「さくらのクラウド」を例に見る ネットワーク仮想化の設計と実装

さくらインターネット研究所
大久保修一 <ohkubo@sakura.ad.jp>

Agenda

- クラウドとは？IaaSとは？
- さくらのクラウドの目指すところ
- ネットワークの設計と実装
- 将来の構想
- 仮想アプライアンス
- まとめ

自己紹介

- 2003年4月 さくらインターネット入社
  - 以後、ネットワークの運用に携わる
- 2009年7月 さくらインターネット研究所
  - 発足と同時に異動
  - クラウド、IPv4アドレス枯渇について研究活動
- 2011年3月 クラウドサービスの開発に従事
  - 主にネットワーク部分を担当

ところで、、

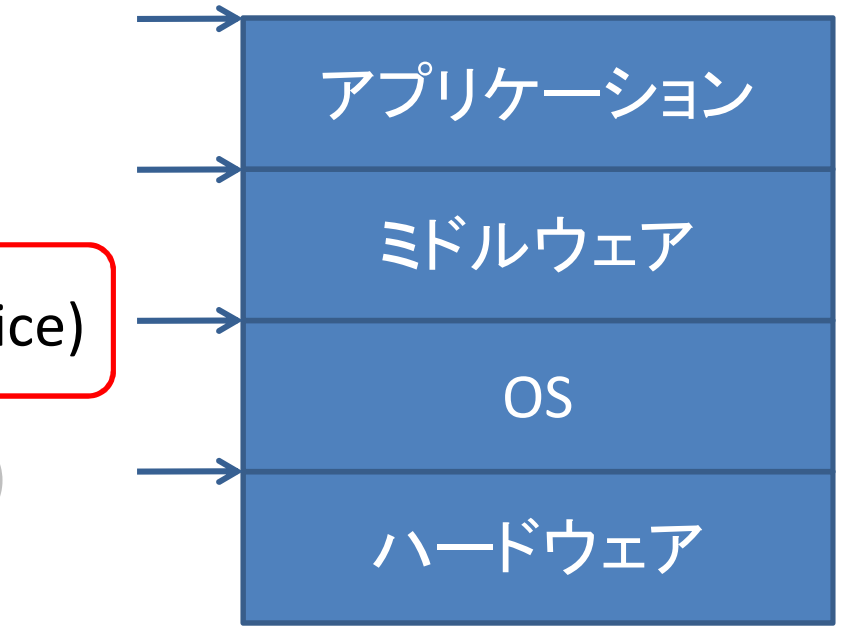
- 昨日(9/27)、Publickeyさんから PR記事を公開いただきました。
- PR:「さくらのクラウド」のアーキテクチャは、意外なほどシンプルだった
- http://www.publickey1.jp/blog/11/pr_sakuracloud.html

クラウドとは？

- NISTの定義がよく引用される
- 五つの特徴
  - On-demand self-service
    必要に応じてコンピュータリソースを利用可能
  - Resource pooling
    マルチテナントモデル、リソース配置場所の隠蔽
  - Rapid elasticity
    スケールアウト可能
  - Broad network access
    様々なプラットフォームから利用可能
  - Measured Service
    計測によるサービスの透明性

IaaSとは？

- 三つのサービスモデル
  - SaaS (Software as a Service)
  - PaaS(Platform as a Service)
  - IaaS(Infrastructure as a Service)
  - HaaS(Hardware as a Service)



現在の「さくらのクラウド」はIaaSです。

IaaSに求められること

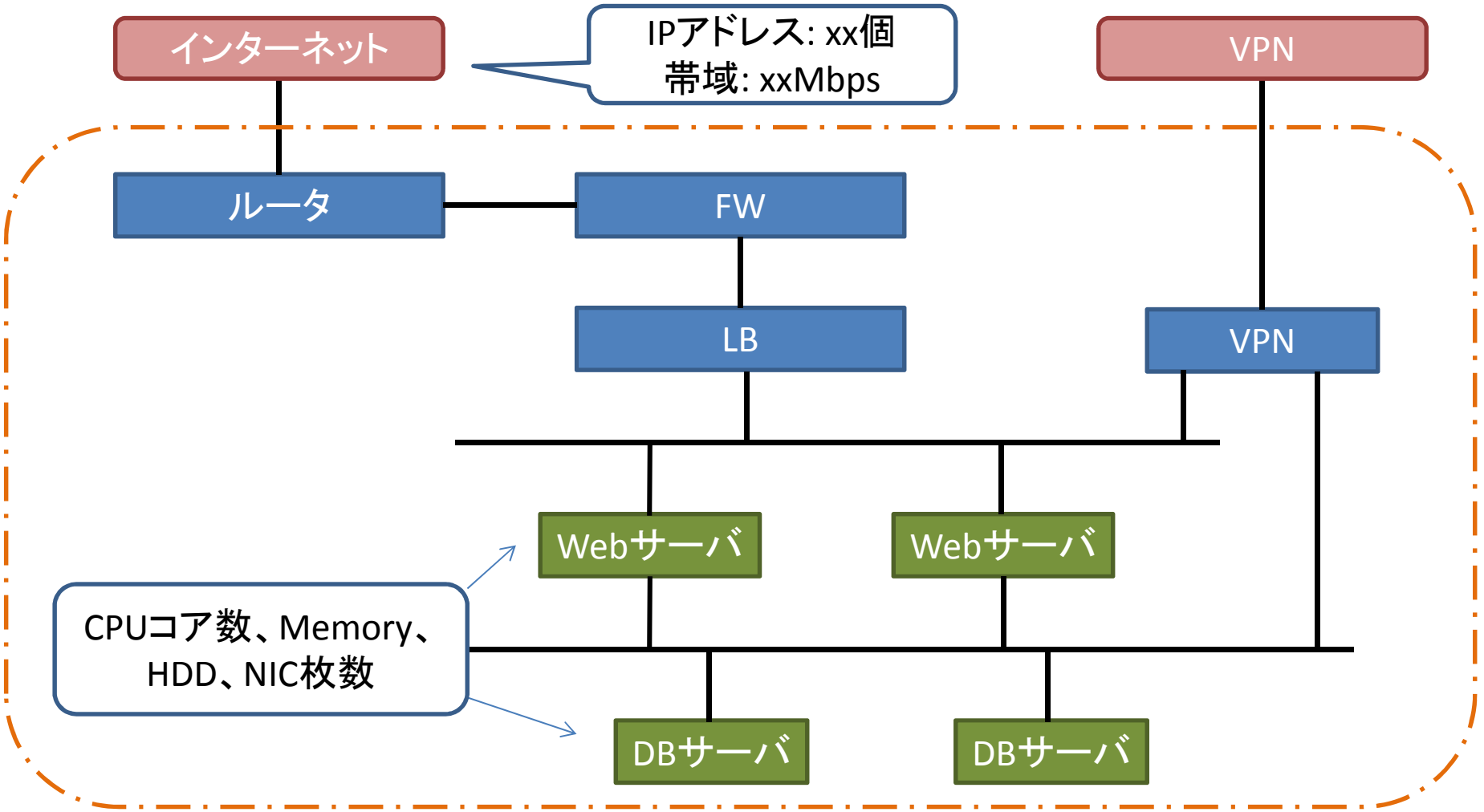
- 仮想マシンを自由に構成して作成できる
  （例）
  - CPUコア数: 2個
  - メモリ: 8GB
  - HDD: 1TB
  - NIC: 4枚
- 仮想ネットワークを自由に構成できる
- 様々な仮想アプライアンスを使える
  - 例：仮想ルータ、仮想FW、仮想LB
- 利用実績に応じた支払い

そういえば・・・・

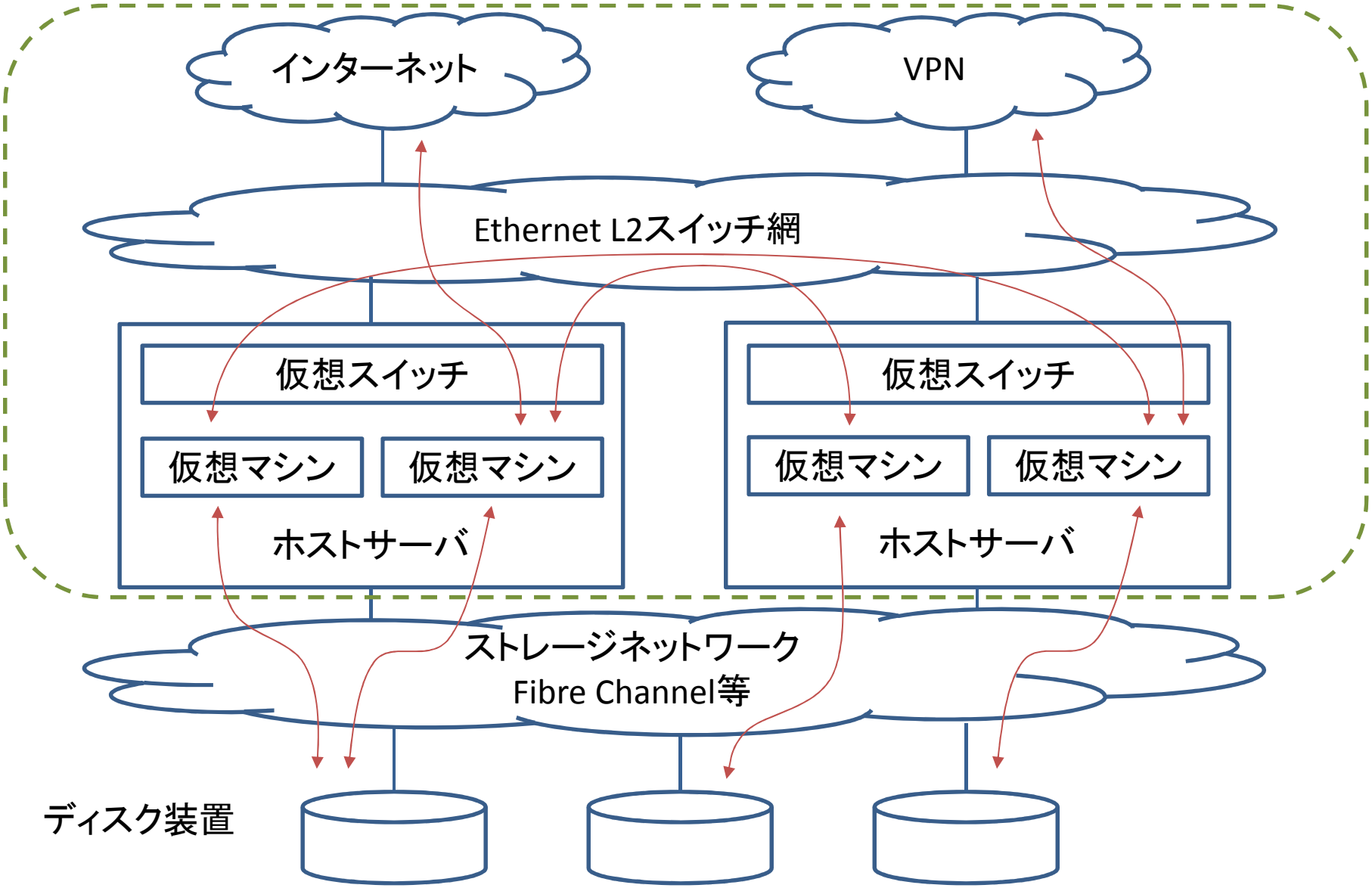
ちょうど１年くらい前に社長がプロトタイプを作っていた

IaaS上でのシステム構築例

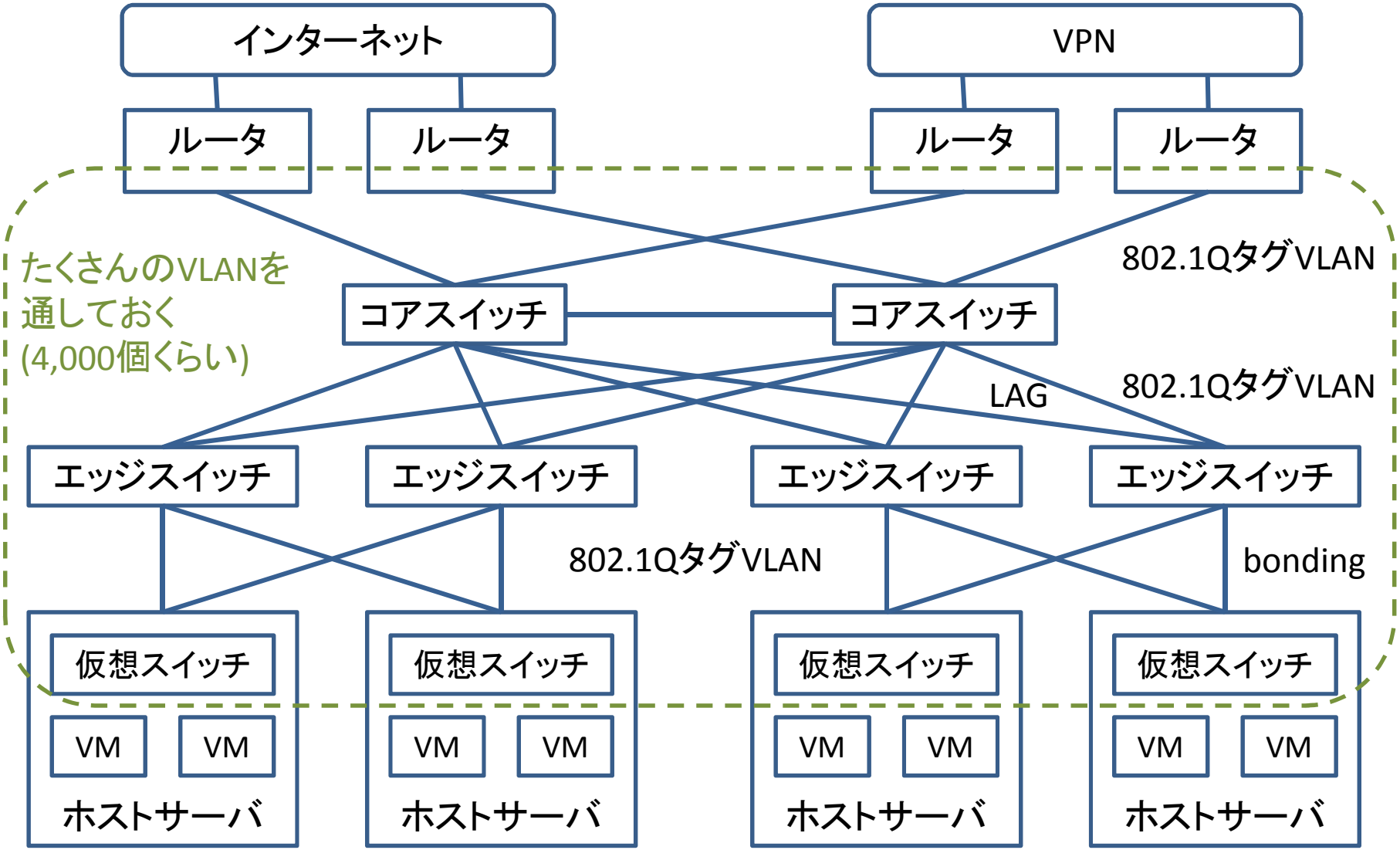
利用者がサーバ構成、ネットワーク構成を定義



IaaSインフラの構成

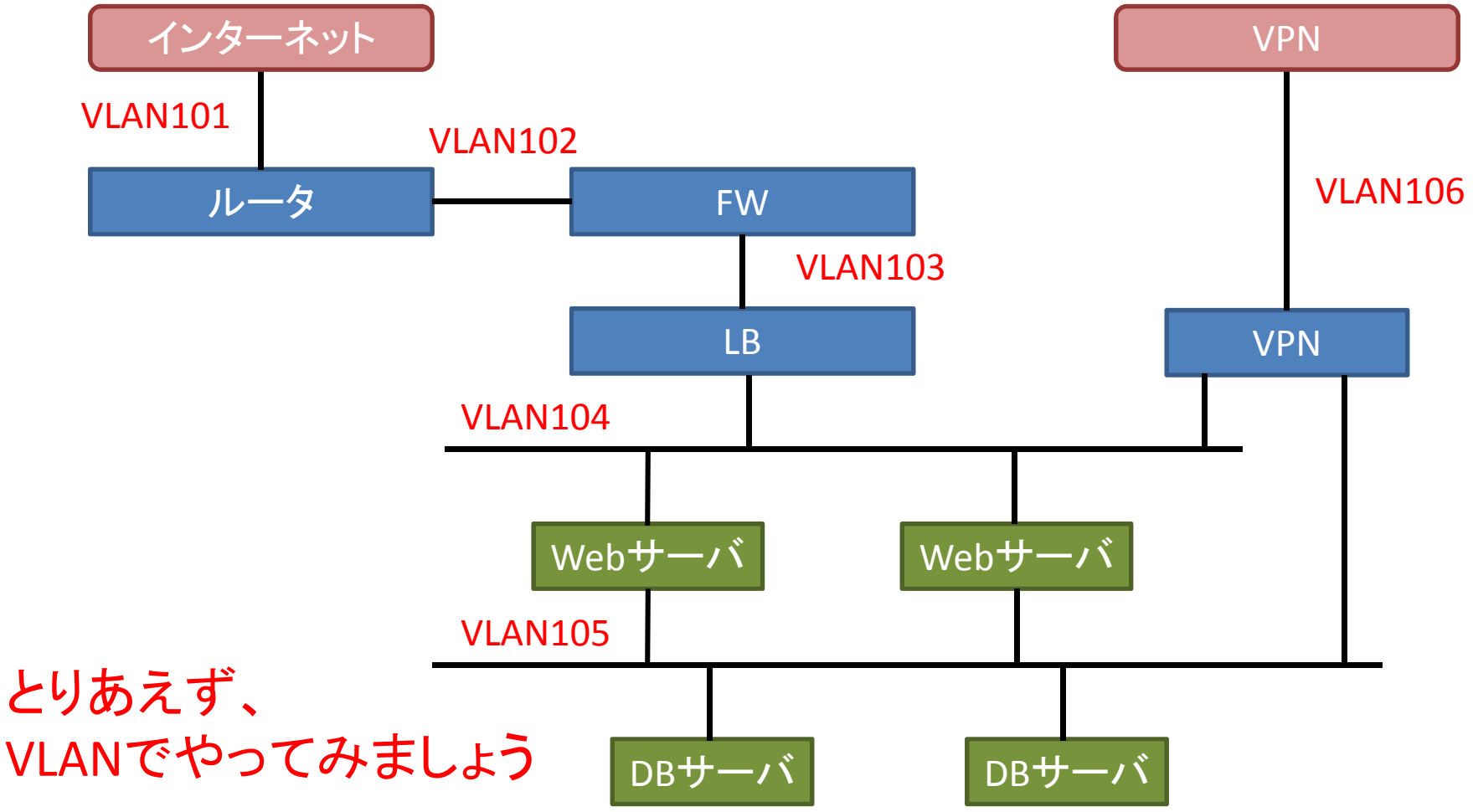


IaaSネットワークの物理構成



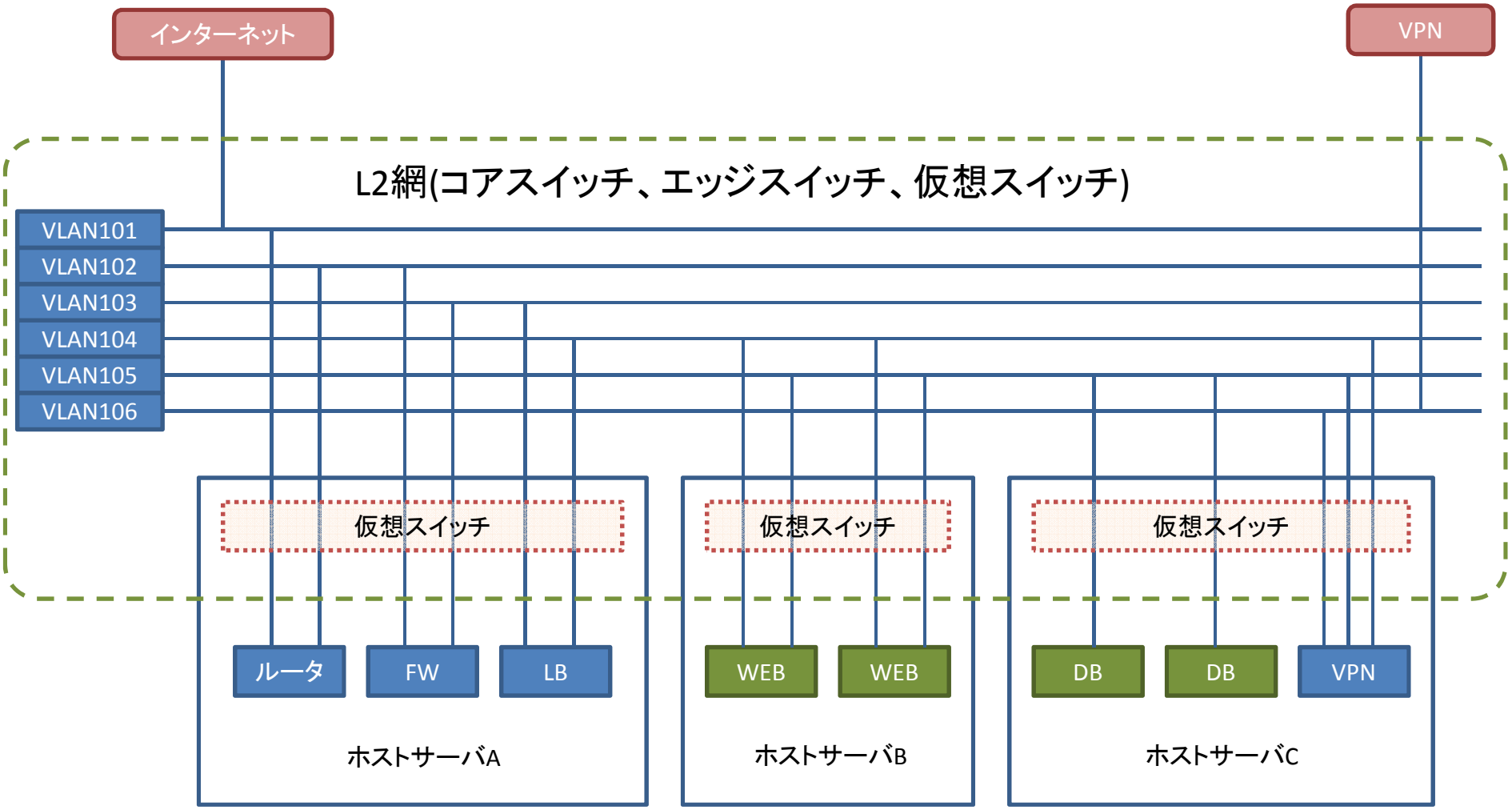
仮想システムプロビジョニング例

これを、IaaSインフラ上に展開してみる



とりあえず、
VLANでやってみましょう

クラウド上に配置したシステム



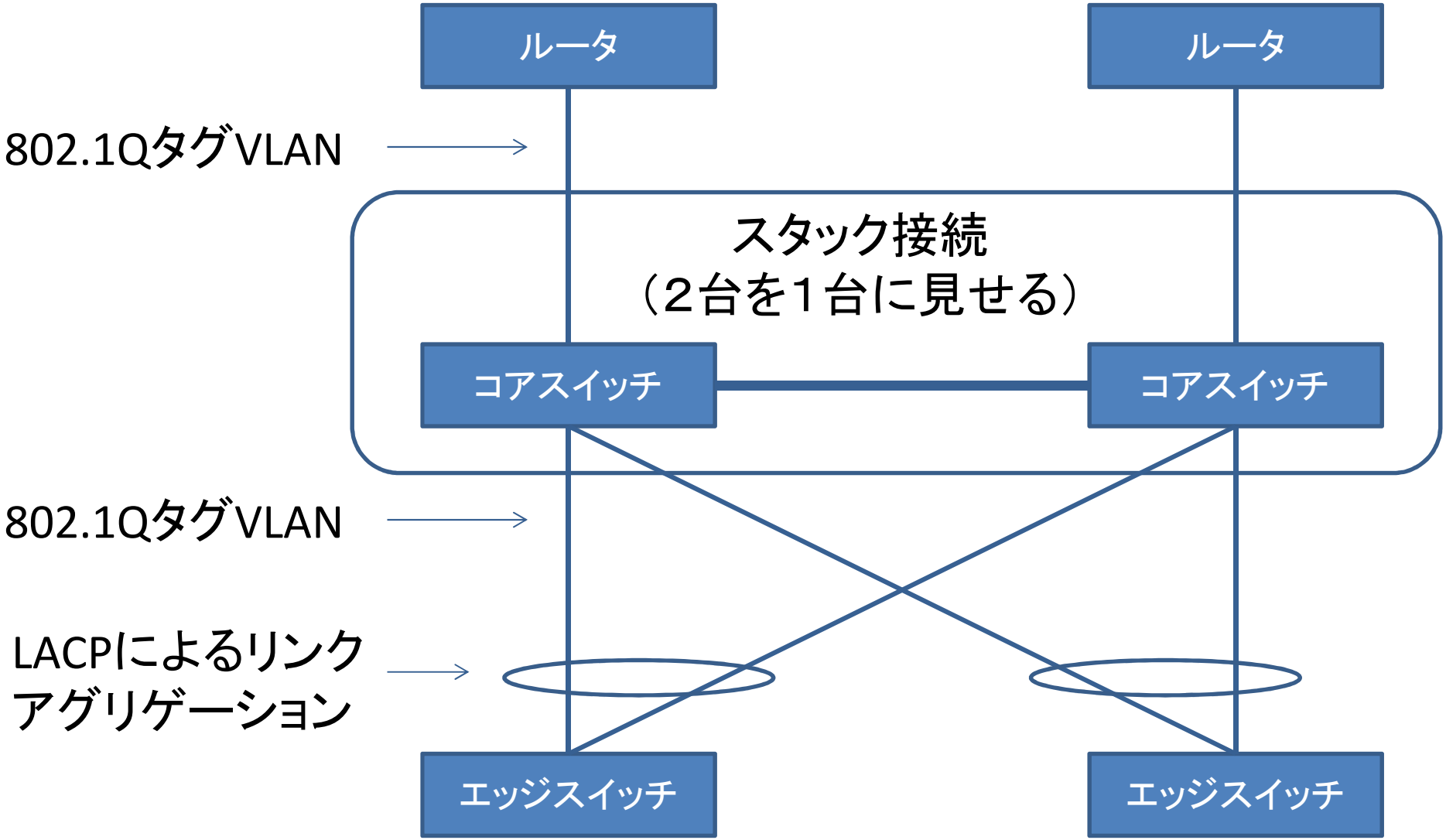
1ゾーンあたりの規模

- VLAN数: MAX4,096 (VLAN ID数の制限による)
- コアスイッチ数: 2台 (2台1セット)
- エッジスイッチ数: 数十台 (2台1セット)
- ホストサーバ数: 数百台
- VM数: MAX8,000
- VMあたり仮想NIC数: 平均2個
- MACアドレス数: 16,000 (8,000×2)

検討が必要な要素

- コアスイッチ
- エッジスイッチ
- 仮想スイッチ
- ルータ

コア・エッジスイッチ間の接続



コアスイッチの要件

- 2台でエッジスイッチやルータを集約する
- 冗長性: スタック機能
- VLAN数: 3,500以上
- MACアドレス数: 16,000以上
- ポート数: 40ポート以上

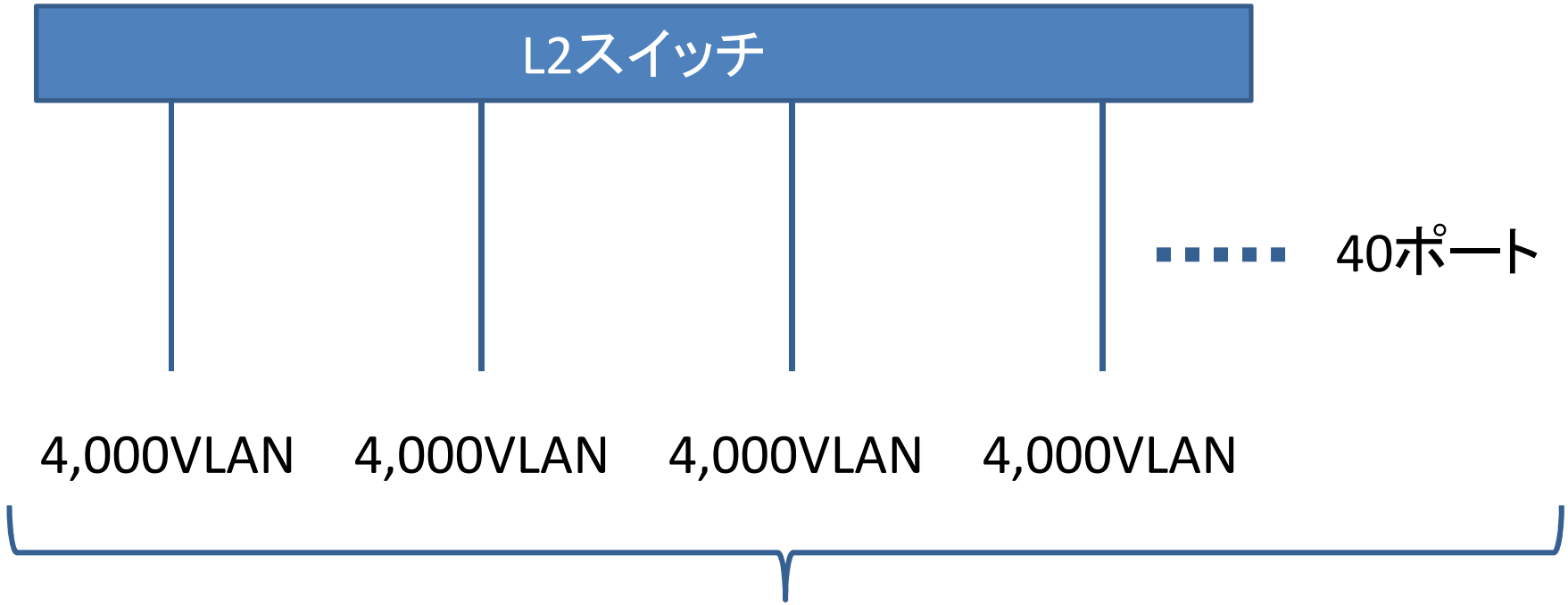
エッジスイッチの要件

- 各ラックに2台ずつ(表、裏) 設置し、ラック内サーバを収容する
- 冗長性: Link Aggregation
- VLAN数: 3,500以上
- MACアドレス数: 16,000以上
- ポート数: 40ポート以上
- ループ検出機能

検証時に発生した問題

- ロジカルポート数の不足
- ショートパケットでワイヤレート出ない
- 使えないVLAN-IDが存在する
- configが長い
- 溢れてないのに学習できないMACアドレスがある
- VLANを設定すると、MACアドレステーブルを消費する
- LACPがダウンする

ロジカルポート数の不足



約4,000×40ポート=約160,000ロジカルポート必要

ロジカルポート数が12,000や24,000の制限があるスイッチ
→採用せず

ショートパケットでワイヤレートでない

Etherのフレームフォーマット

最小フレーム長でワイヤレート出るには？
14.88Mpps = 10,000,000,000 / 8 / 84

ワイヤレートでないスイッチはよくある
→ 現実的に問題ない範囲ならOKとする。

使えないVLAN IDが存在する

- VLAN 1002番
  - 某箱では、FCoEのIDとして予約されている
- VLAN 1002～1005と1006以降いくつか
  - 某箱では、routed interfaceに割り当てられたり、予約されていたりする(show vlan internal usage)
- VLAN 3584～
  - 某箱では使用できない
- Configできても、正式サポート数が少ない(2000VLANとか。。)
- クラウドコントローラにて、お客様に割り当てないようにする必要あり

Configが長くなる

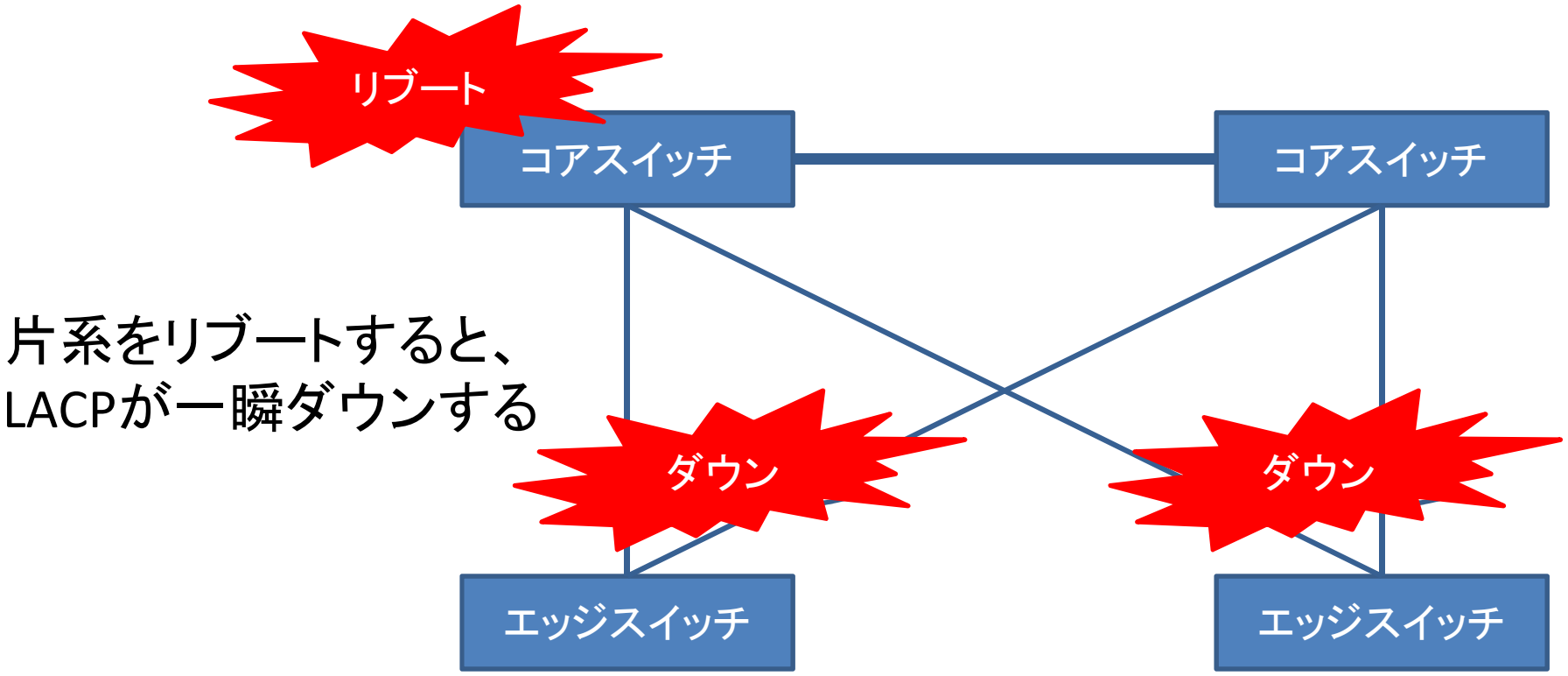
```
interface Vlan 1
  no shutdown
!
interface Vlan 2
  no shutdown
!

・・・省略・・・

interface Vlan 1998
  no shutdown
!
interface Vlan 1999
  no shutdown
!
interface Vlan 2000
  no shutdown
```

1VLAN毎に設定が必要

VLAN設定をまとめれるとうれしい

LACPがダウンする

冗長性が確保できているはずだが・・・



片系をリブートすると、
LACPが一瞬ダウンする

スイッチのバグ → 修正してもらった

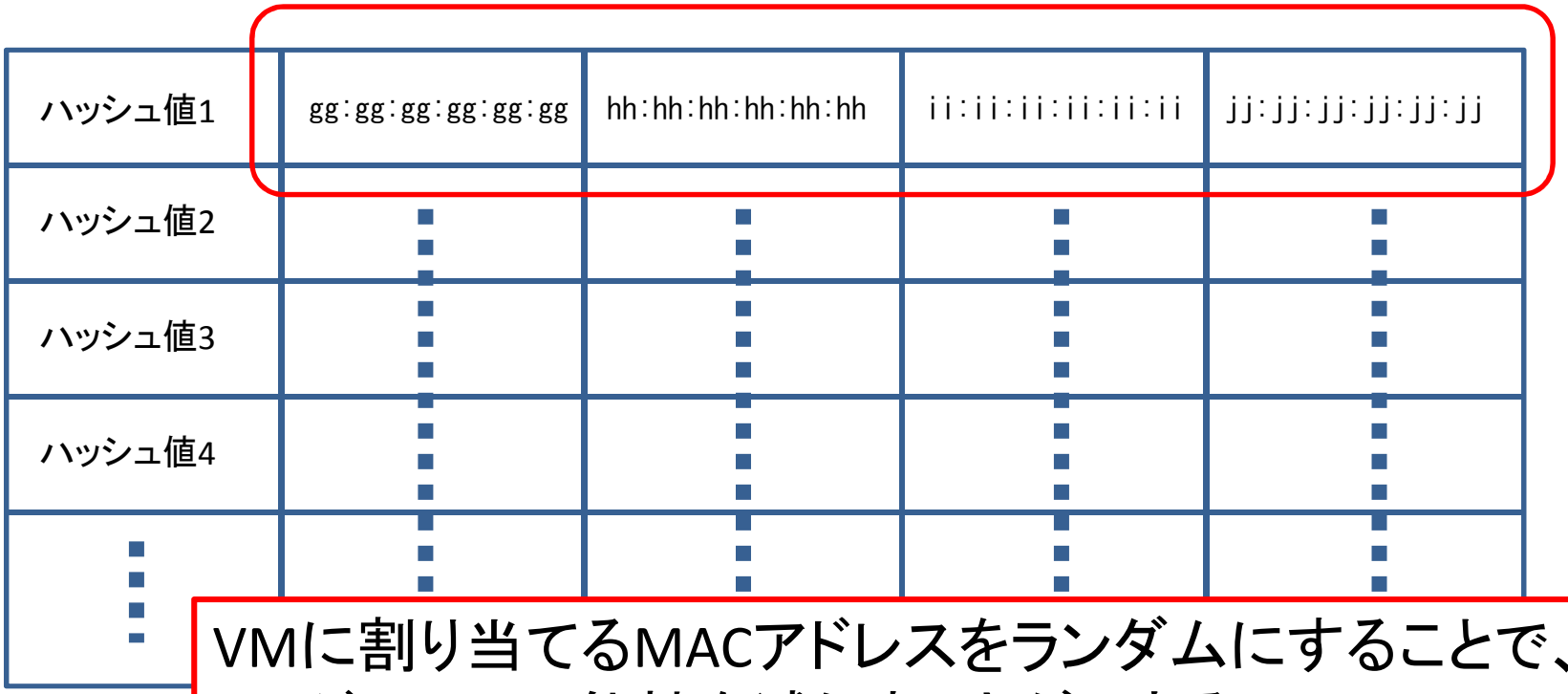
MACアドレスハッシュコリジョン問題

- 多くのL2スイッチは、MACアドレスの学習にハッシュテーブルを使用している
- ハッシュ値のコリジョンが発生し、ハッシュテーブルの深さを超えると、学習できないMACアドレスが発生する
- 学習できないMACアドレスは、Unknown Unicast扱いとなりフラッディングする（設定によっては帯域制限がかかり、通信に影響）

MACアドレスハッシュコリジョン問題

同じハッシュ値をとる5つ目のMACアドレスが
来ると学習できない

4段の例



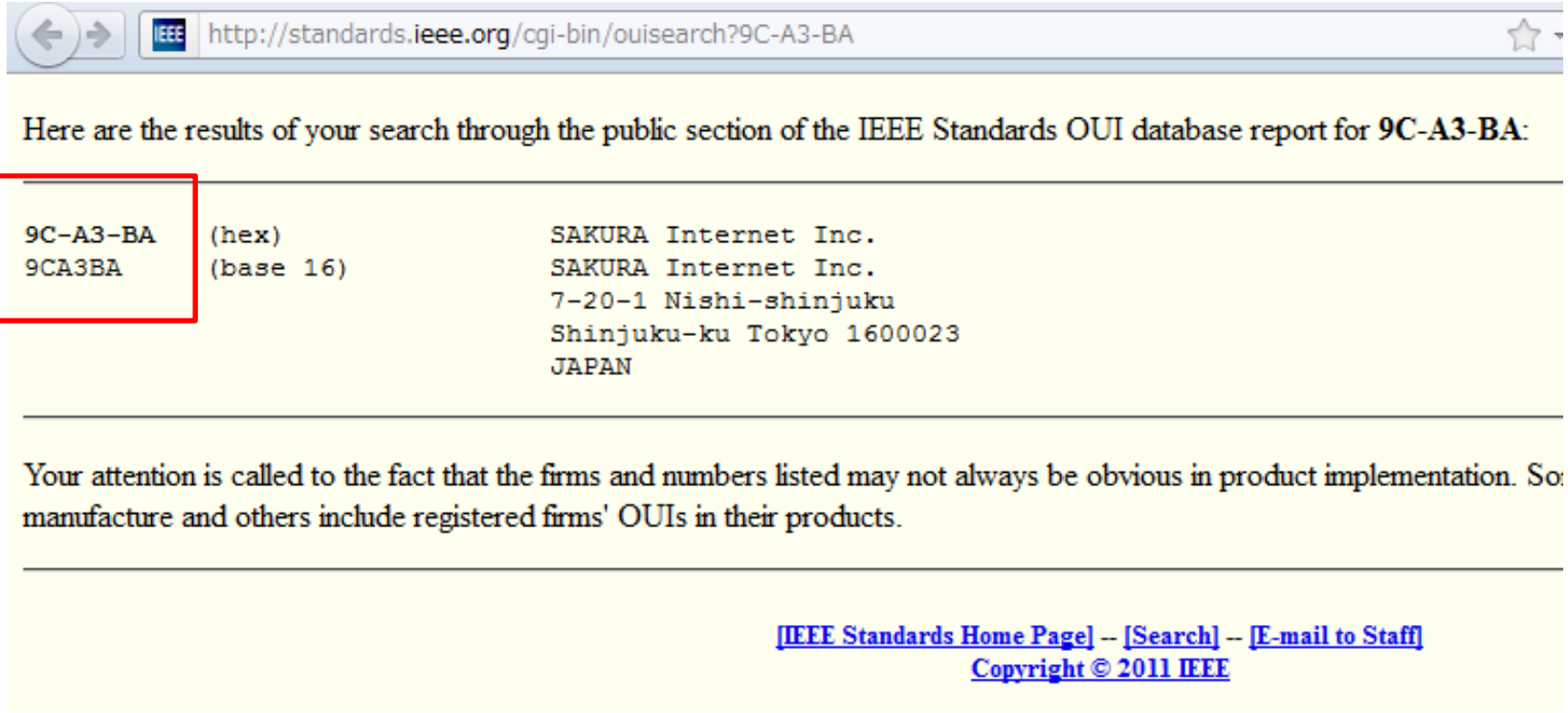
VMに割り当てるMACアドレスをランダムにすることで、
コリジョンの可能性を減らすことができる

VLANでMACエントリを消費

- 某B社のスイッチ
- スペックでは32,000のMACアドレステーブル
- ただ、1VLAN設定すると、3エントリ消費
- CPUに吸い上げる特殊なMACアドレスを自動で登録しているもよう
- 4,000VLAN設定すると、12,000エントリ消費
- 残り、20,000エントリしか使えない。。。

余談：OUI取得しました

http://standards.ieee.org/cgi-bin/ouisearch?9C-A3-BA

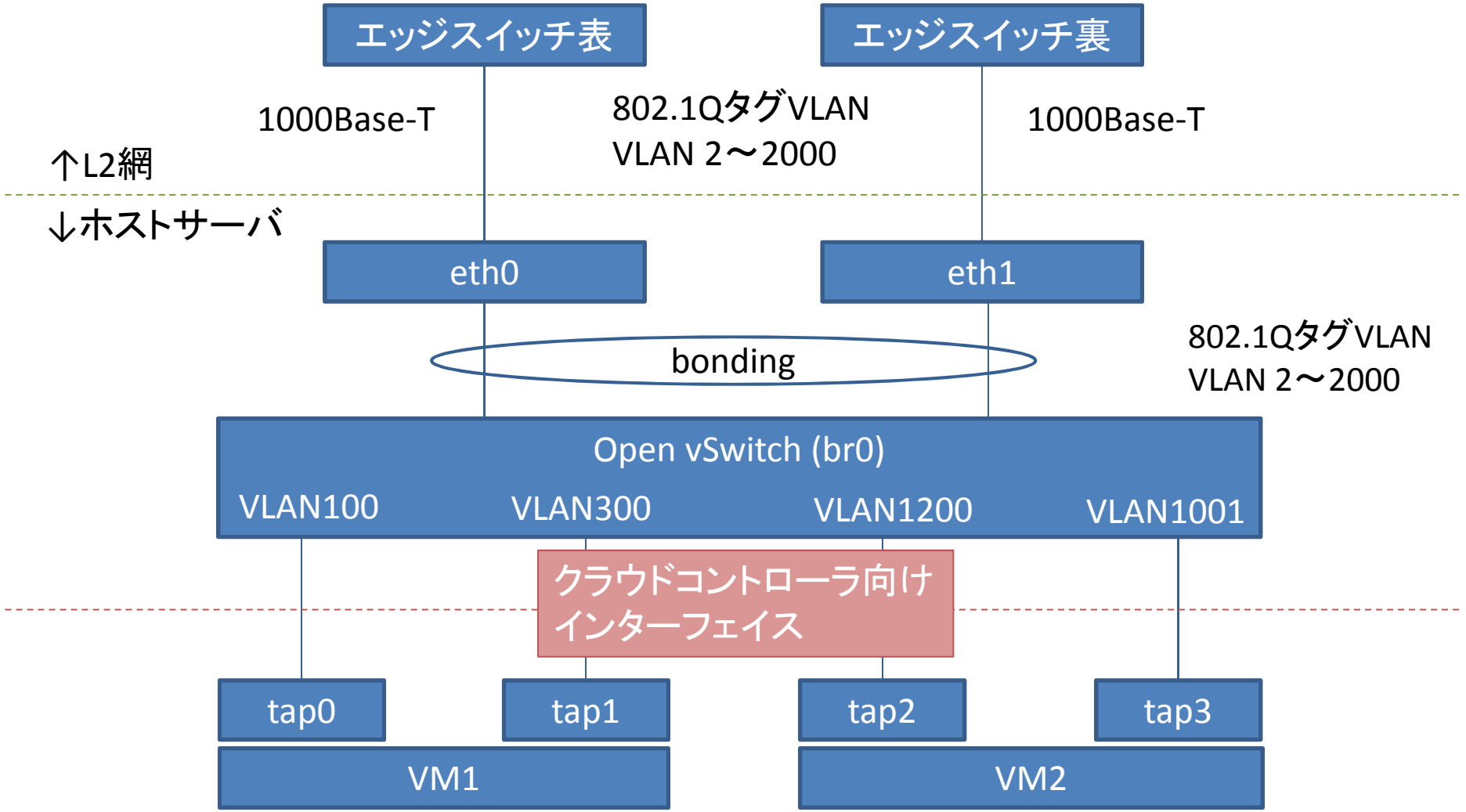


他のクラウドや他の物理ネットワークとの
L2レベルでの相互接続も問題なし！

仮想スイッチの要件

- ホストサーバ上で動作し、物理ネットワークとVM間の通信の橋渡しを行う。
- 冗長性: Bonding
- VLAN数: 3,500以上
- MACアドレス数: 16,000以上
- VM向けにフィルタ設定が必要
  - MACスプーフィング対策、ARPスプーフィング対策、
  - IPスプーフィング対策
  - その他
- 仮想ポートで帯域制限
- 弊社では、Open vSwitchを採用

弊社の実装例


エッジスイッチ表
エッジスイッチ裏
1000Base-T
802.1Qタグ VLAN
VLAN 2～2000
1000Base-T
↑L2網
↓ホストサーバ
eth0
eth1
bonding
802.1Qタグ VLAN
VLAN 2～2000
Open vSwitch (br0)
VLAN100
VLAN300
VLAN1200
VLAN1001
クラウドコントローラ向け
インターフェイス
tap0
tap1
tap2
tap3
VM1
VM2

Open vSwitchのMAC学習数

- デフォルトで2,048になっていた
  全然足りない
- ソースコードを変更して32,768にして使用

```
% vi lib/mac-learning.h

//#define MAC_MAX 2048
#define MAC_MAX 32768
```

共有セグメントのスプーフィング対策

Open vSwitchでは、Open Flowのエントリとして
フィルタを設定する。

```
% ovs-ofctl add-flow br0 in_port=8,actions=drop
% ovs-ofctl add-flow br0 in_port=8,dl_src=52:54:00:f4:c7:47,ip,
  nw_src=10.0.140.1,actions=normal
% ovs-ofctl add-flow br0 in_port=8,dl_src=52:54:00:f4:c7:47,arp,
  nw_src=10.0.140.1,actions=normal
% ovs-ofctl add-flow br0 in_port=8,dl_src=52:54:00:f4:c7:47,udp,
  nw_src=0.0.0.0,tp_src=68,nw_dst=255.255.255.255,
  tp_dst=67,actions=normal
```

帯域制限

- とっても簡単
  - 10Mbpsの例

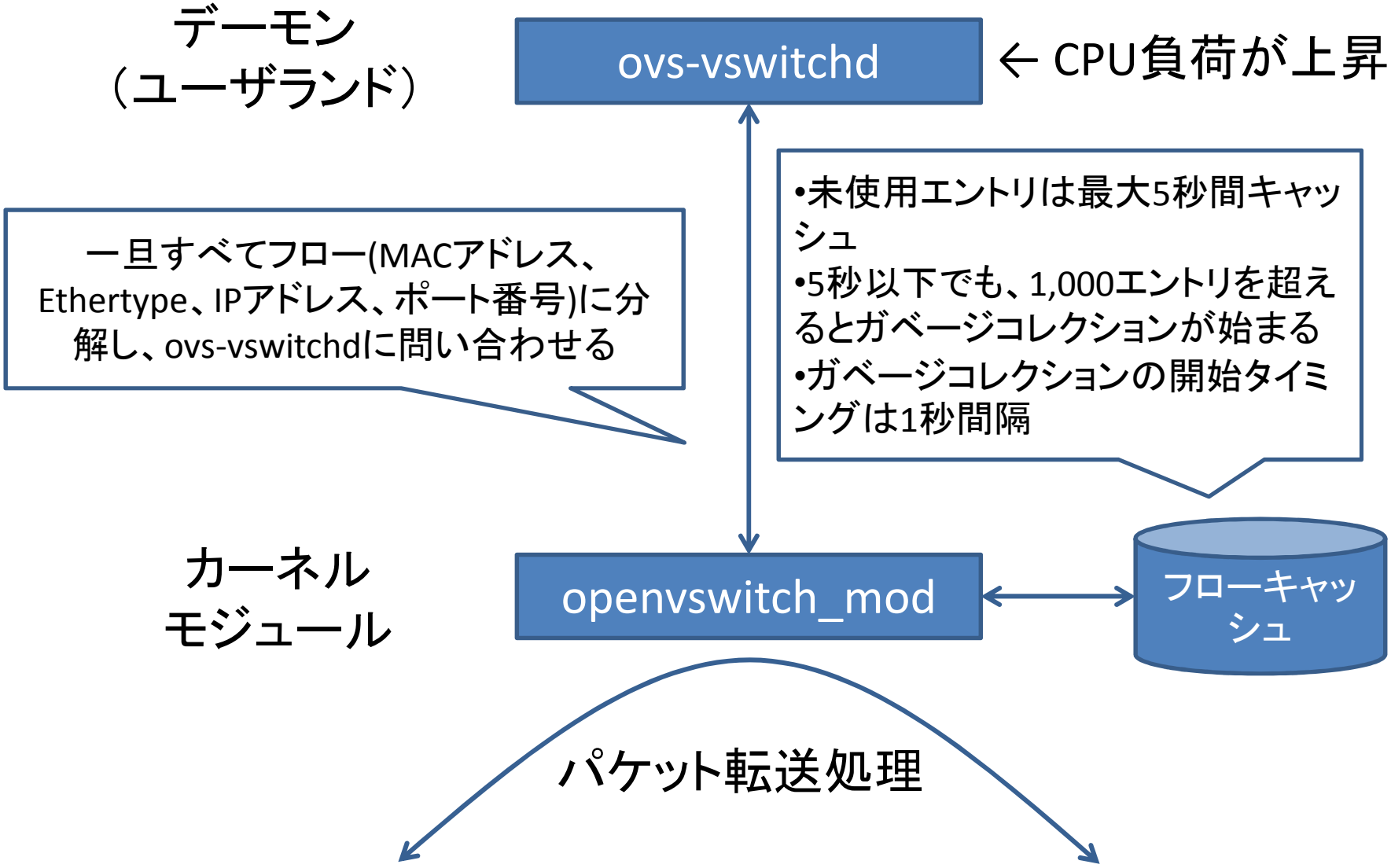
```
% ovs-vsctl set Interface tap2 ingress_policing_rate=10000
```

運用中に発生した問題

- Open vSwitchは、フローベース(Open Flow)のアーキテクチャ
- DoSアタックなど、大量のフローが発生すると、転送性能が極端に低下する
- 同一ホストの別のお客様のVMに影響が及ぶ
- フローキャッシュを見る方法

```
% ovs-dpctl dump-flows br0
in_port(2),eth(src=52:54:00:15:3e:a3,dst=52:54:00:b3:d4:57),eth_type(
0x0800),ipv4(src=220.205.27.17,dst=61.12.213.224,proto=17,tos=0),udp(
src=49766,dst=19968), packets:0, bytes:0, used:never, actions:1
in_port(2),eth(src=52:54:00:15:3e:a3,dst=52:54:00:b3:d4:57),eth_type(
0x0800),ipv4(src=200.254.249.181,dst=14.232.213.232,proto=17,tos=0),u
dp(src=54759,dst=10904), packets:0, bytes:0, used:never, actions:1
```

Open vSwitchのアーキテクチャ



とりあえずの対策

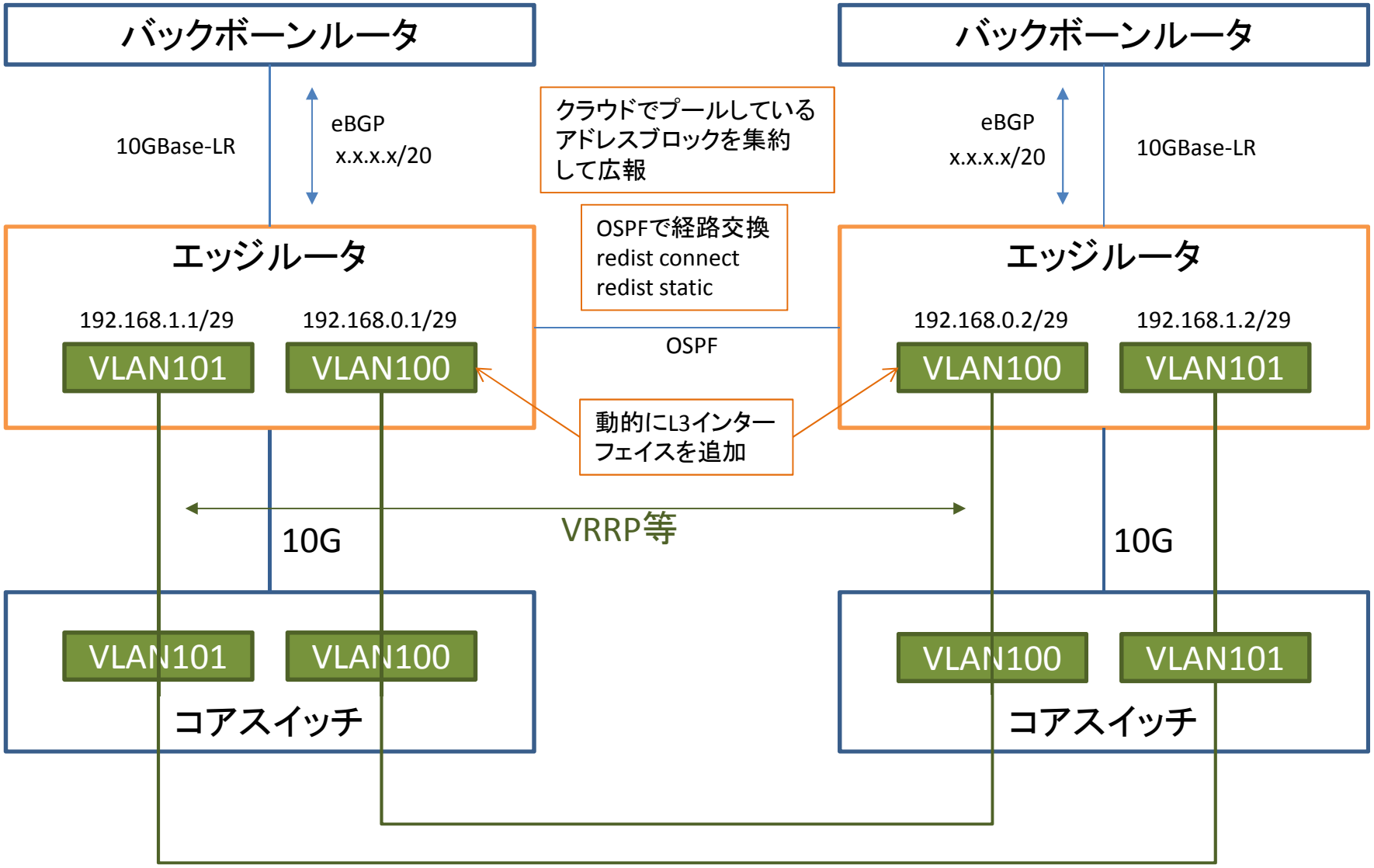
- Open vSwitch 1.1.0から1.2.1にバージョンアップ
- フローのセットアップ性能が50倍に
  (約800flows/sec → 約40,000flows/sec)
- バージョンアップに伴い、ガベージコレクション開始の閾値を変更可能に
- dp内フロー数の最大値が131,072なので、120,000に設定

```
% ovs-vsctl set bridge br0 other_config=flow-eviction-threshold=120000
```

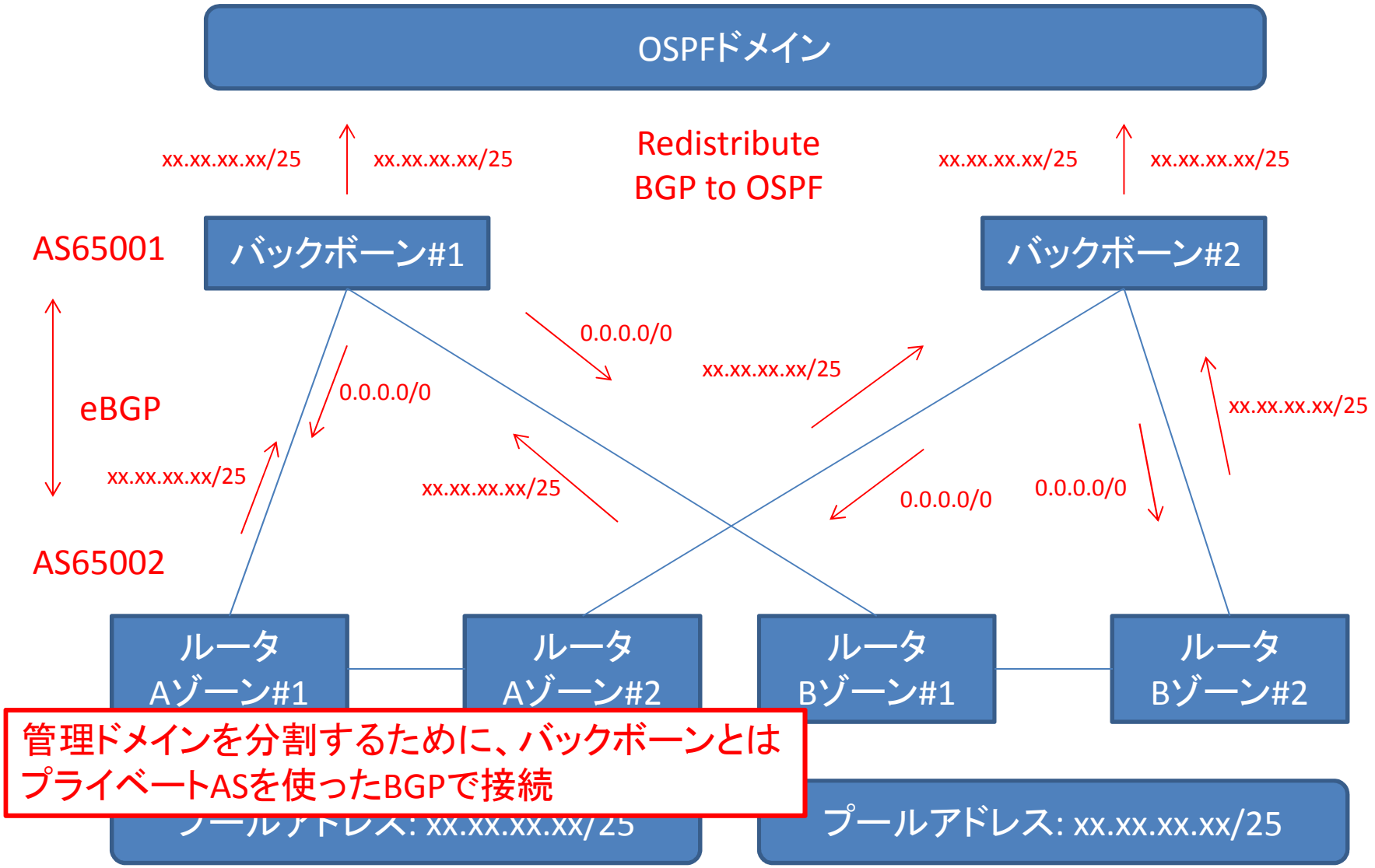
今後検討中の対策

- かなり緩和したが、根本的な解決策ではない。
- Open vSwitchがフローベースのアーキテクチャである以上、根本的な解決は難しそう。
- 定期的に仮想ポート毎のフロー数をモニタし、異常なフロー数が観測された場合、帯域制限を自動的に実施する機構の実装を検討。
- Open vSwitchにて、仮想ポート毎のフロー数制限を設定できるようになると吉か。

インターネット接続構成



ルーティング構成


OSPFドメイン
xx.xx.xx.xx/25
xx.xx.xx.xx/25
Redistribute
BGP to OSPF
xx.xx.xx.xx/25
xx.xx.xx.xx/25
AS65001
バックボーン#1
バックボーン#2
eBGP
0.0.0.0/0
xx.xx.xx.xx/25
0.0.0.0/0
xx.xx.xx.xx/25
xx.xx.xx.xx/25
xx.xx.xx.xx/25
0.0.0.0/0
0.0.0.0/0
AS65002
ルータ
Aゾーン#1
ルータ
Aゾーン#2
ルータ
Bゾーン#1
ルータ
Bゾーン#2
管理ドメインを分割するために、バックボーンとは
プライベートASを使ったBGPで接続
プールアドレス: xx.xx.xx.xx/25
プールアドレス: xx.xx.xx.xx/25

ルータの要件

- クラウドに配置した仮想サーバを、インターネットやVPN等の外部ネットワークと接続する。
- VLAN数: 500以上
- VRRPインスタンス数: 500以上
  - VRRP以外のプロトコルも可
- MACアドレス数: 8,000以上
- ARPエントリ数: 4,000以上
- プロトコル: OSPF、BGP

検証時に発生した問題

- VRRPインスタンスを大量に作成するとフラップ
- ルータ宛てDoSに弱い
- ARPのエントリを大量に持つと重くなる

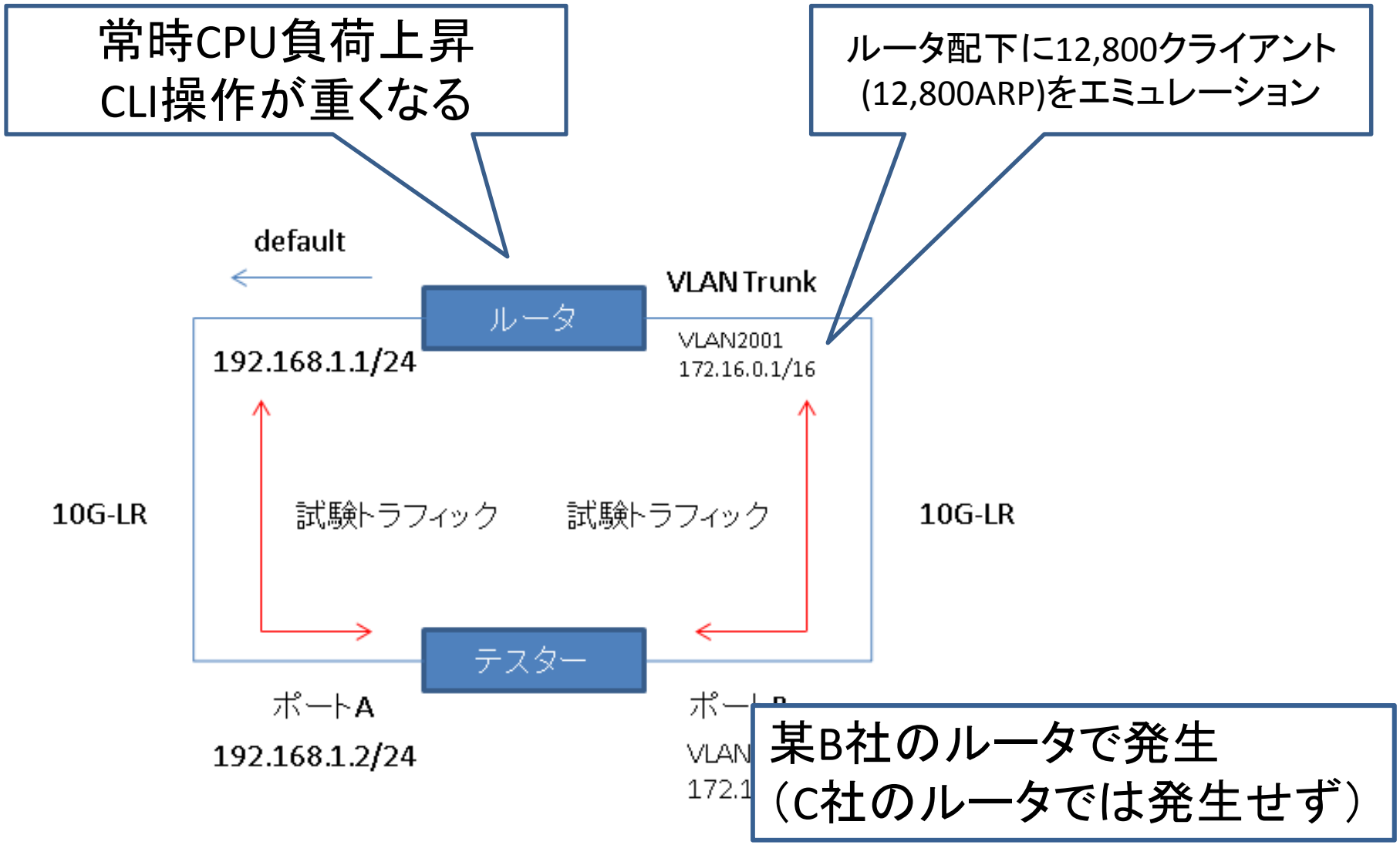
VRRPインスタンスがフラップ

- 某B社のルータで発生（C社のルータでは発生せず）
- 2,000VLANにIPアドレスを設定し、VRRPを構成
- → 毎秒 2,000パケットのVRRP Hello（マルチキャスト）
- バックアップ側で受け取れずに、Masterに昇格するVLANが続出
- Helloのインターバルを伸ばすことで解決（5秒/15秒など）

ルータ宛てDoSアタックに弱い

- 某B社のルータで発生（C社のルータでは発生せず）
- TCP 23番ポート、ルータ宛てのパケットを10Gショートパケット、ワイヤレートで送出
- ルータの操作ができなくなった(CLIが固まる)。
- ACLの設定で回避可能

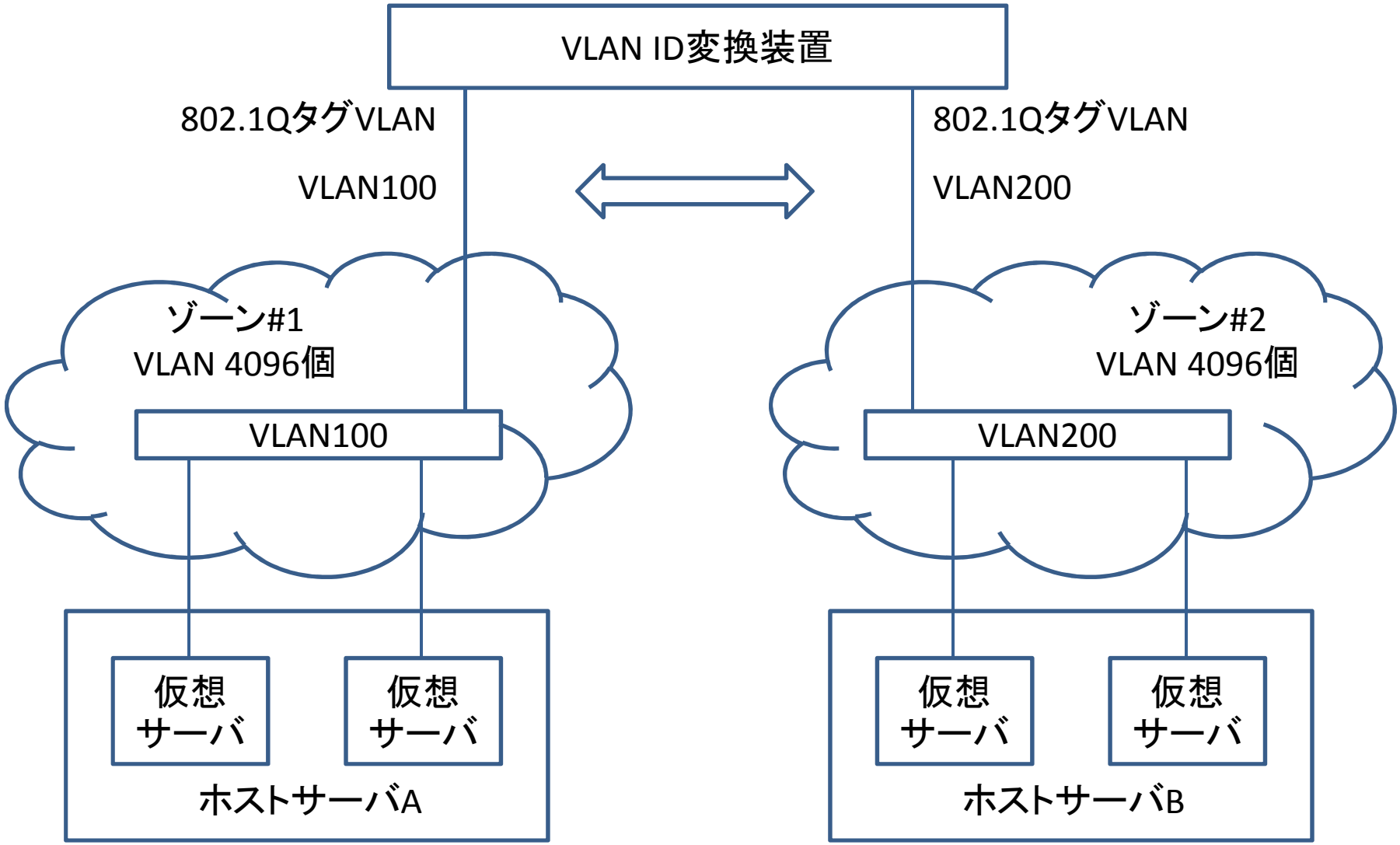
大量のARPエントリで重くなる



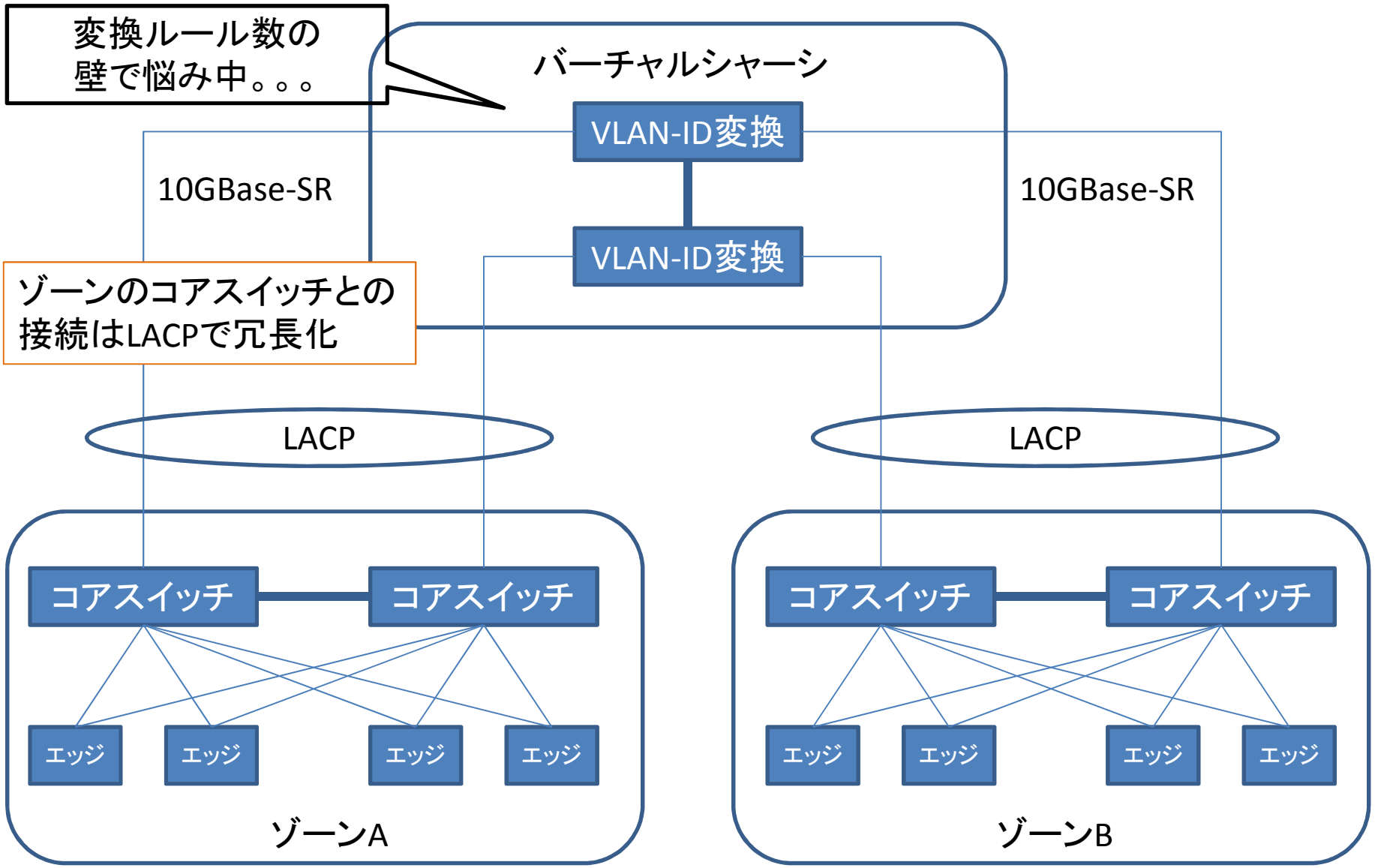
VLAN ID数制限の克服

- 802.1QタグVLANのVLAN IDは12bit(4096)
- 1ゾーンあたり最大4,096VLANしか使用できない(実際はもっと少ない) 。
- VLANが不足したら、別のゾーンを作成（ルータ、コアスイッチを新設）
- 別ゾーンに収容されたお客様が相互接続したい場合はどうするか？
- ゾーン間のL2接続を行う「ブリッジ」機能を実装

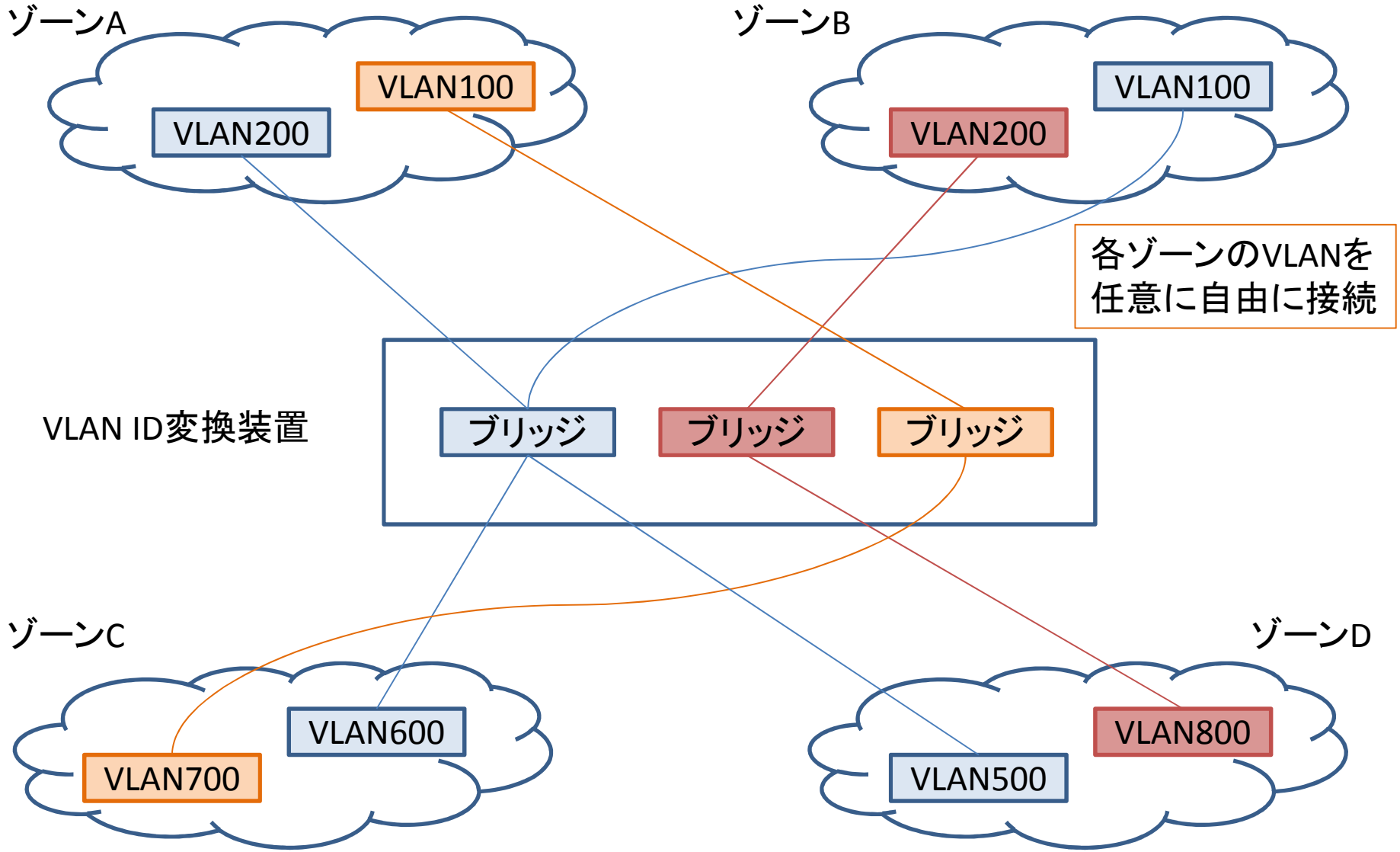
ブリッジのイメージ


VLAN ID変換装置
802.1Qタグ VLAN
VLAN100
802.1Qタグ VLAN
VLAN200
ゾーン#1
VLAN 4096個
VLAN100
ゾーン#2
VLAN 4096個
VLAN200
仮想
サーバ
仮想
サーバ
ホストサーバA
仮想
サーバ
仮想
サーバ
ホストサーバB

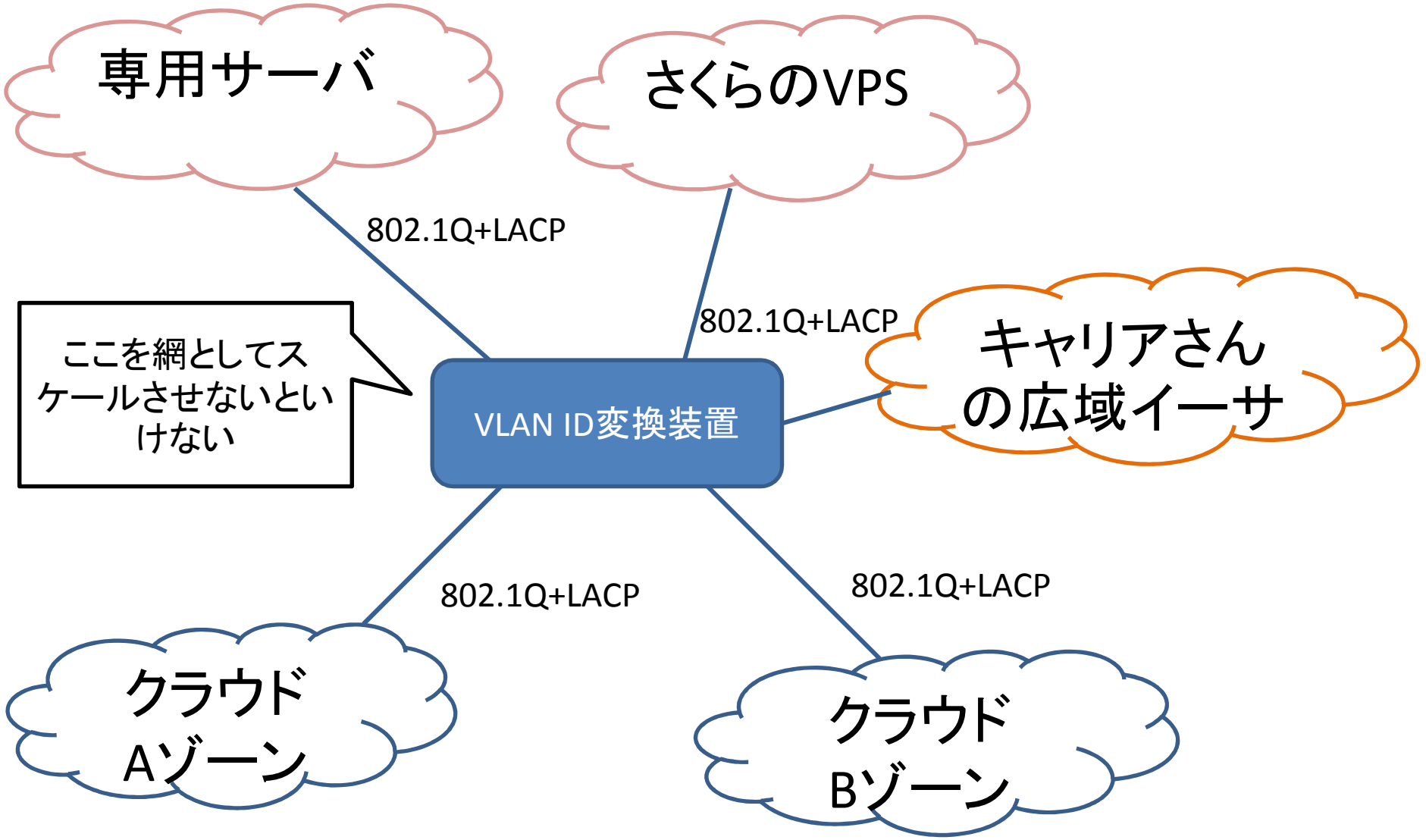
現在の構成


変換ルール数の
壁で悩み中。。。
バーチャルシャーシ
VLAN-ID変換
VLAN-ID変換
10GBase-SR
10GBase-SR
ゾーンのコアスイッチとの
接続はLACPで冗長化
LACP
LACP
コアスイッチ
コアスイッチ
コアスイッチ
コアスイッチ
エッジ
エッジ
エッジ
エッジ
エッジ
エッジ
エッジ
エッジ
ゾーンA
ゾーンB

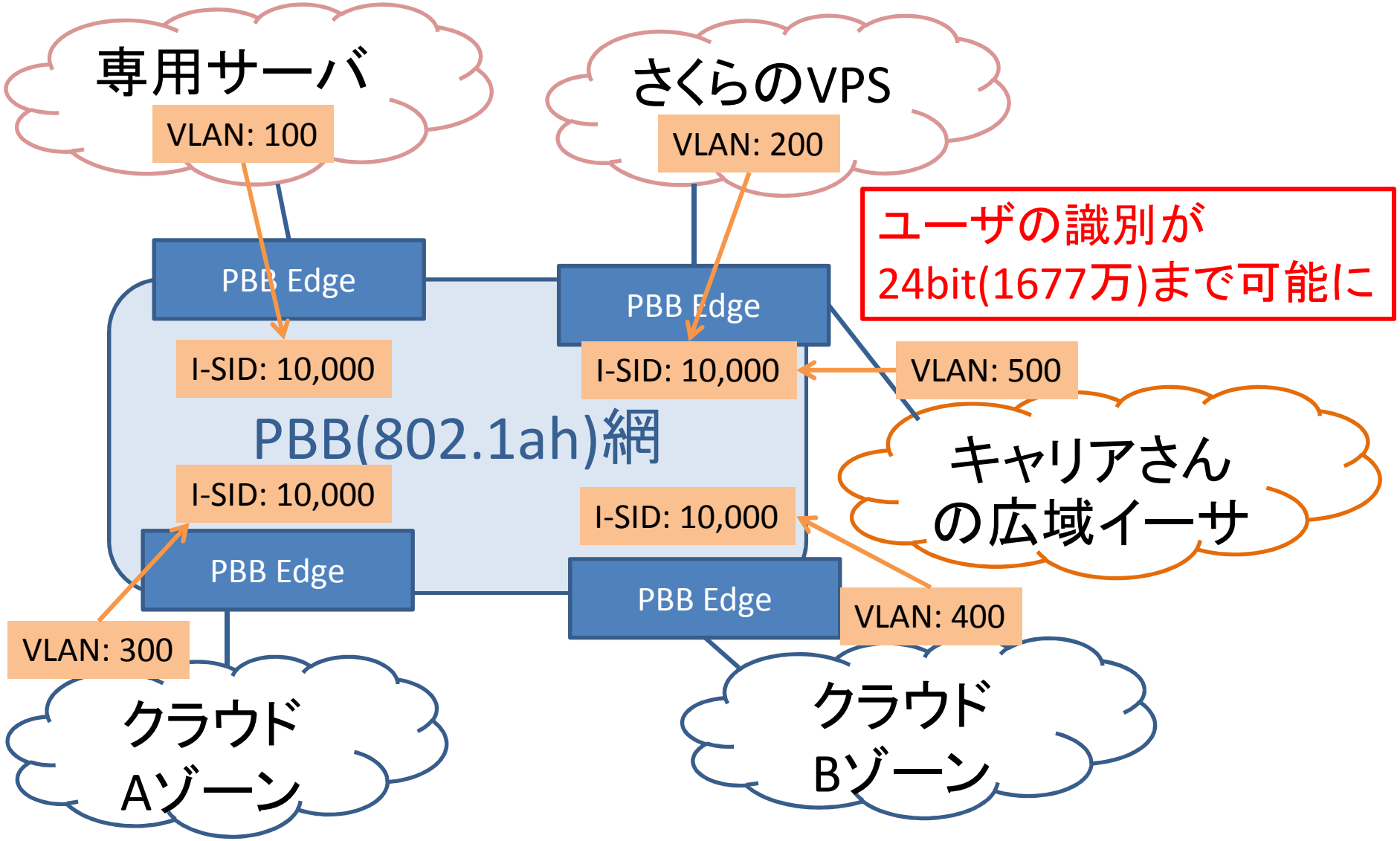
マルチポイントでの相互接続



将来的には・・・


専用サーバ
さくらのVPS
802.1Q+LACP
802.1Q+LACP
ここを網としてスケールさせないといけない
VLAN ID変換装置
キャリアさんの広域イーサ
802.1Q+LACP
802.1Q+LACP
クラウドAゾーン
クラウドBゾーン

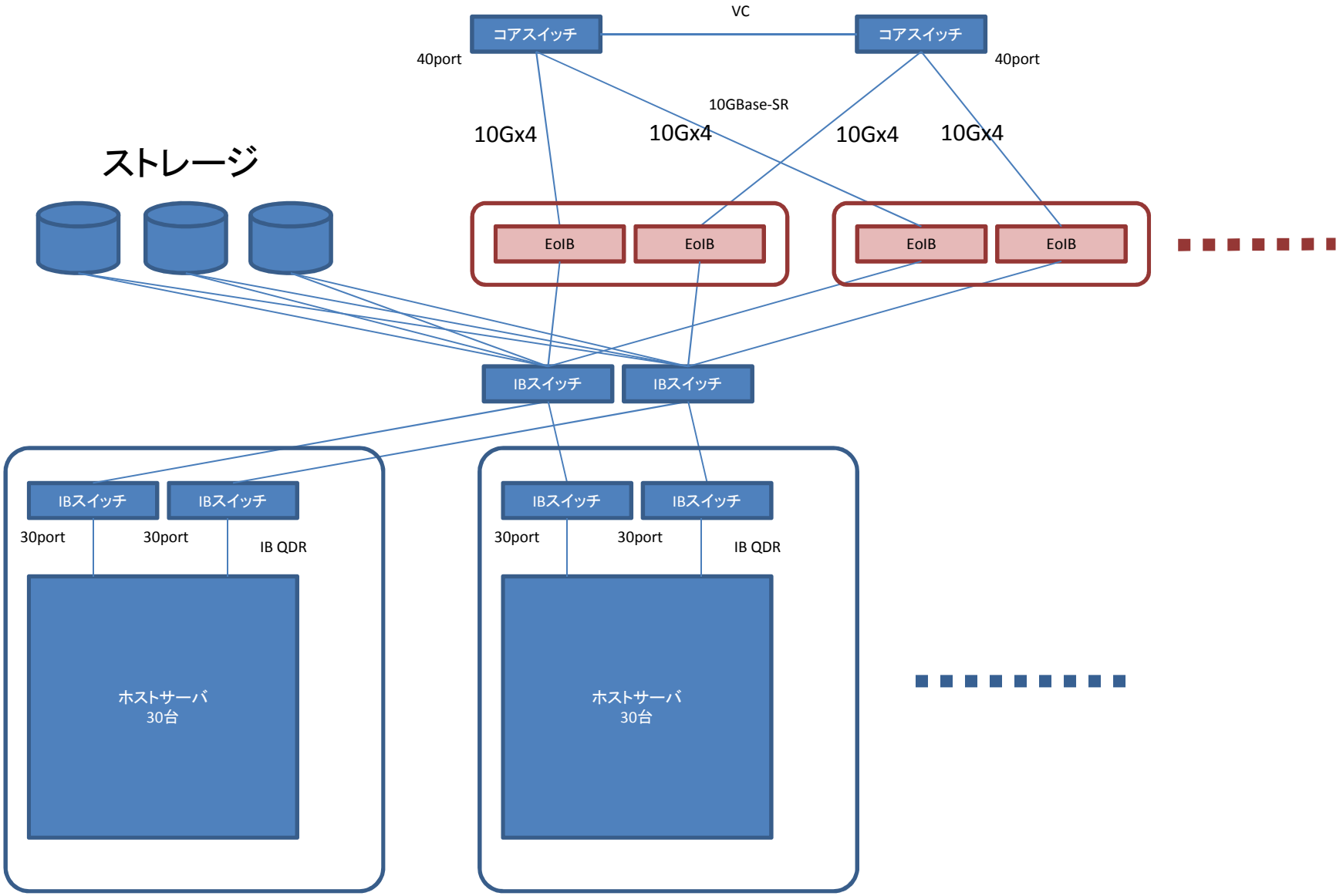
PBBとか、EoEとか


専用サーバ
VLAN: 100
さくらのVPS
VLAN: 200
ユーザの識別が
24bit(1677万)まで可能に
PBB Edge
PBB Edge
I-SID: 10,000
I-SID: 10,000
VLAN: 500
PBB(802.1ah)網
キャリアさん
の広域イーサ
I-SID: 10,000
I-SID: 10,000
PBB Edge
PBB Edge
VLAN: 400
VLAN: 300
クラウド
Aゾーン
クラウド
Bゾーン

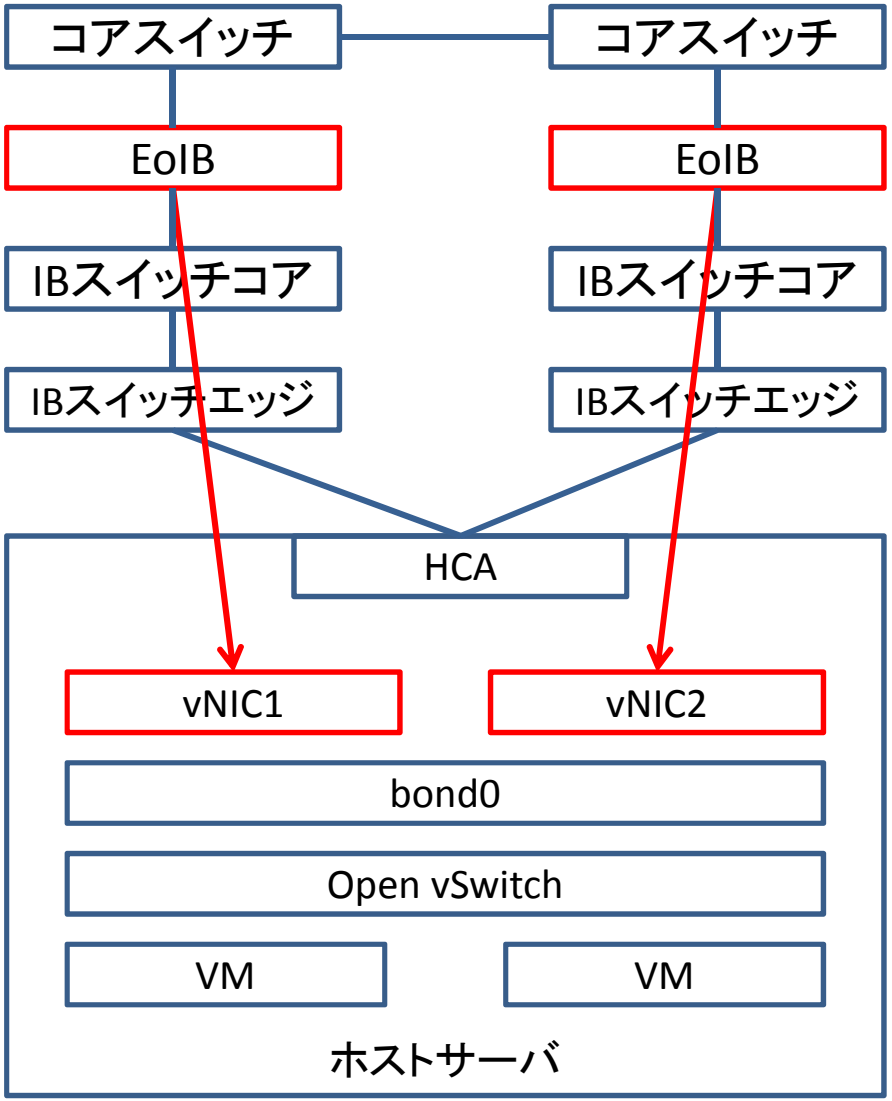
今後のネットワークプラン

- Ether over InfiniBand(石狩から導入)
  - 配線数の削減
  - コストそのまま、帯域増加
- 完全トンネル方式(２～３年後？)
  - VM間のL2通信を、ホストサーバにてIPトンネル
  - バックボーンをL3で組める
  - スケールする構成に

Ether over InfiniBand



ホスト内の構成



802.1Q
VLAN1～4096

802.1Q
VLAN1～4096

802.1Q
VLAN1～4096

完全トンネル方式

- Open vSwitch
  - GREでEtherをIPで飛ばす機能が載っている
  - 某社が頑張っているらしい・・・
- 富士通研究所さんの事例
  - http://www.ieice.org/ken/paper/20100805B000/
  - Linux GRETAPを使った方式
- 新しいプロトコル仕様(IETFに提案されている)
  - http://tools.ietf.org/html/draft-sridharan-virtualization-nvgre-00
  - http://tools.ietf.org/html/draft-mahalingam-dutt-dcops-vxlan-00
  - どちらも、ユーザの識別は24bit、IP上にオーバレイする

Etherのスイッチを使わずに、Etherのネットワークが組めるようになる時代が来る！

おまけ：仮想アプライアンス

- 単純なネットワーク到達性だけでは機能不足。
- 従来のハードウェアアプライアンスにかわるものを、クラウドで提供する。

種別	ソフトウェア
仮想ルータ	Vyatta SEIL/x86（実装中）
仮想ファイアウォール	pfSense その他実装中
仮想ロードバランサ	実装中
仮想ファイルサーバ	実装中

まとめ

- クラウドのネットワークはL2との戦い
  - L2到達性が必要で、L3到達性ではダメだと認識
  - MACアドレス、VLAN数、スケーラビリティ
- 今後、どちらかの方向に進まないといけない
  - L2機器でがんばる
  - ホストサーバでトンネルする
- 他サービスとの相互接続も検討中
- さくらのクラウドでは、既存の全物理サービスの置き換えを目指したい
- 仮想アプライアンスの充実、QoSの強化を目指す