

「さくらのクラウド」におけるL2ネットワークの課題

さくらインターネット株式会社 研究所 大久保 修一

- 2003年4月 さくらインターネット入社
 - 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

- IaaS
- 来月(2011/11)中旬リリース予定

- 主な機能
 - 仮想サーバ
 - 仮想ディスク
 - 仮想スイッチ
 - 仮想アプリケーション



The screenshot displays the SAKURA Internet Cloud Console interface. At the top, the header includes the SAKURA Internet logo, the text "Cloud Console", and the user information "root@nba56849 設定 ログアウト". Below the header is a navigation bar with icons for "大阪β", "ステータス", "サーバ", "ディスク", "スイッチ", and "テンプレート", along with a search box labeled "検索".

The main content area is titled "スイッチリスト" (Switch List). On the left, a sidebar menu shows "スイッチリスト" and "新規スイッチ" (New Switch). Below this, a list of switches is shown: "os1a" (大阪βゾーンA), "local-switch", "local-switch2", "os1b" (大阪βゾーンB), and "local-switch10".

The main panel displays details for "os1a (大阪βゾーンA)". It contains two switch cards:

- local-switch**: ローカルネットワーク, 接続数: 4
- local-switch2**: ローカルネットワーク, 接続数: 4

Below this, the section "os1b (大阪βゾーンB)" contains one switch card:

- local-switch10**: ローカルネットワーク, 接続数: 5

The section "ブリッジ一覧" (Bridge List) is currently empty. A blue information banner at the bottom states: "このリージョンにはブリッジが作成されていません" (No bridges are created in this region).

SAKURA Internet Cloud Console

root@nba56849 設定 ログアウト

大阪B ステータス サーバ ディスク **スイッチ** テンプレート 検索

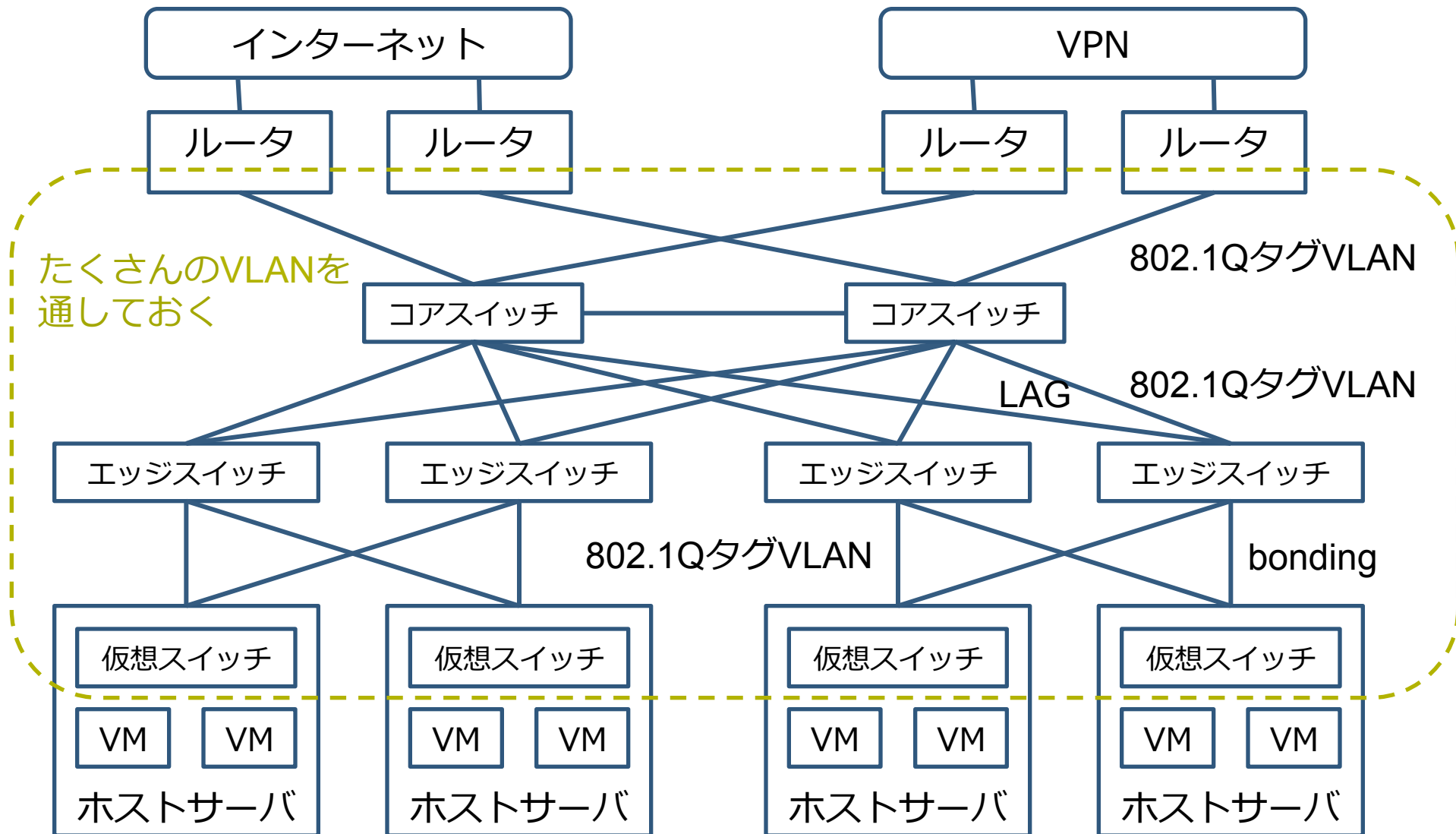
スイッチリスト
+ 新規スイッチ

- os1a
- local-switch**
- local-switch2
- os1b
- local-switch10

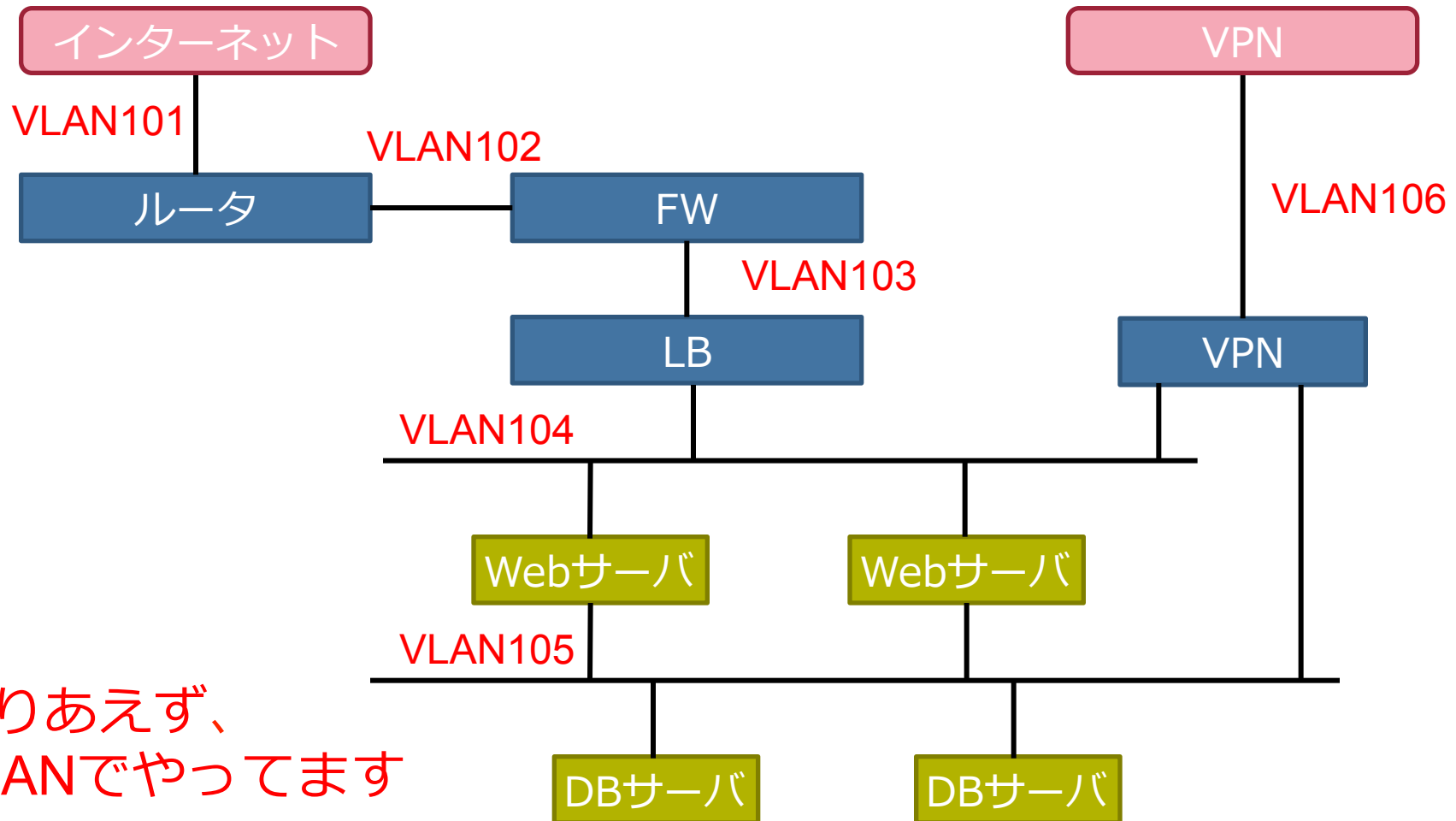
マップビュー » local-switch

戻る | 編集 | 更新

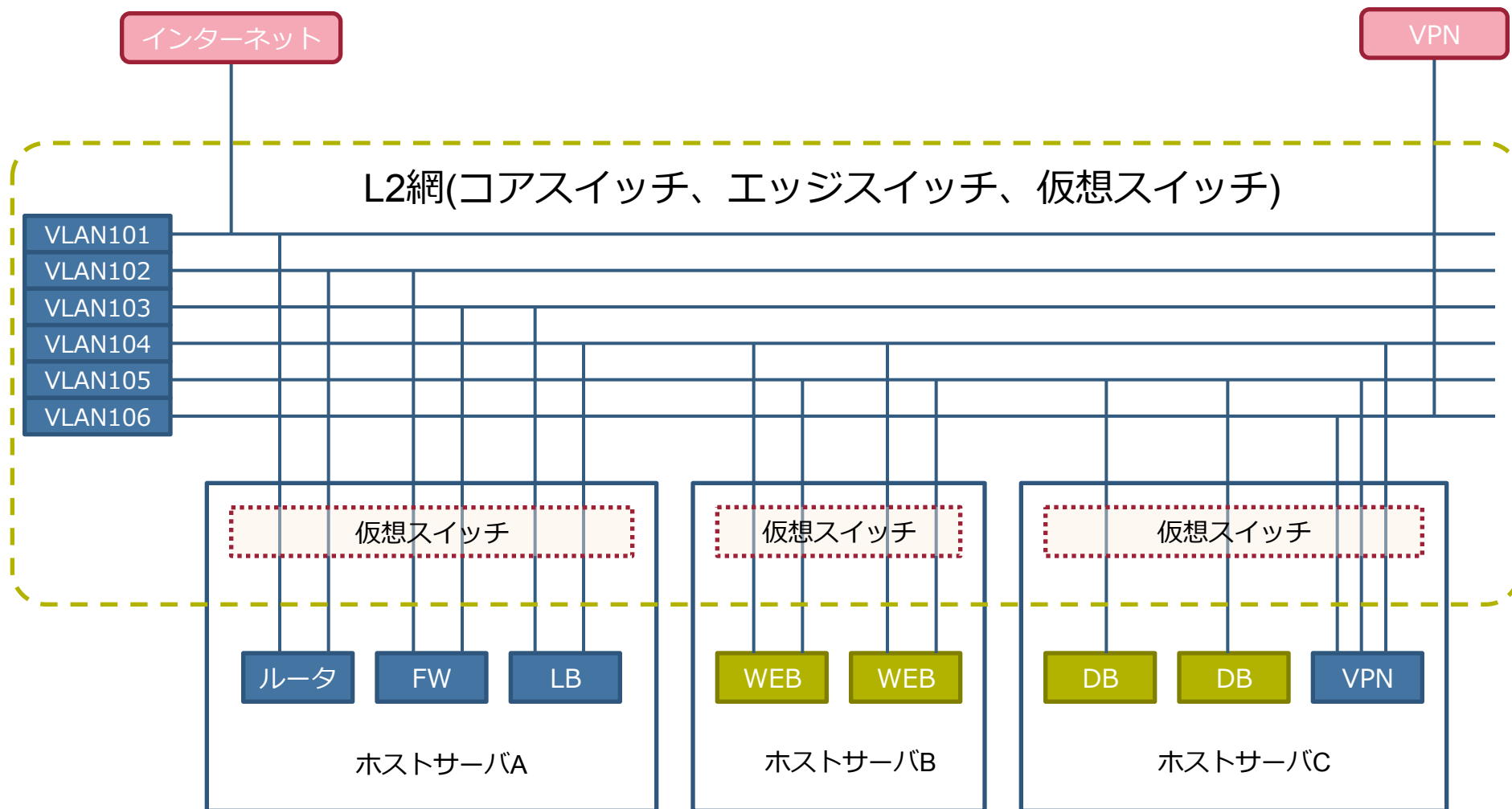
```
graph TD; Internet[共有セグメントインターネット] --- pfsense; Internet --- vyatta; Internet --- slb1; pfsense --- local-switch; vyatta --- local-switch; slb1 --- local-switch; local-switch --- Ubuntu[Ubuntu 11....]; local-switch --- seil; local-switch --- local-switch2[local-switch...]; local-switch2 --- gentoo-test; local-switch2 --- gentoo-32;
```



これを、IaaSインフラ上に展開してみる



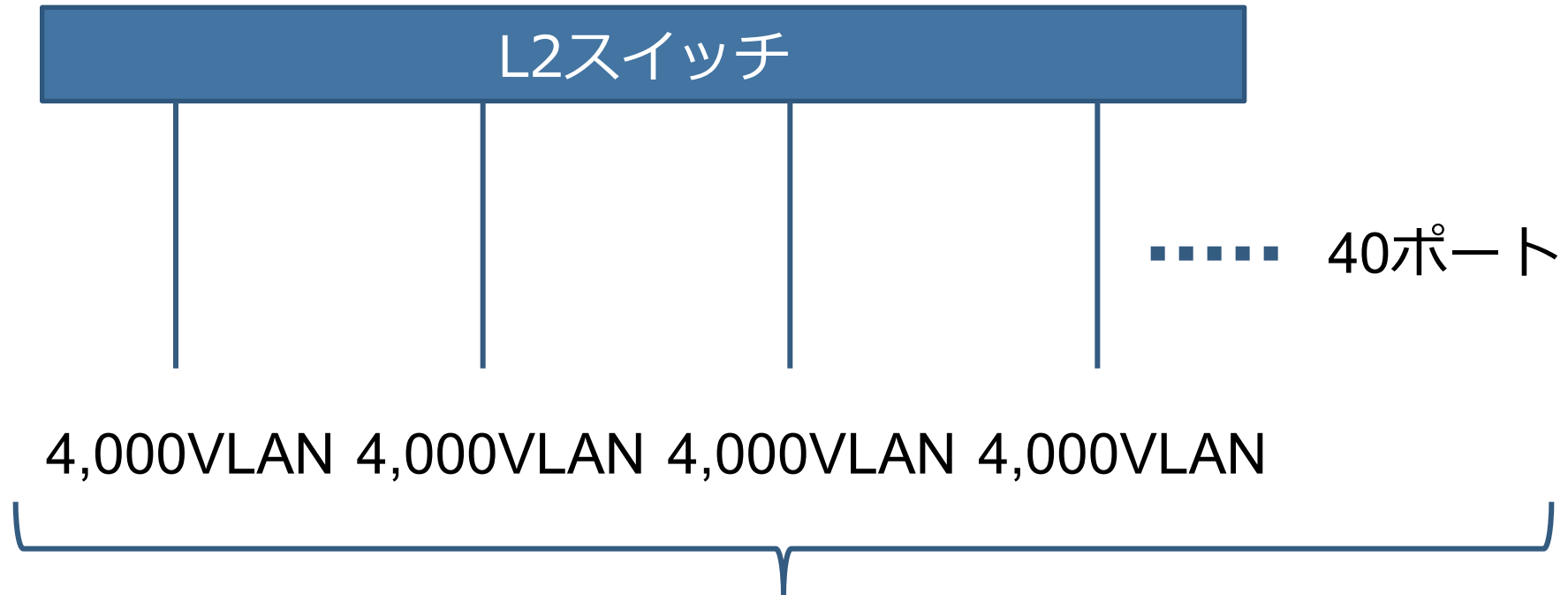
とりあえず、
VLANでやっています



- 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 (想定値)
 - MACアドレス数: 16,000 (8,000×2)
-
- ルータ、コアスイッチ、エッジスイッチ、仮想スイッチに収容できる能力が必要。

- 実際に使用できるVLAN数が意外と少ない
- ショートパケットでワイヤレート出ない
- 実際に収容できるMACアドレス数が少ない
- configが長くなってオペレーションに支障がある

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L2スイッチもある

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

Etherのフレームフォーマット

IFG (12)	プリアン ブル(8)	宛先MAC (6)	送信元 MAC(6)	タイプ(2)	ペイロード (46-1500)	FCS(4)
-------------	---------------	--------------	---------------	--------	--------------------	--------

最小 64Bytes

最小 84Bytes

最小フレーム長でワイヤレート出るには？

$$14.88\text{Mpps} = 10,000,000,000 / 8 / 84$$

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。

```
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
```

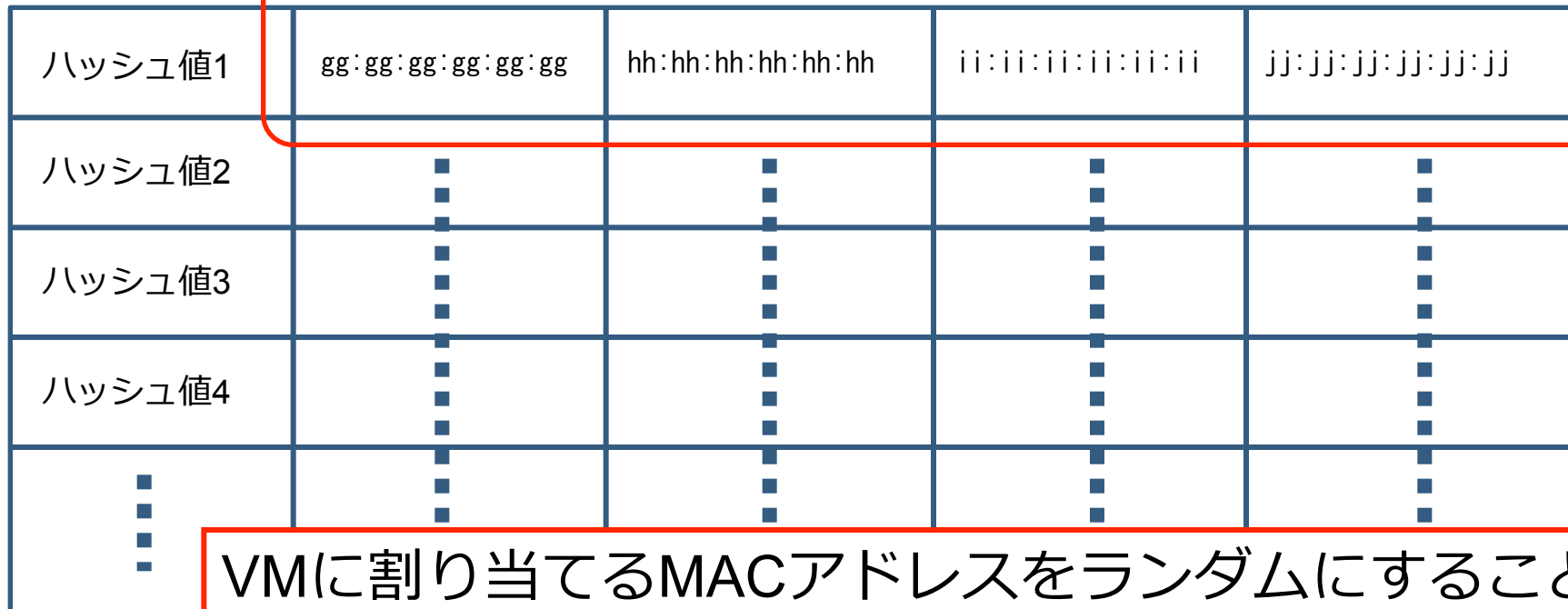
1VLAN毎に設定が必要

VLAN設定をまとめられるとうれしい

MACアドレスの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



ハッシュ値1	gg:gg:gg:gg:gg:gg	hh:hh:hh:hh:hh:hh	ii:ii:ii:ii:ii:ii	jj:jj:jj:jj:jj:jj
ハッシュ値2	⋮	⋮	⋮	⋮
ハッシュ値3	⋮	⋮	⋮	⋮
ハッシュ値4	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮

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

- スペックでは32,000のMACアドレステーブル
- ただし、1VLAN設定すると、3エントリ消費
- 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。。。。

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



Here are the results of your search through the public section of the IEEE Standards OUI database report for **9C-A3-BA**:

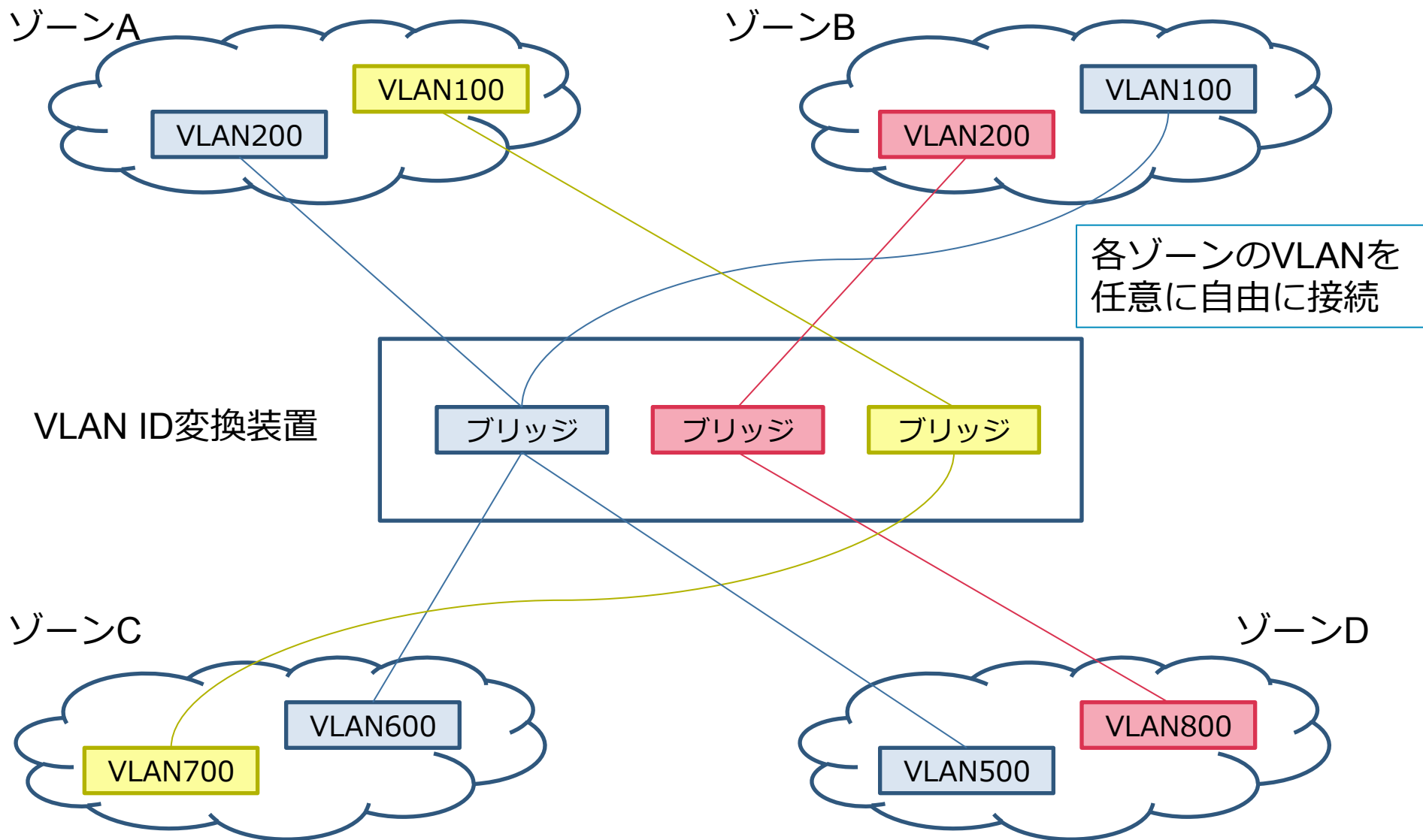
9C-A3-BA	(hex)	SAKURA Internet Inc.
9CA3BA	(base 16)	SAKURA Internet Inc. 7-20-1 Nishi-shinjuku Shinjuku-ku Tokyo 1600023 JAPAN

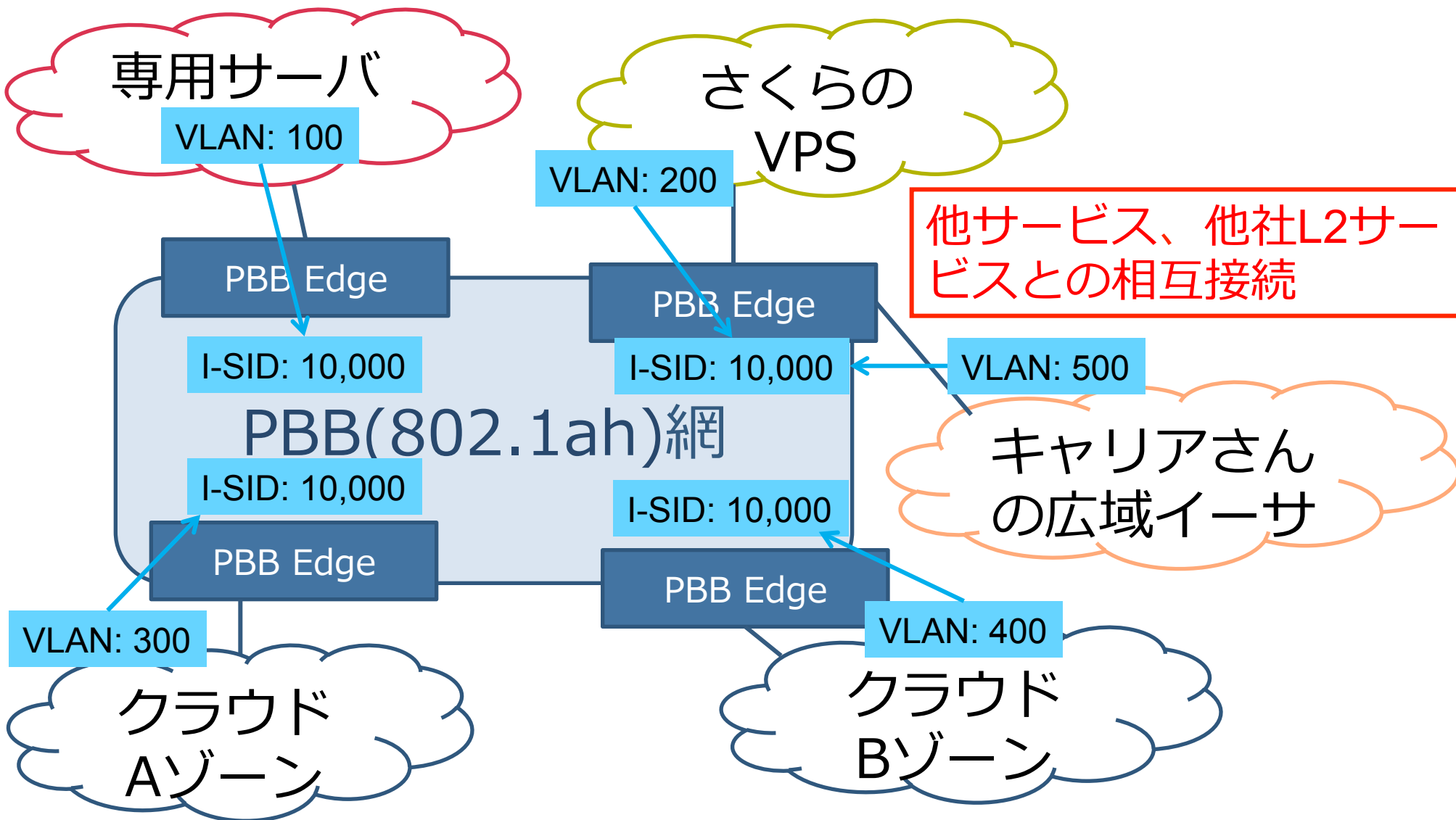
Your attention is called to the fact that the firms and numbers listed may not always be obvious in product implementation. Some manufacture and others include registered firms' OUIs in their products.

[\[IEEE Standards Home Page\]](#) -- [\[Search\]](#) -- [\[E-mail to Staff\]](#)
[Copyright © 2011 IEEE](#)

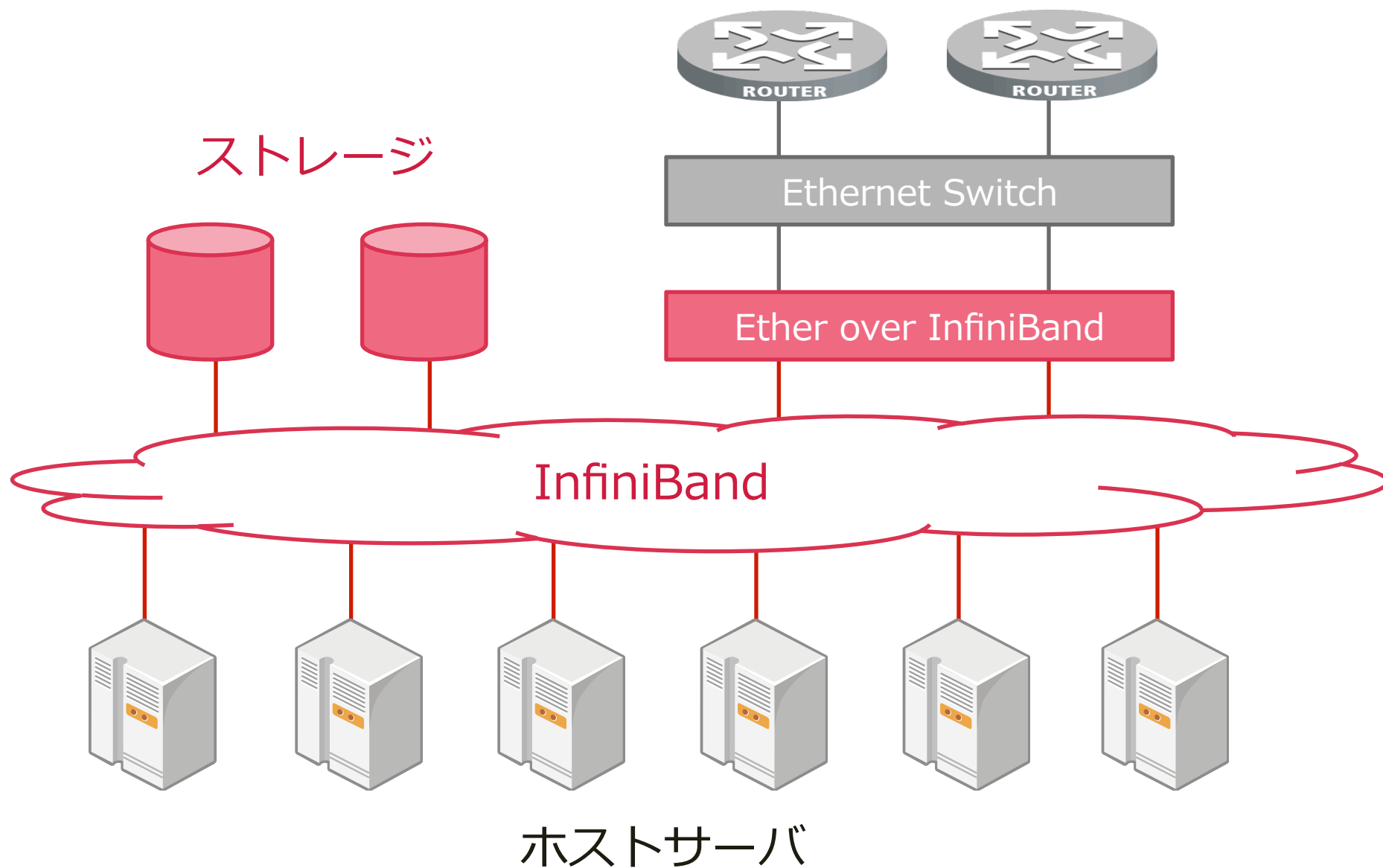
他のクラウドや他の物理ネットワークとの
L2レベルでの相互接続も問題なし！

- 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





- 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
 - VLAN数、MACアドレス数の制限がほぼなくなる
 - スケールする構成に



- 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 GRE-TAP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！

- 現在はVLANを用いて実装している。
- VLAN数とMACアドレス数がネックになる。
- VLAN数の制限については「ゾーン」という単位で網を分割している。
- ゾーン間をVLAN IDを変換する「ブリッジ」という機能で接続できる。
- 将来はトンネル方式に移行したい。