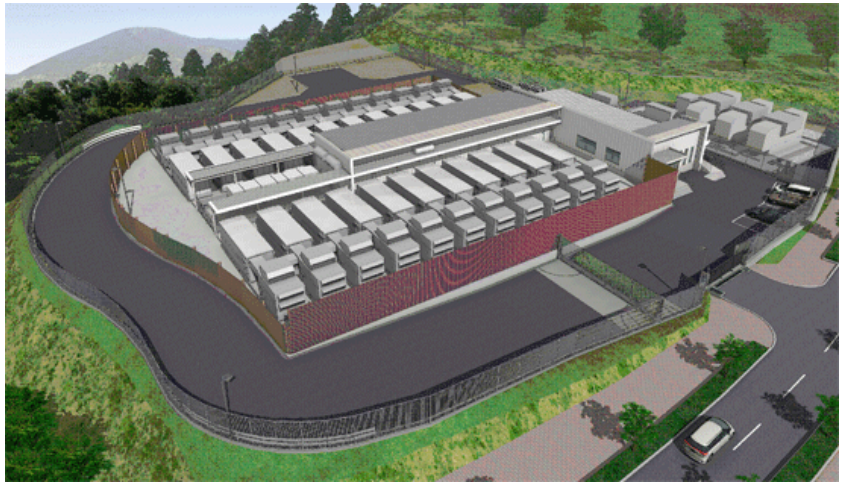
アドレス管理とクラウドネットワーク

- **L2にこだわらず、L3で広げる戦略をとるとどうなる？**
  - 同一テナントのリソースでも積極的に複数ゾーンへ配置し、L3でつながる広大なリソースでひとつのリソースプールを構成する
  - 同一L2上に配置が必須のリソースだけ特別扱いとする
  - リソースプールの拡大は可能になる
- **アドレスがリソースのポータビリティを縛る**
  - サーバにアドレスをアサインすると、そのサーバは特定のゾーンに縛り付けられる
  - 静的にアドレスをアサインすると、在庫や負荷の平準化は実現できない
  - 障害やメンテナンスのために、別ゾーンのハイパーバイザへVMを移動させることもできない
  - つまり、最初に仮想サーバを生成するときはどこにあってもよいが、その後移動できなくなるということ
- **クラウド外とのコネクティビティを阻害**
  - ユーザーが自由にプライベートアドレスを設定できないことも意味する
  - クラウド上のリソースとオンプレシステムを接続するには、一手間かかり使い勝手が悪くなる

クラウドネットワークの方法論

1. **L2ネットワークをできるだけ広くとる**
   - MACアドレス数かVLAN数の上限はあるものの、物理サーバで1000台程度の規模ならばこれが一番よい
   - クラウドベンダとしての競争力は乏しい
2. **アドレス管理に制約を設ける**
   - アドレスを動的にアサインする。グローバルアドレスを直接サーバにアサインしない
   - どのゾーンに収容してもよいことになるので楽。ただし利便性を考えると難あり
3. **ネットワークを仮想化（あるいはその類似の技術）**
   - 数万台規模でクラウド基盤を構築する手段はこれしか残らない
   - ただし、グローバルネットワークはテナント固有のL2が不要であることと、IPv6を提供するために仮想化されていない

コンテナデータセンターと仮想ネットワークの相性



相性は抜群です

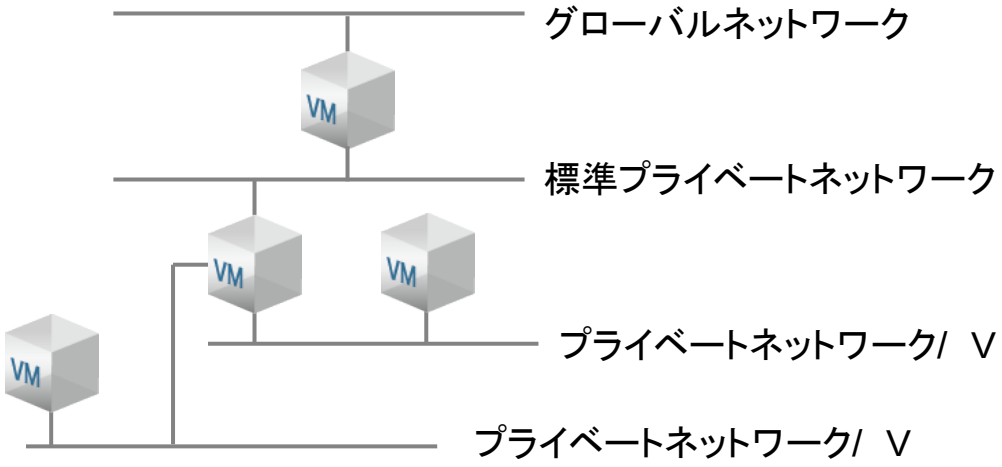
仮想ネットワークで生きるコンテナデータセンター

- **コンテナのメリット**
  - 必要なタイミングでモジュールを追加することで、段階的な投資が可能
  - その時代における最新技術をモジュールに反映することで設備の陳腐化を防ぐ
  - 標準化されたモジュールを用いることで、迅速な構築が可能
  - 省スペース、高効率空調による省エネ化
- **コンテナのデメリット**
  - L2のサイズが小さく、物理的なL2の延伸が困難
  - 省スペースゆえ、人的DC作業に制限
- **ネットワークを仮想化すればメリットを生かしつつ、デメリットがデメリットでなくなる**
  - 物理的にはコンテナ間をL3で接続していくだけでよい

GIO P2 VMに接続されているネットワーク

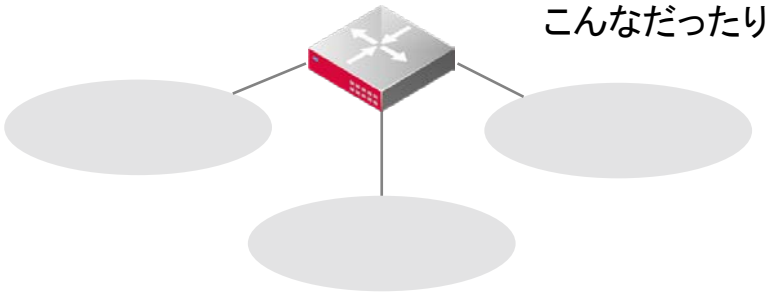
- **グローバルネットワーク ..... ひとつ**
- **標準プライベートネットワーク ..... ひとつ**
- **プライベートネットワーク/V 最大5つ**
  - サブネットを自由に作成できる

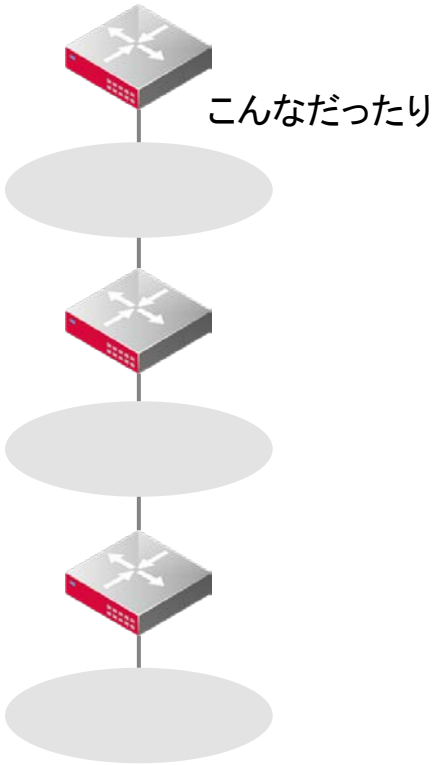
- **以降ではプライベートネットワーク/Vの実装について説明します**
  - 実装上は標準プライベートとプライベートネットワーク/Vは同じ仕組み（フィルタの内容が異なる）



理想的なプライベートネットワーク

- いくつでも作れる
- 自由にネットワークアドレスを設定できる
- 自由に構成できる

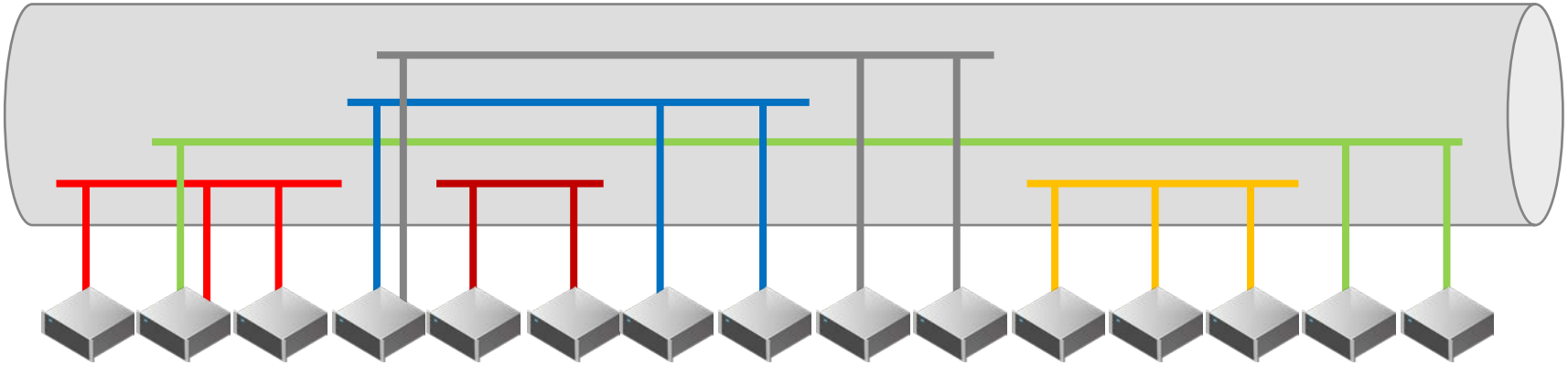




- 自由にVMを配置できる
  - 足の数も自由
- 物理ネットワーク実装の制約を受けない！

自由なプライベートネットワークを実現するためには

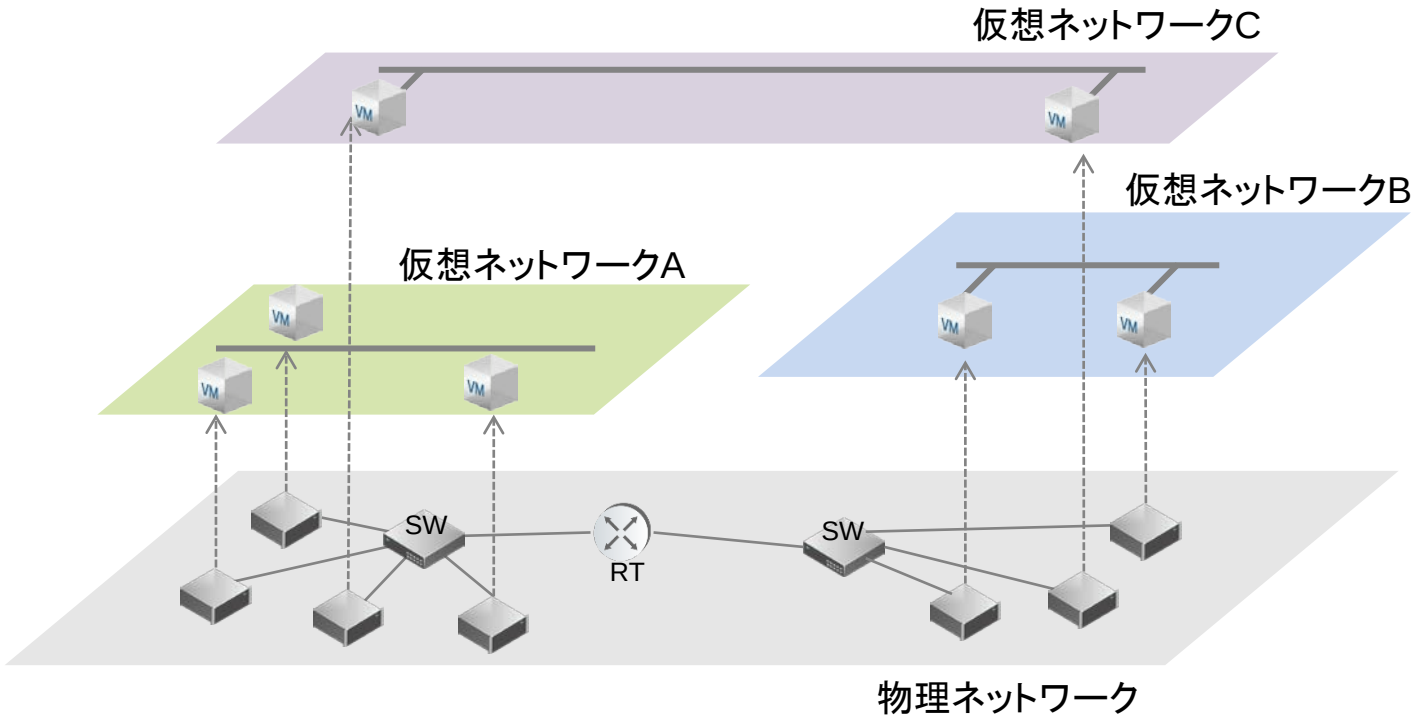
- ひとつのL2ネットワークをVLANで分割する



- 数千台の物理サーバを単一のネットワークに接続するのは無理
  - VLAN 4094 の壁
  - MAC学習の壁
  - ブロードキャストの壁

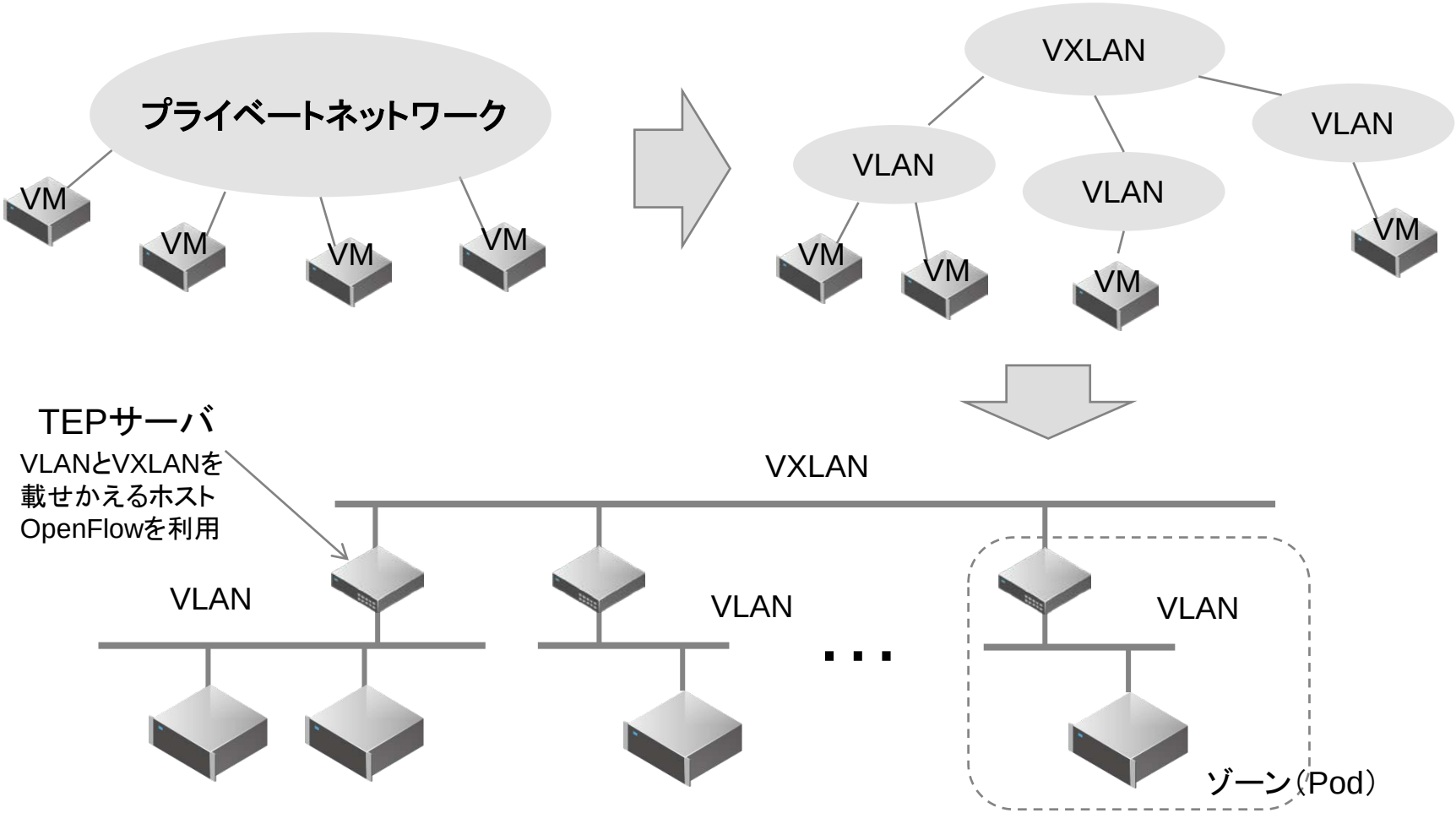
オーバレイネットワークで解決

- L3ネットワーク上にL2VPNを張って仮想L2ネットワークを構築する



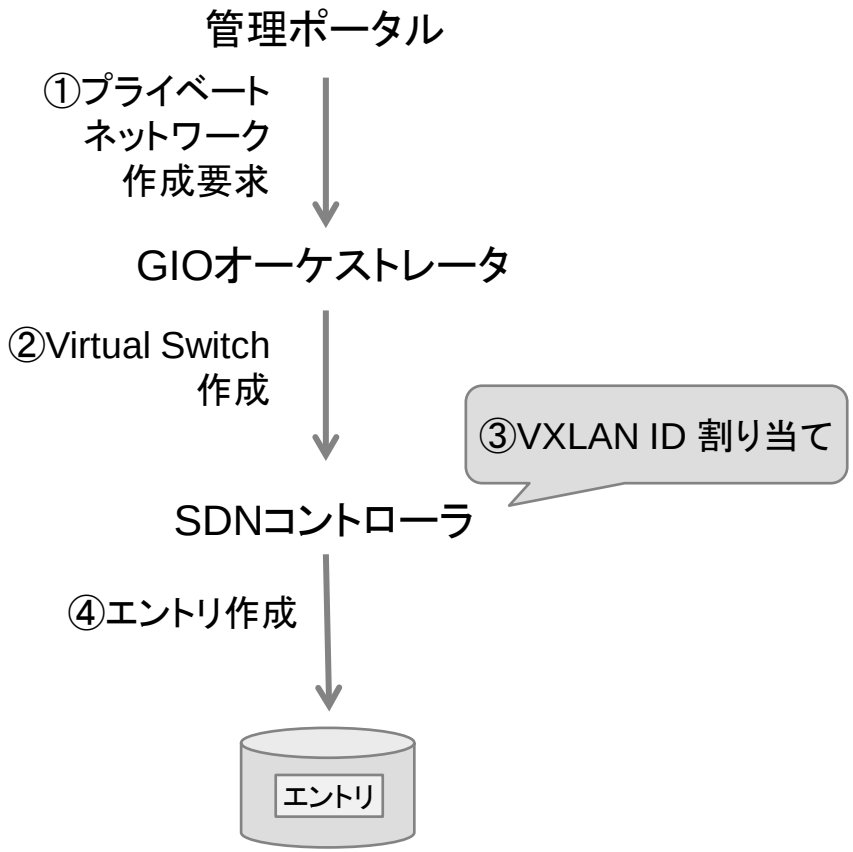
GIO P2のオーバレイネットワーク

- **VLANとVXLANを組み合わせて構築する**



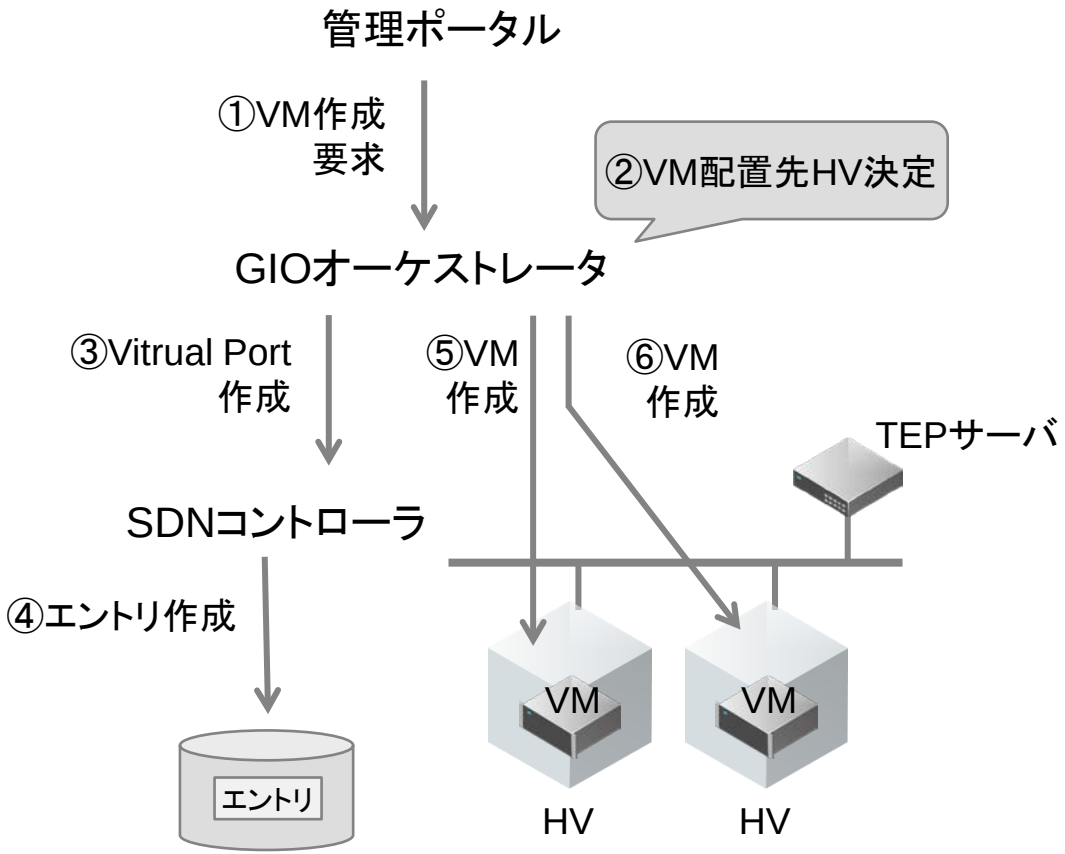
VMとTEPの配置シーケンス１

- プライベートネットワーク新規作成



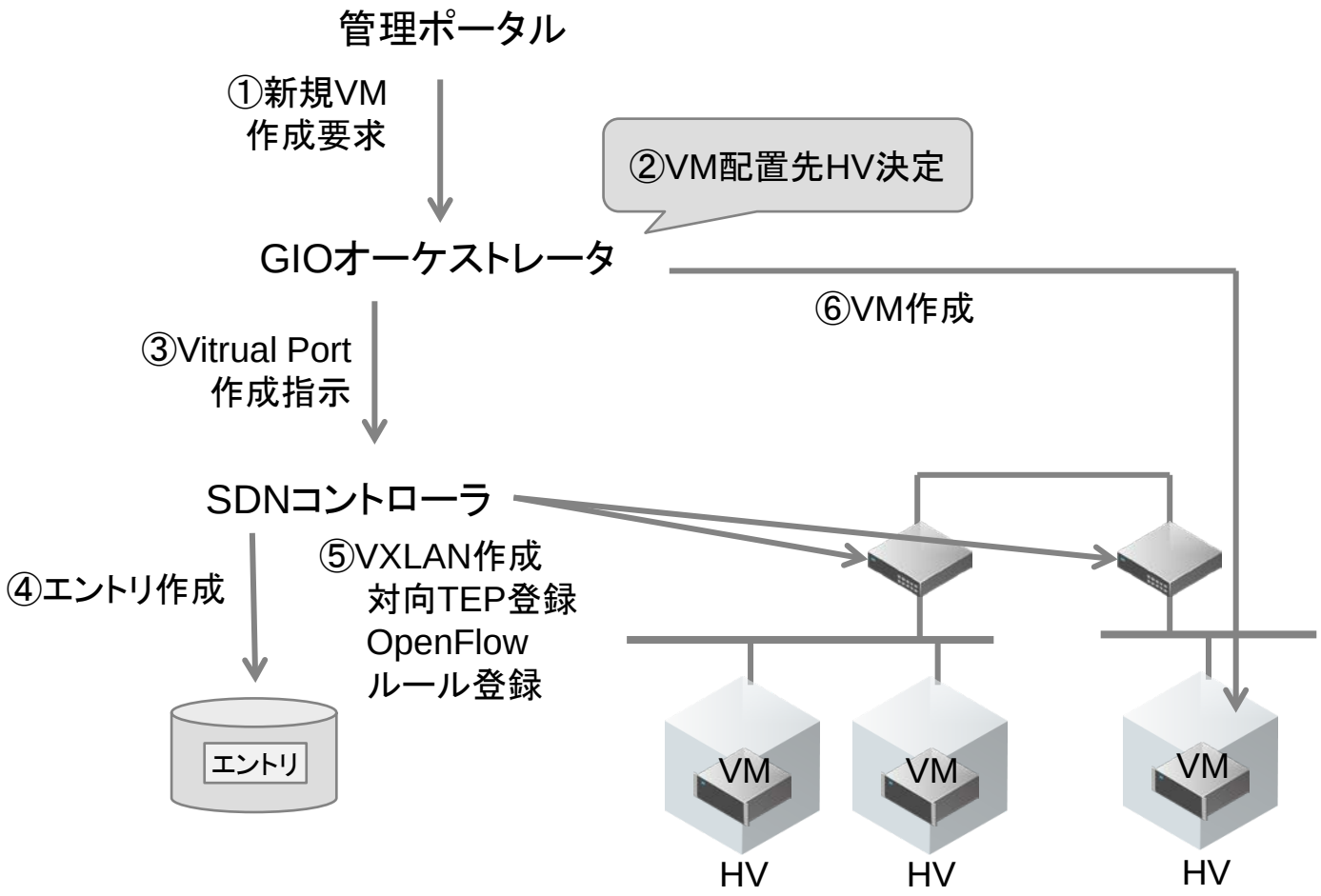
VMとTEPの配置シーケンス２

- **VM作成（おなじゾーンに配置されるパターン）**



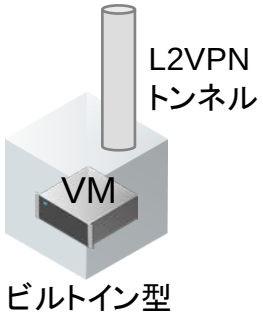
VMとTEPの配置シーケンス３

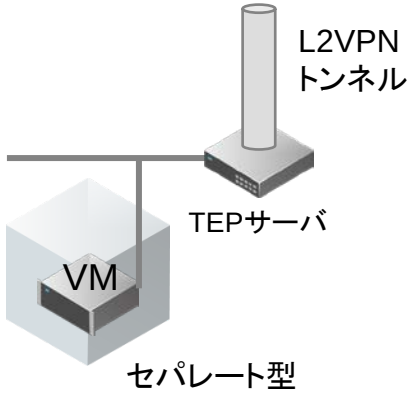
- **VM作成（異なるゾーンに配置されるパターン）**



VLANを使う理由

- **総トンネル数を減らす工夫**
- **VXLANの構成パターン**
  - ビルトイン型
    - HVにTEPがある
  - セパレート型
    - HVとTEPサーバが分離している



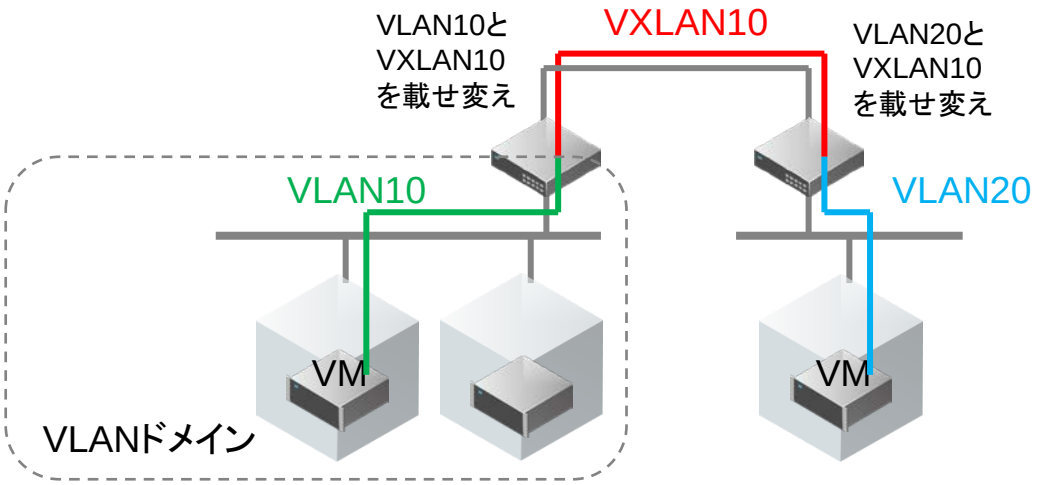


- **VXLAN：知らない宛先へのパケット**
  - マルチキャスト
  - Head End Replication
- **TEPが増えると**
  - （コピーされた）ブロードキャストパケットが倍増
  - 各ノードのTEP更新コストが無視できない
- **セパレート型を採用**
  - 「ゾーン」をひとつのVLANネットワークとする
  - （HVを複雑化したくなかった）

VLANの問題はどうした？

- **4094の壁**
  - ゾーン毎にVLAN ID空間を分ける
  - VXLANとVLANとの間でID変換を行う
  - VM配置を工夫すれば4094の壁はない
- **MAC学習の壁**
  - １ゾーンの物理サーバは480台なので問題なし
- **ブロードキャストの壁**
  - 次項で説明

VXLANとVLANのID変換の例

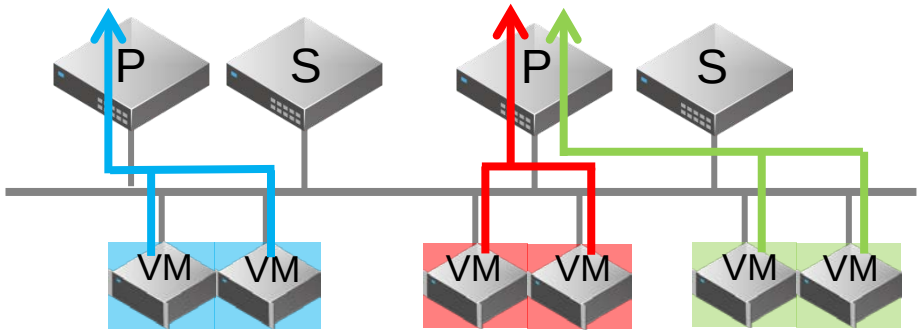


オーバレイってお高くつくんでしょ？

- **ブロードキャストがコピーされて倍増するんでしょ？**
  - プライベートネットワークを１ゾーン内に収めるようにVMを配置
  - ブロードキャストを広めない
  - ブロードキャストARPをユニキャストARPに書き換え
- **ヘッダが挿入されてフラグメント化するんでしょ？**
  - GIO設備内はジャンボフレーム対応で統一した

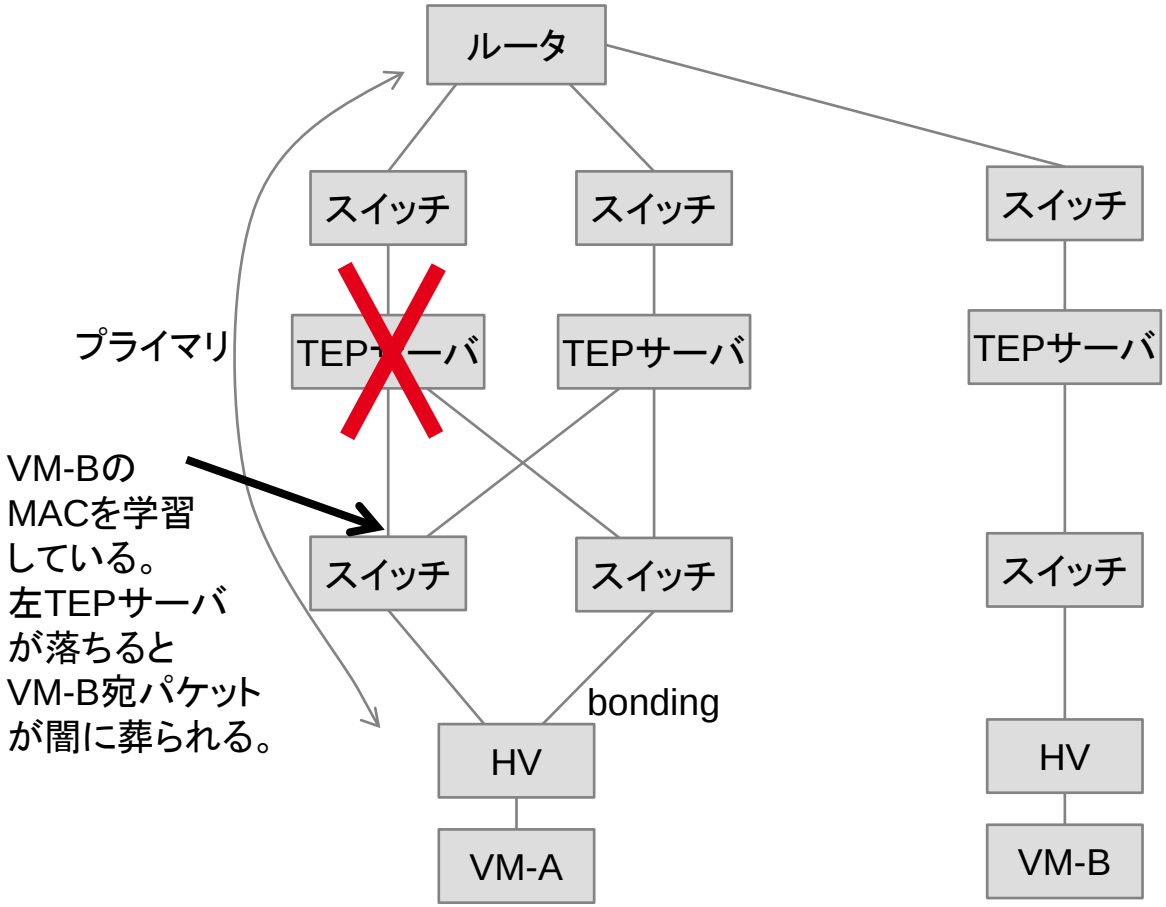
TEPサーバの冗長化とスケールアウト

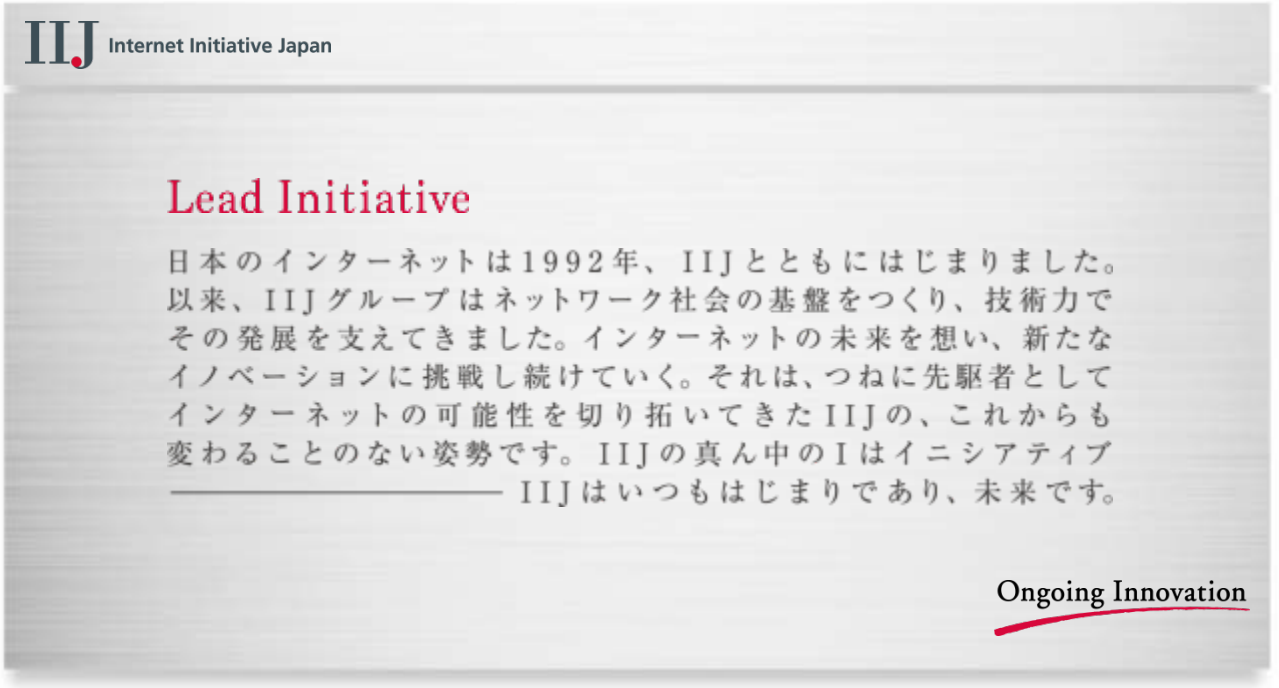
- **TEPサーバはプライマリ／スタンバイ構成**
- **TEPサーバの処理能力が足りなくなった場合はスケールアウト**
  - 吸い込むVLAN IDが異なる



レガシーネットワークとSDNを接続する注意点

- **スイッチのMAC学習テーブルを更新する必要がある**
  - HSRPのGratuitous ARP送信と同じ





お問い合わせ先　IIJインフォメーションセンター
TEL：03-5205-4466 （9：30～17：30 土/日/祝日除く）
info@iij.ad.jp
http://www.iij.ad.jp/