\Orchestrating a brighter world NEC

「クラウドNWの展望 -5Gとの融合-」
-JANOG44のその後-

2020年1月24日

中原一彦
日本電気株式会社

大橋憲昭　山本幸治
NECネッツエスアイ株式会社

目次

- はじめに
- クラウドNWの展望
- Janog43と44のおさらい – HaaSの実現 -
- 仮想と物理の融合
- ディスカッション

はじめに

はじめに

今クラウド基盤に必要な構成は何か？

マルチクラウド・５Ｇ・ベアメタルクラウドという言葉から、IaaSテナントの要求を実現する基盤がどういったものかを検討してきた

JANOG44では

「クラウド基盤の仮想化NWを利用した構築自動化」と題して、HaaSレイヤの仮想NWの現実解としてEVPN/VxLANを採用することで、物理サーバのプール化を実現した

今回ついにIaaSと物理レイヤの融合を果たす

IaaSテナントの要求に応じて仮想サーバ・VNFを提供したうえで、物理機器もプールからの提供を行う

→ クラウドNWの展望の考察と今回構成したクラウドNW構成がマッチするのか議論を行いたい

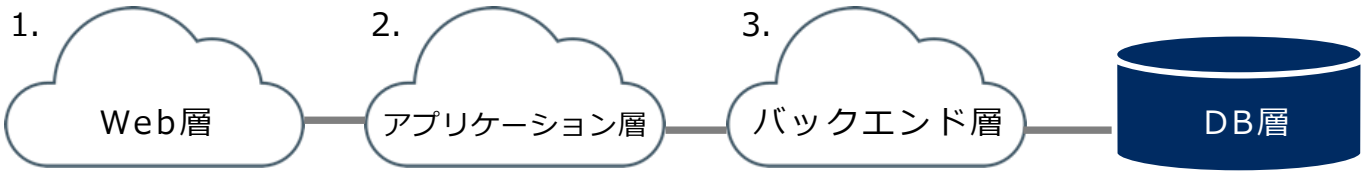
クラウドNWの展望

クラウドNWの展望　５Gが入るとどうなるか考察してみた

直近の5Gが商用スタート！
では5Gが入ると
クラウド基盤はどうあるべきか？

クラウドNWの展望　５Gが入るとどうなるか考察してみた

まず「IaaSテナントがどんな構成になるか」
について、それぞれ考えてみた



ポイント：5Gでデータ量が膨大になる！

クラウドNWの展望　５Gが入るとどうなるか考察してみた

1. Web層

パブリッククラウドの利用により、Web層の規模縮小・拡大を容易に可能にする

パブリッククラウドの例）

マルチクラウド化が必要

検討中。。。

クラウドNWの展望　５Gが入るとどうなるか考察してみた

2. アプリケーション層

VMが一般化したが、コンテナの導入を進めていかなければならない

コンテナ管理プラットフォームの導入

- kubernetesをはじめとする複数のコンテナ管理を容易にできるプラットフォームの登場により、コンテナ技術のメリットをより強化

Good Point

開発面：ソフトウェア開発、サービスリリースのスピードUP!

運用面：環境起因のシステム障害発生のリスクを軽減

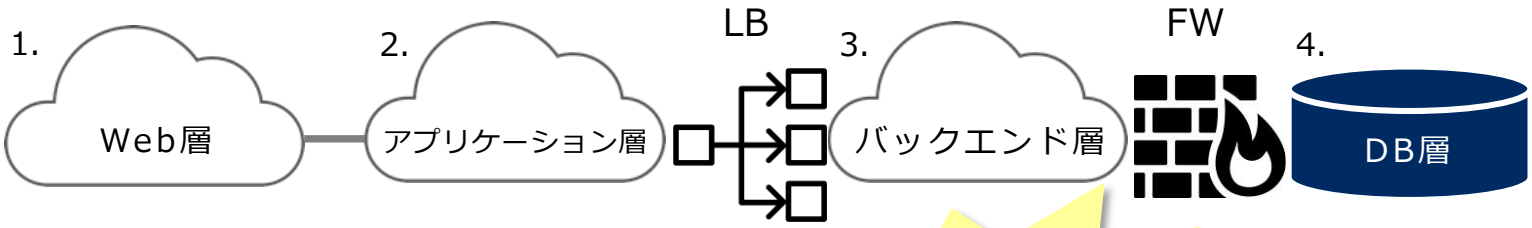
仮想サーバ基盤の多様化の考慮が必要

今回は未着手。。。

クラウドNWの展望　５Gが入るとどうなるか考察してみた

3. バックエンド層

5Gで収集した大量のデータを扱うためには、高い処理速度や能力が必要不可欠



大容量のデータの
プレ・ポスト処理が発生

処理速度・能力を考慮し、
ベアメタルを導入する

アクセラレータの限界
DPDK、SR-IOVでは結局は物理構成（NUMA）の考慮が必要

柔軟性は仮想サーバが圧倒的だが、性能面では物理には及ばない

↓

物理と仮想の融合が必須！！

コンセプト

JANOG43 クラウド基盤の仮想化NWの設計と課題より

HaaSレイヤを定義し、オンデマンドにリソースの提供を実現
ー IaaSレイヤでは物理リソースの受給管理から解放 ー

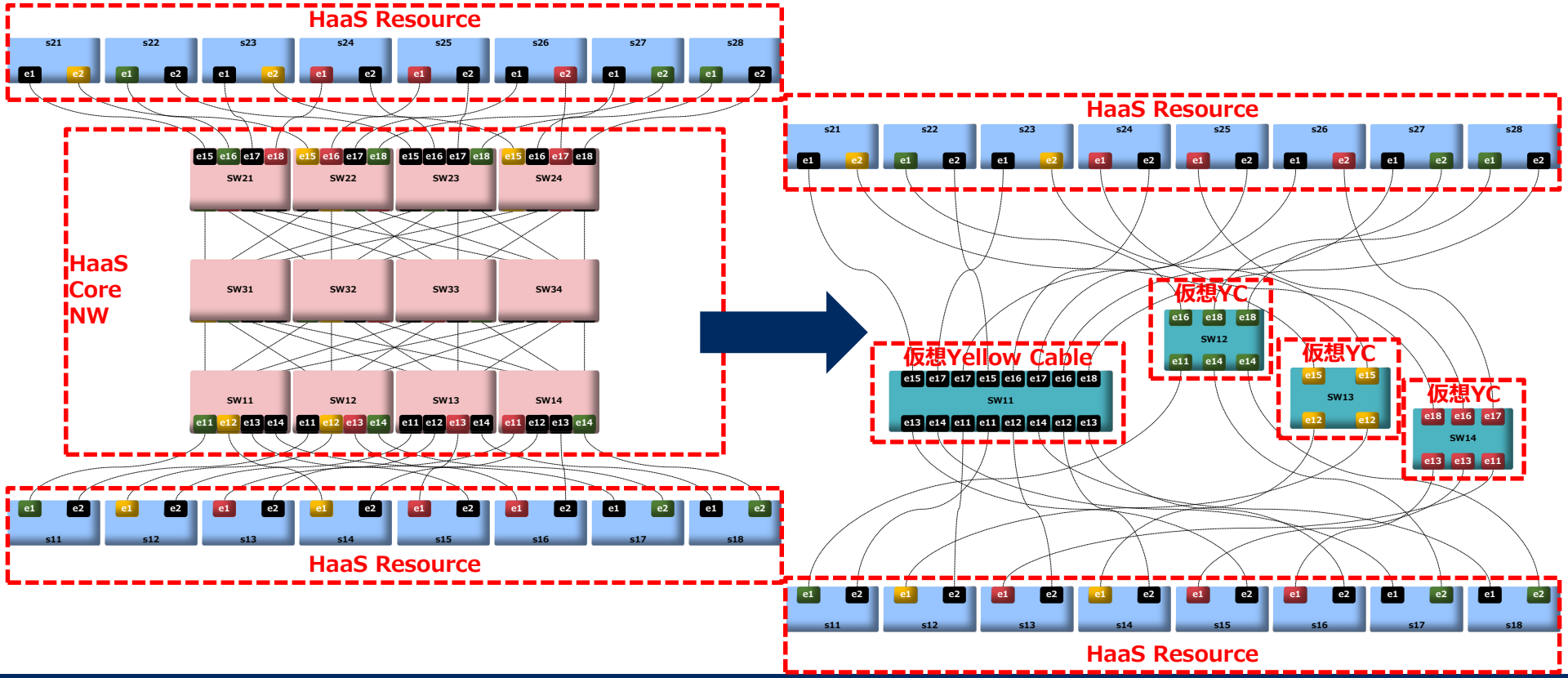
JANOG43と44のおさらい

やってみたシリーズ１　スイッチで物理を仮想化

JANOG43クラウド基盤の仮想化NWの設計と課題より

物理サーバーやストレージを仮想化したSW装置へ接続

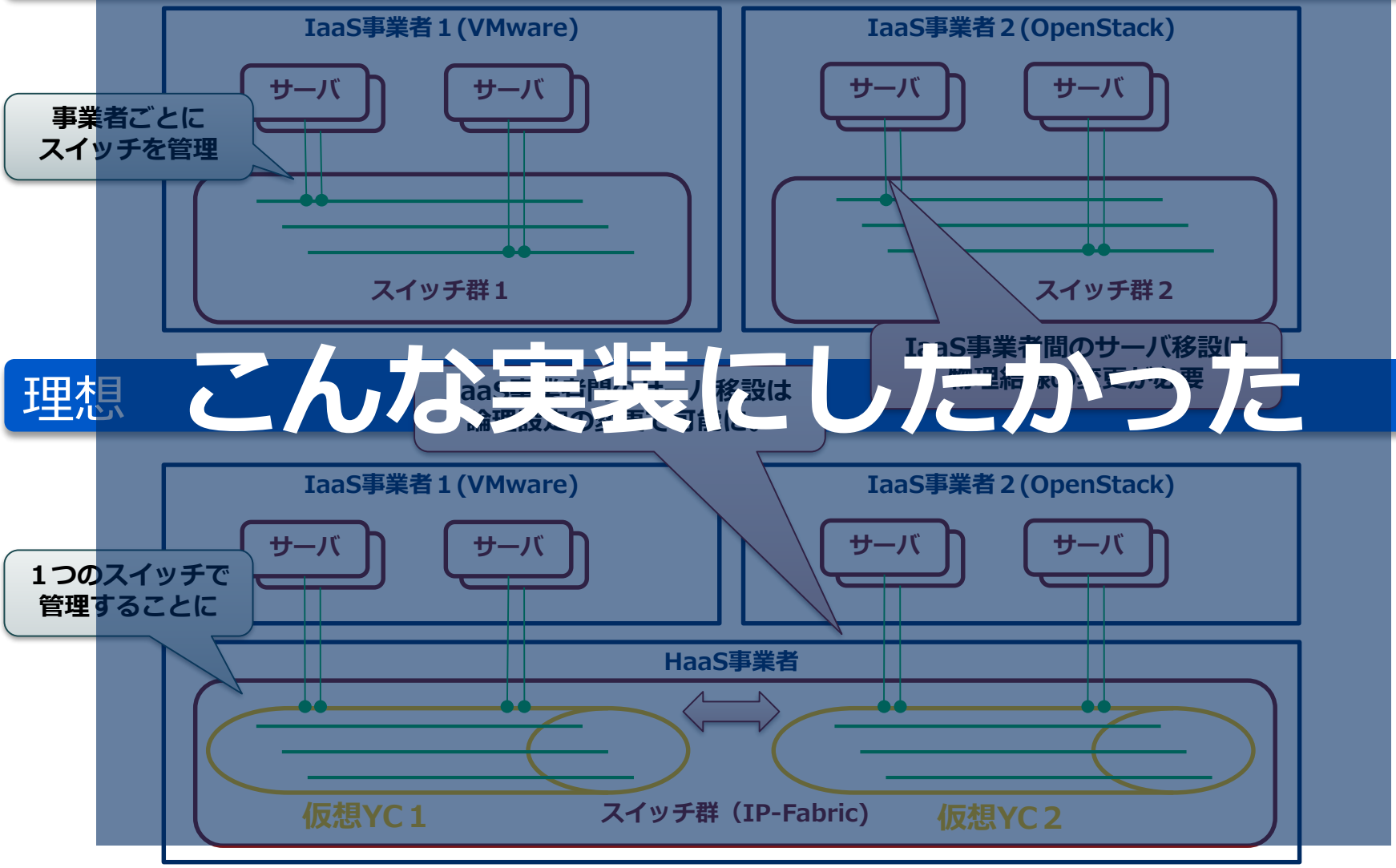
- HaaS Resource – 物理サーバーやルーター、ストレージ…
- HaaS Core NW – 物理SW装置で接続したNW
- 仮想Yellow Cable – 仮想化したSW装置



やってみたシリーズ１　スイッチで物理を仮想化

JANOG44クラウド基盤の仮想化NWの設計と課題-その後-より

HaaSレイヤがない場合



理想

こんな実装にしたかった

リソースを自由にプール化して HaaSテナントに払い出す

JANOG43 クラウド基盤の仮想化NWの設計と課題より

オーバーレイにEVPN/VxLANを採用（仮想YCを考えた）

- VLAN-Based Service
  仮想ネットワーク（物理サーバポートからでてくるVLAN）の収容数が４Kに限られる。

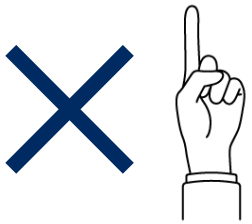
- VLAN Bundle Service
  仮想ネットワーク（物理サーバポートからでてくるVLAN）の収容数が４K x ４K可能であるが、MAC重複は許されない

- **VLAN-Aware Bundle Service**
  仮想ネットワーク（物理サーバポートからでてくるVLAN）の収容数が４K x ４K可能でMAC重複も可能

	案1 VLAN-Based Service	案2 VLAN Bundle Service	案3 VLAN-Aware Bundle Service
EVPN instance (EVI) Bridge domain (BD)	1EVI ： 1BD 1BD ： 1VLAN NG) EVI数の消費量が膨大	1EVI ： 1BD 1BD ： 多VLAN NG) EVI数の消費は抑えられるが、1BDに収容するためにBUMフラッディングの影響範囲が大きい。	1EVI ： 多BD 1BD ： 1VLAN EVI数の消費も抑えられ、BUMフラッディングの影響範囲も対象VLANのみである。
EVPN-VXLAN対応	1BDに1VLAN ⇔ 1VXLANマッピングする形になる。	1BDに複数VLAN ⇔ 1VXLANマッピングする形になる。 NG) 複数VLANをマッピングする場合はQ-in-Qを利用することになるが、Q-in-QではMAC重複ができない。	1BDに1VLAN ⇔ 1VXLANマッピングする形になる。
考察結果	1EVIに1VLANの紐付けとなるため収容数が少ない。拡張には機器のスケールアップが必要となる。	1EVIに複数VLANを紐づけできるが、BDが複数VLANで1つになること、および複数VLANを紐づけする際にQ-in-Qを利用するがQ-in-QでMAC重複ができない制限があるため、マルチテナントでの利用には制限が出てしまう。	1EVIに複数VLANを紐付け可能であるため、収容数を拡張可能である。

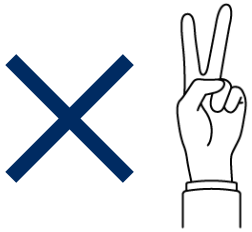
実装にあたって

実装にあたって再度案1～3を検討



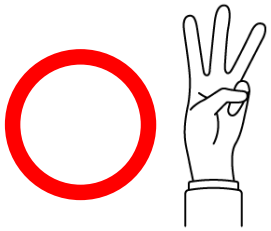
VLAN-Based Service

VLANが重複しないよう管理が必要



VLAN Bundle Service

VM用仮想MACなどの管理をIaaS事業者またぎで行う



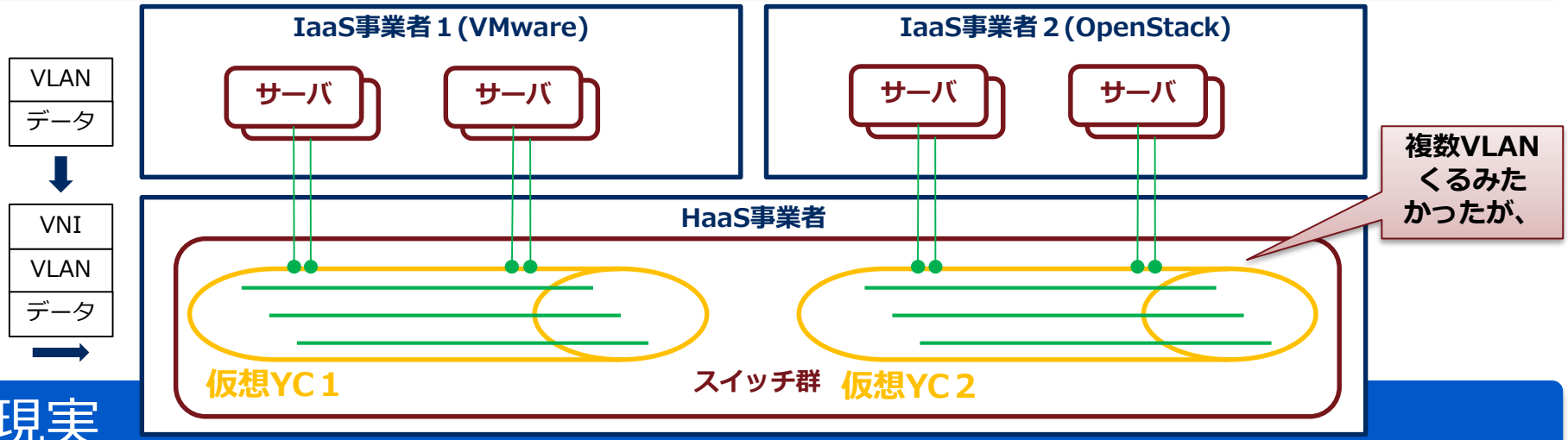
VLAN-Aware Bundle Service

VLANの重複管理・EVI数・Q-in-Qの制限がかからない

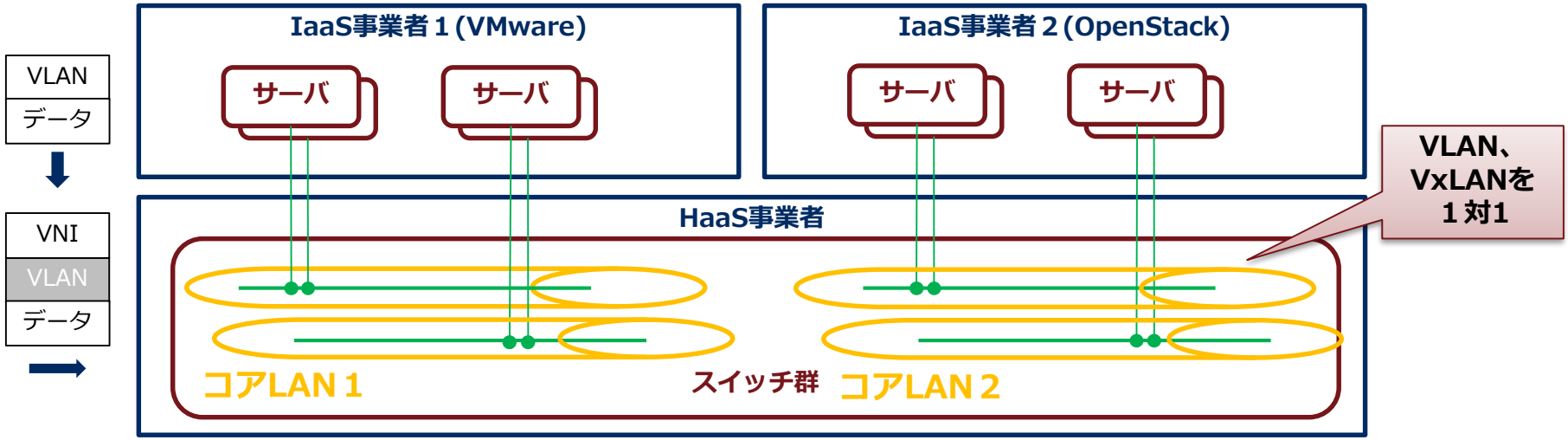
論物分離の理想と現実

JANOG44クラウド基盤の仮想化NWの設計と課題-その後-より

理想



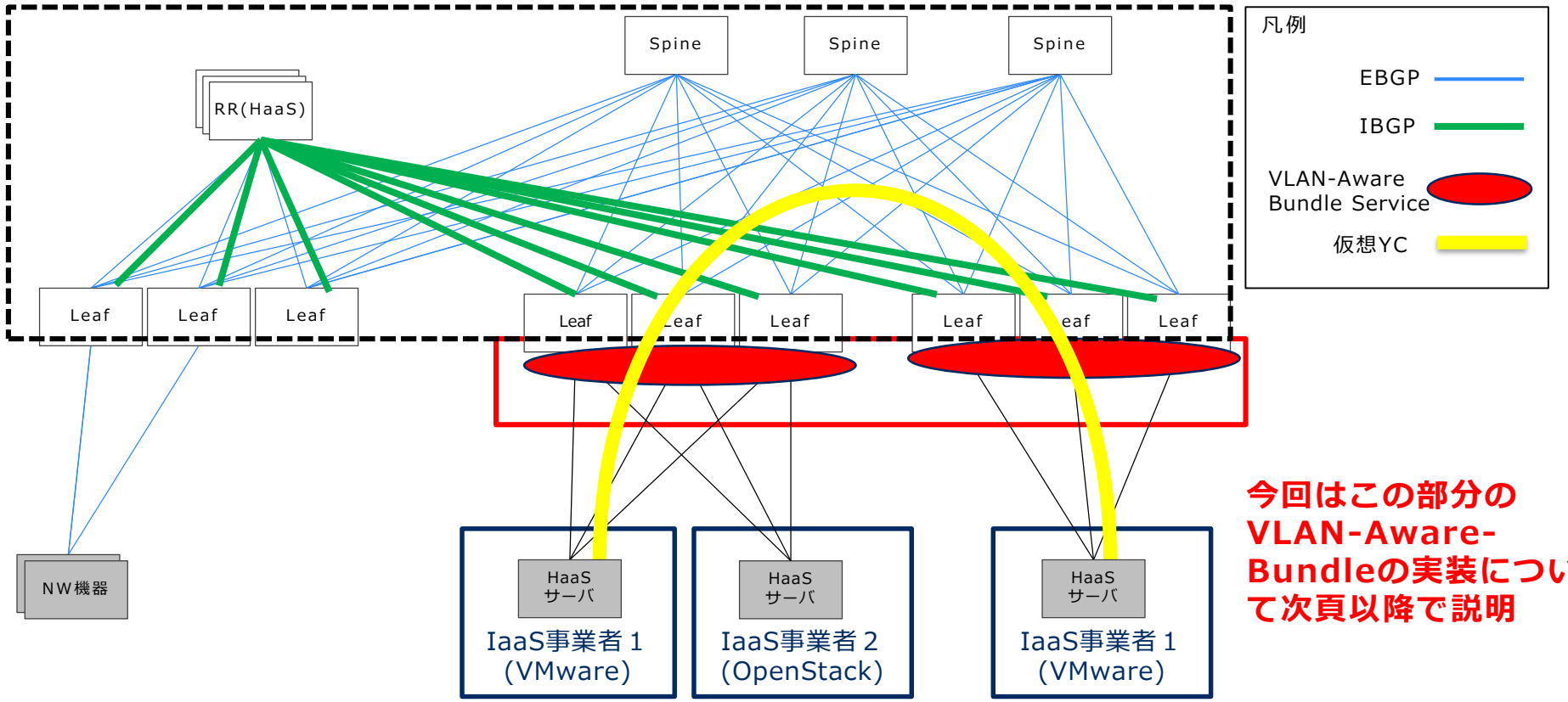
現実



仮想YC実装の全体構成概要

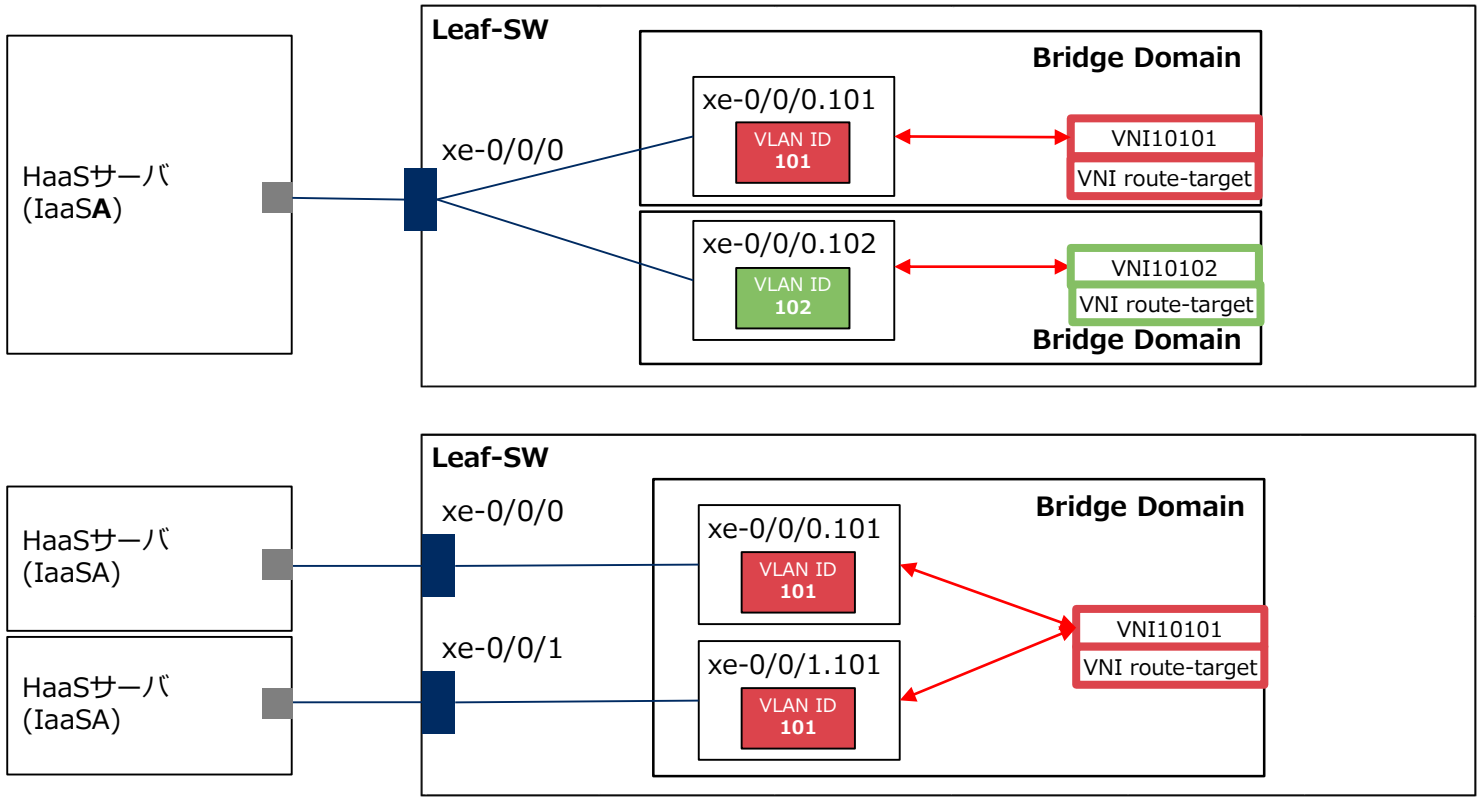
- IP-Fabric構成を採用
- EVPN/VXLANの経路情報交換にIBGPを使用
- IaaS事業者を分離しIBGPのRRに情報を乗せ、仮想YCが伸びる

仮想YCを延伸する部分についてはJANOG44で発表済みのため省略



今回はこの部分のVLAN-Aware-Bundleの実装について次頁以降で説明

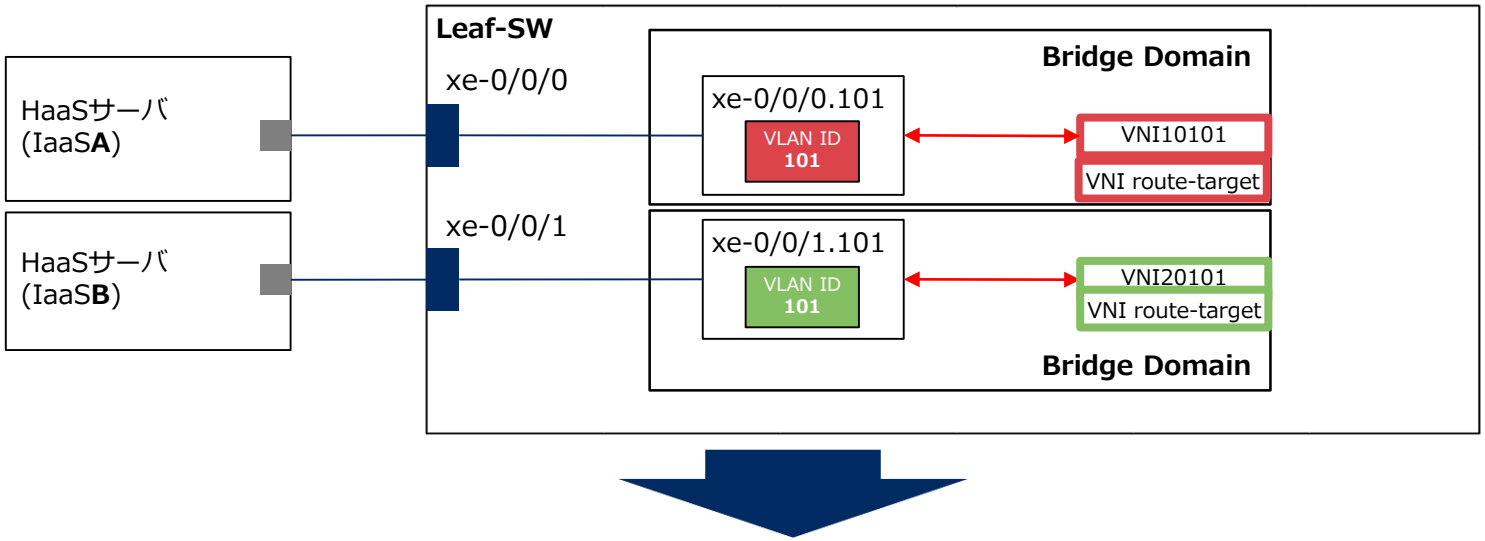
VLAN-Aware-Bundleの実装



VLAN Aware Bundle
（1EVI:複数Bridge Domain:1VLAN）を実現！

VLAN awareで実装した図

・別IaaS事業者のサーバが1Leaf-SW上で混在
・VLAN IDが重複



物理機器（サーバ・NW機器・アプライアンス等）に
マルチテナンシーを提供できる仕組みが完成

実装時の課題

マルチベンダー化

- シングルベンダー構成だと大規模障害の発生原因になったりするのでは？

Spineはベンダフリーに！一方Leafは・・・

ストーム対策

- テナントが発生させたBUMストームをどう防ぐか？

本当はサブインターフェースでストームコントロールを掛けたかった・・・

マルチベンダ機器でのIP管理

- BGP Unnumberedで簡略化したい！

マルチベンダでは実現できず・・・

仮想と物理の融合

やってみたシリーズ２　VMwareを物理とおしゃべり

ん、そういえばVMwareのVXLANは物理とは直接しゃべれない

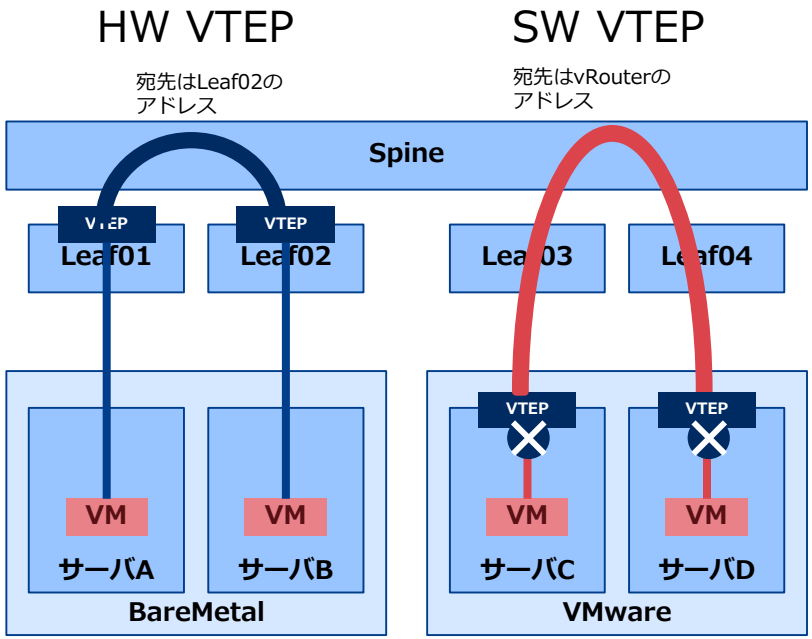
→具体的には・・・

NSXのVXLANの仕様（Geneve…）

IP fabricのRoute ReflectorとiBGPができない

(参考) HW VTEPとSW VTEPの違い

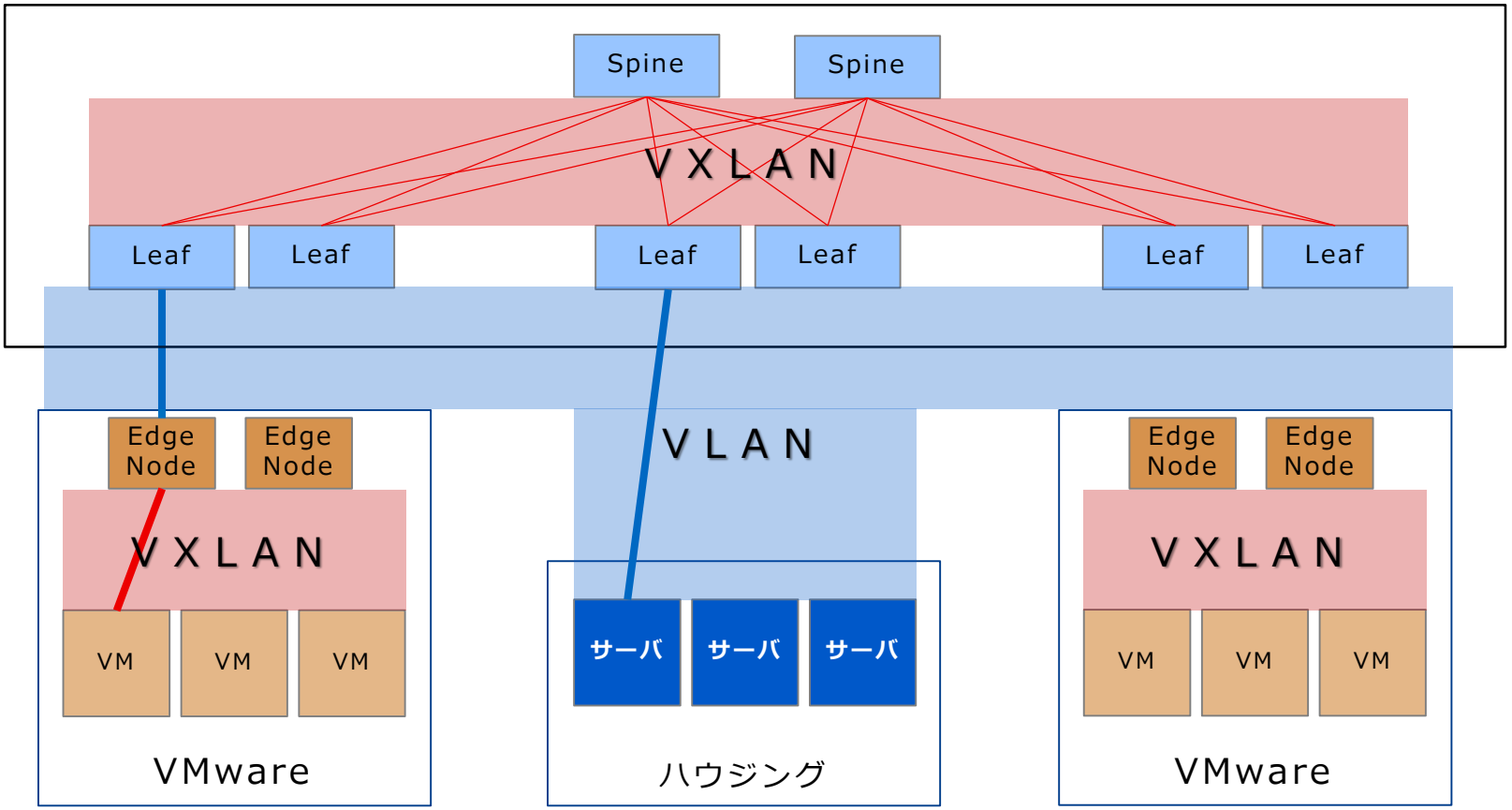
- HW VTEP
  - LeafでVTEPを行う
  - Leafに接続した機器から届いたパケットにVXLANヘッダを付与する
  - VXLANパケットの宛先IPはMACテーブルに基づいて宛先機器の直上のLeafのアドレス（loopback）になる
- SW VTEP
  - サーバ上でVTEPを行う
  - SDNによって制御される
  - VXLANパケットの宛先IPは宛先VMが存在するサーバのアドレスになる



やってみたシリーズ２　VMwareを物理とおしゃべり

仕方ない、Edge Node（L2BR）使ってVLAN変換してもらおう。。。

HaaS構成群



やってみたシリーズ２　課題

帯域

- Edge Node（L2BR）にトラフィックが集中する

理論上１仮想LAN当たり10Mbps程度・・・

Leaf SWのVLAN上限

- 一つのインターフェースでVLANが大量に必要

VLAN４K制限でEdge Nodeは２台くらいしか収容出来ない・・・

Edge Node（L2BR）の性能

- NIC性能分だけ全て処理性能は出るのか？

評価中・・・

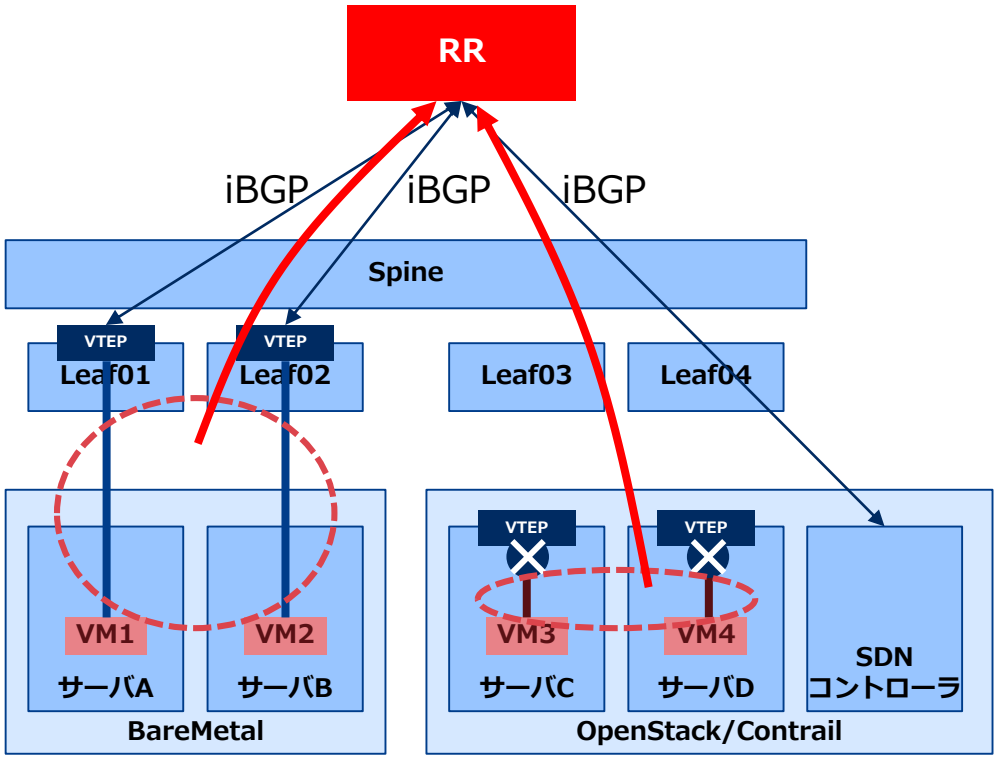
やってみたシリーズ３　OpenStackを物理とおしゃべり

仮想基盤をOpenStackに変えたらどうだろう

- 標準のNeutron制御では物理とはVLAN・・・
- ContrailならVLAN aware bundleを話せる！

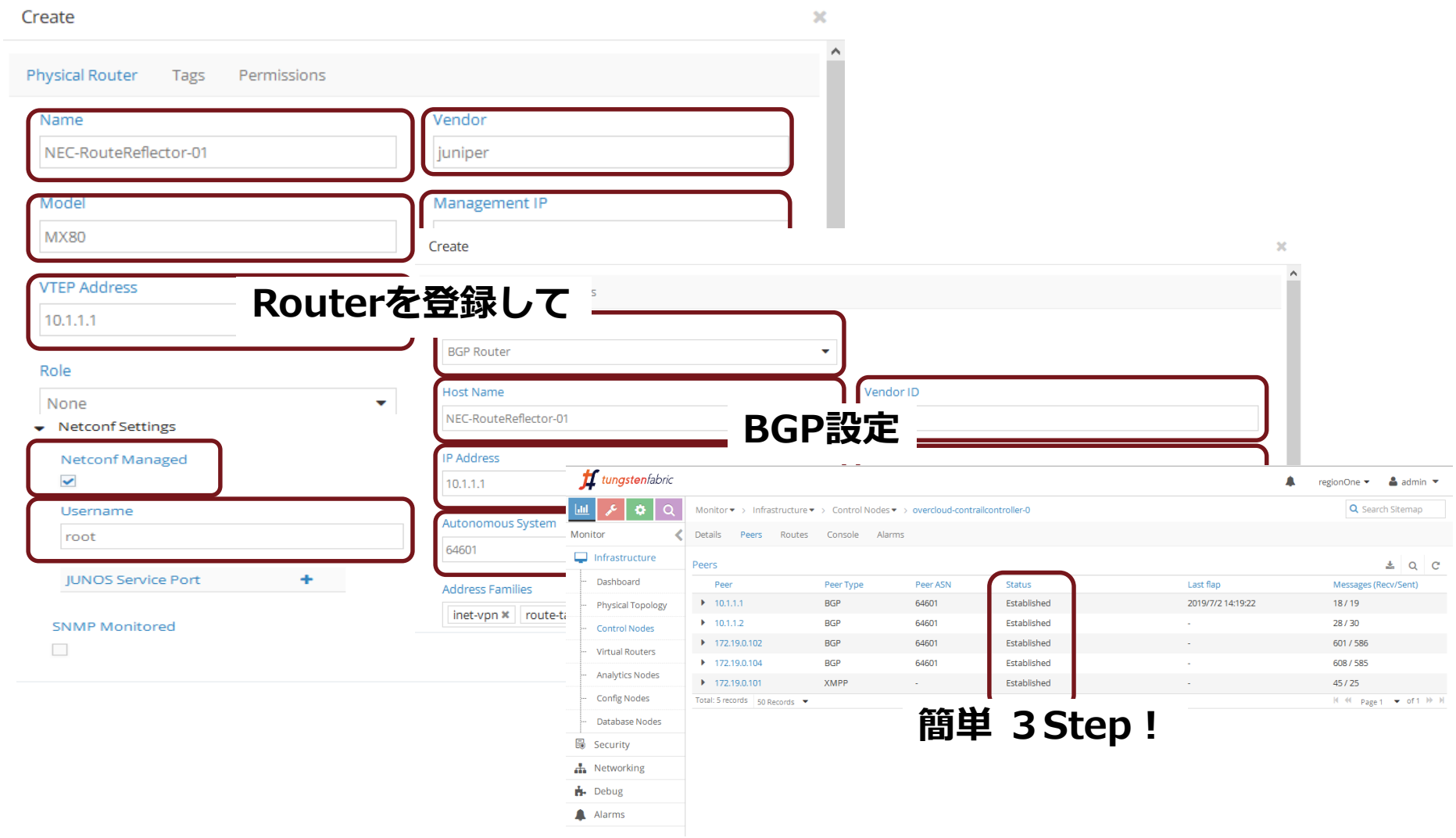
RRとSDNコントローラの連携(iBGPによる接続)によるVTEP接続

- NFVI内の仮想NW情報とLeafの情報を持つRRに送るために、SDNコントローラが代理でRRとiBGPを張る必要がある。



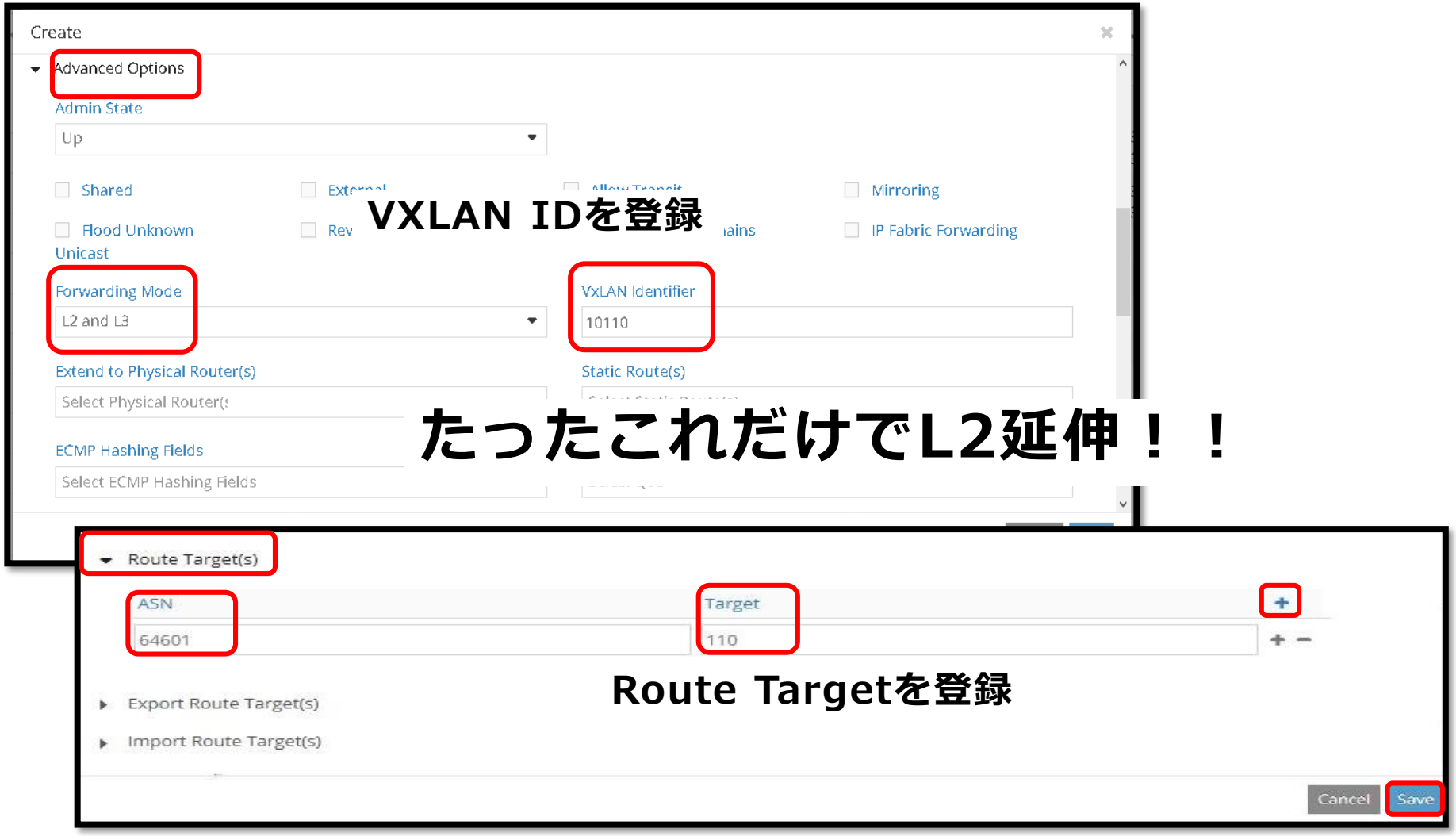
やってみたシリーズ３　OpenStackを物理とおしゃべり

IP FabricのRoute ReflectorとiBGP



やってみたシリーズ３　OpenStackを物理とおしゃべり

L2延伸も簡単



やってみたシリーズの考察

つまり仮想と物理が直接しゃべれるVLAN aware bundleって最高！

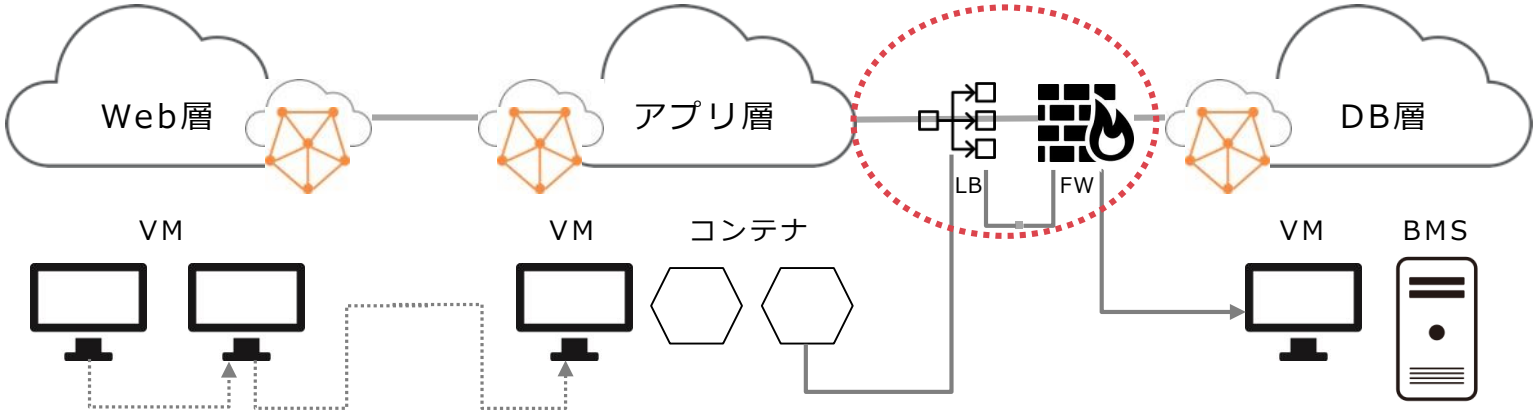
メリット１

物理機器に対してマルチテナンシーを適用できる！

メリット２

物理機器と仮想サーバをL2接続できる！

物理サーバと仮想サーバのL2接続はあまりなさそうだが、
物理NW Functionと仮想サーバをL2で接続することは多くあるはず。。。

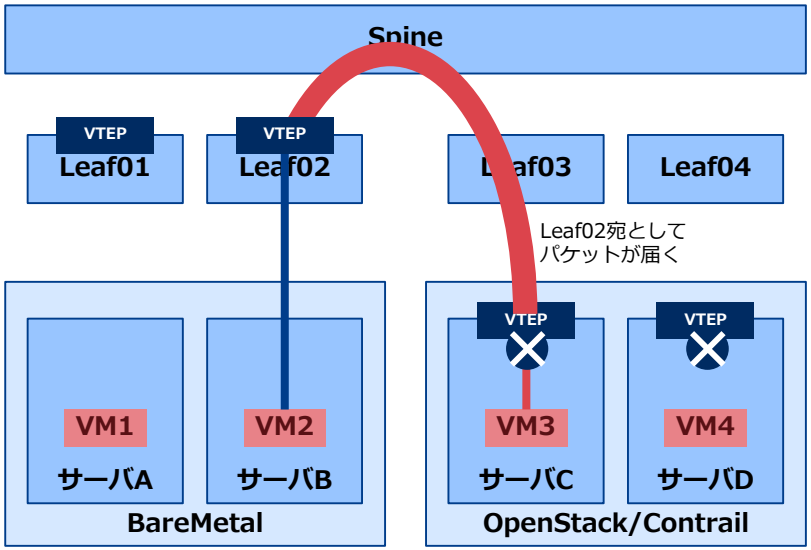


やってみたシリーズ３にも課題はある

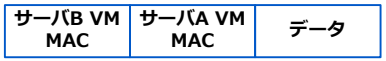
SW VTEPとHW VTEPがIPリーチャブルな構成が必要

- 前提

・HW VTEPを行うLeafはMAC情報（L2情報）に基づいて宛先LeafのIPに転送する。
・SDNによって制御されるSW VTEPで管理されたパケットは、Leaf上で宛先IP情報（L3情報）に基づいて転送する必要がある。
・next hop（GW IP）が必要となる。

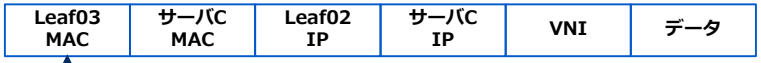


HW VTEPで転送するパケット（Leaf02に届くパケット）



← LeafはL2情報に基づいてVNIをつけて転送

SW VTEPされたパケット（Leaf03に届くパケット）



← LeafはL3情報に基づいて転送

サーバCのIPが所属するVLANのGWのMAC

<u>HaaSで管理するIPとIaaS側で付けるIPの連携が必要</u>

ディスカッション

ディスカッションポイント

- VMwareを物理とつなぐポイントはどうするのがBest Answerか？
- ベアメタルとのNW融合方法は、本当にVLAN aware bundleだけなのか？
  - 他のクラウド事業者様ではどのように対策を取られているのだろう？
- VLAN aware bundleは今後業界に広がりを見せるのか？
  - 本問題についてSWベンダー様はどのような対策を考えているのか？


Orchestrating a brighter world
NEC