QinQとVLANを変換するサーバで
トラフィックを運んでみた

JANOG39 Meeing in Kanazawa

さくらインターネット株式会社
伊東 宏起

SAKURA internet

自己紹介

伊東 宏起

- **所属**

  さくらインターネット株式会社
  技術本部 ミドルウェアグループ

さくらのVPSチームプロデューサー

- JANOGとわたし
  - JANOG37 Meeting Nagoya PC
  - JANOG38 Meeting Okinawa PC/登壇
    - そんなMQで大丈夫か? ーより良いMQの使い所を考えるー

Agenda

- プログラムの概要と目的
- L2NW 接続サービスの概要
- QinQ VLAN 変換サーバ version1
  - 変換サーバ version1 の実装
  - 変換サーバ version1 冗長化構成
- QinQ VLAN 変換サーバ version2
  - 変換サーバ version2 の実装と冗長化
- まとめ
- 会場への質問と議論

SAKURA internet

プログラムの概要と目的

プログラムの概要

・任意の QinQ と VLAN ヘッダーを相互変換し、
　トラフィックを運ぶ Linux サーバを作り・運用した話

プログラムの目的

- 「Linux サーバでトラフィックを変換し運ぶ」とりくみを
  共有をさせて頂き、新たな NW を考えるきっかけを作りたい
- 巨大化する L2NW に対する新しいアプローチを見つける

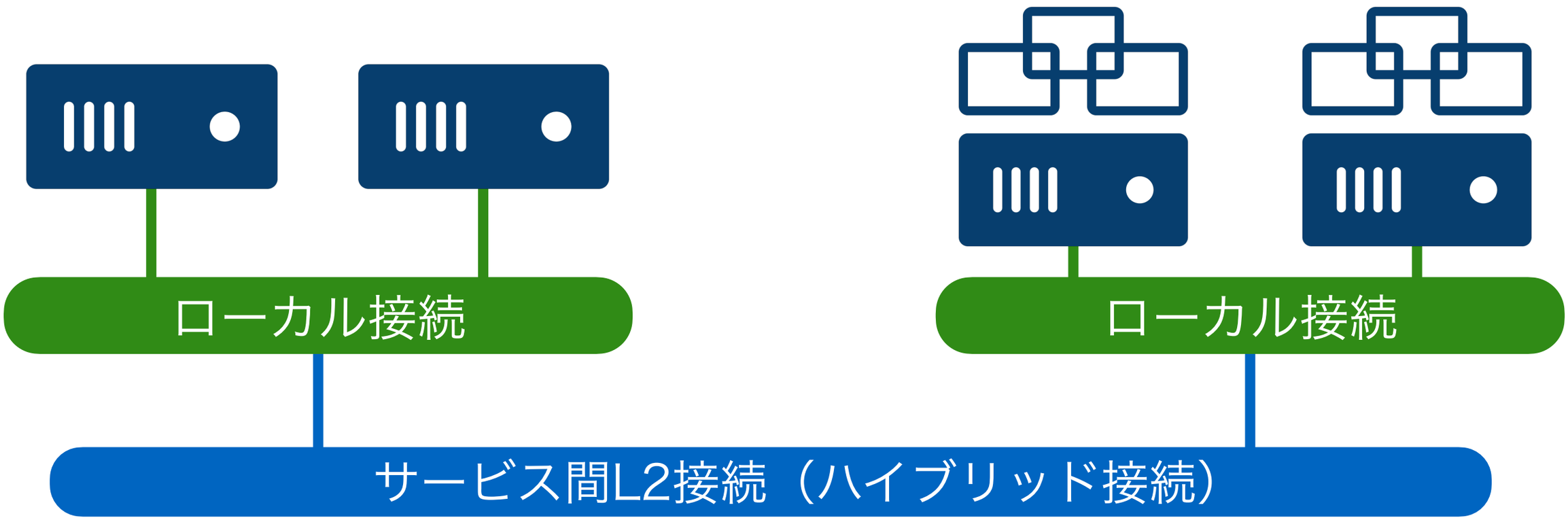
SAKURA internet

L2NW接続サービスの概要

L2NW接続サービスの概要

・VPS/クラウド/専用サーバなど仮想サーバから物理サーバまで多岐にわたったホスティングサービスを展開

- サービス内でサーバを L2 接続するローカル接続オプション
- 異なるサービスのサーバをL2接続するハイブリッド接続サービス



SAKURA internet

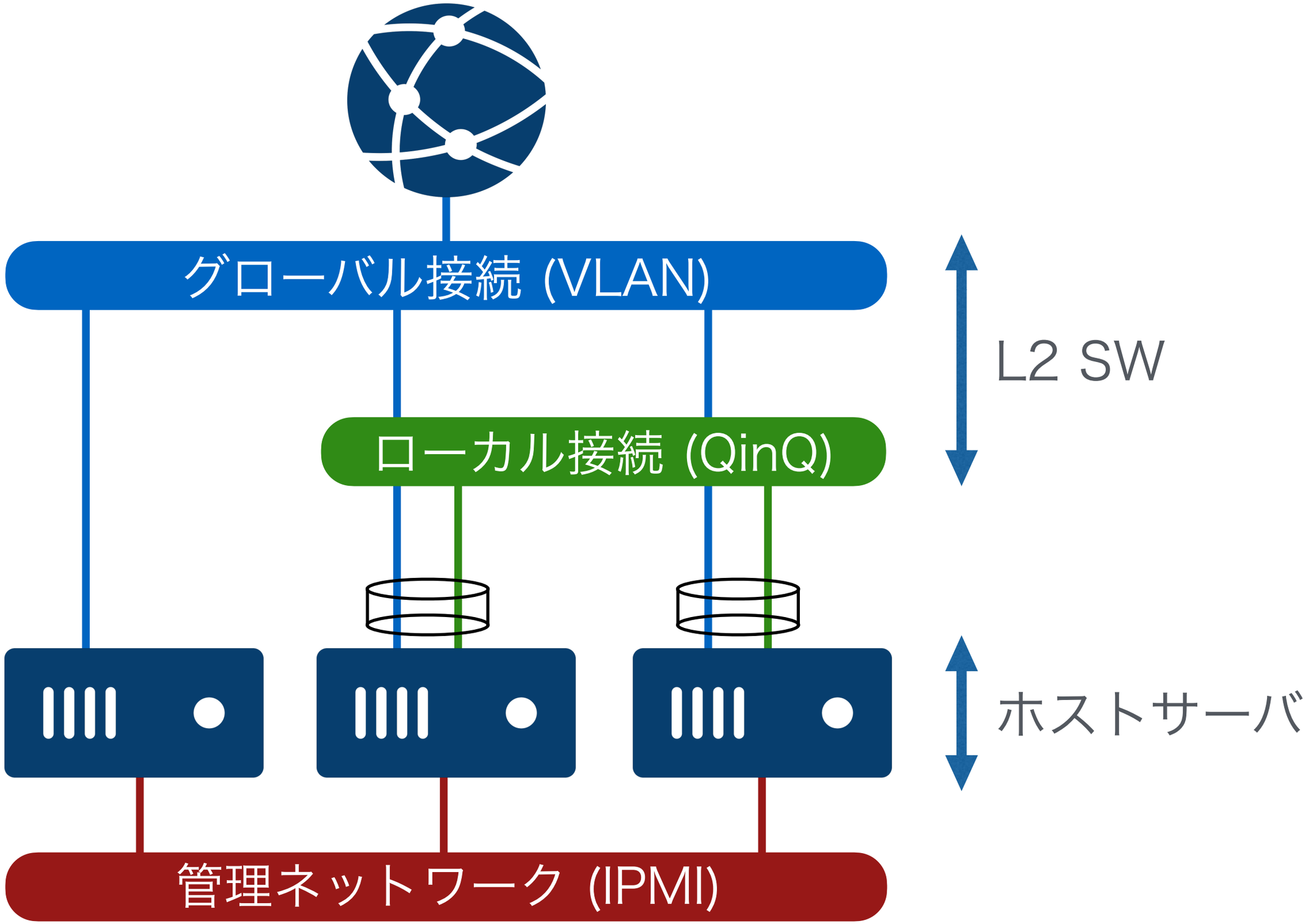
QinQ VLAN 変換サーバ version1

QinQ VLAN 変換サーバ version1

- さくらのVPS（仮想サーバ）とベアメタルプラン（物理サーバ）のローカル接続の為に開発
- さくらのVPSで QinQ/ベアメタルプランで VLAN を利用し、ローカル接続を実装
- これらを接続するために QinQ と VLAN を相互変換する「何か」が必要、という課題がきっかけとなり実装
- 現在1ペア2台が本番稼働中

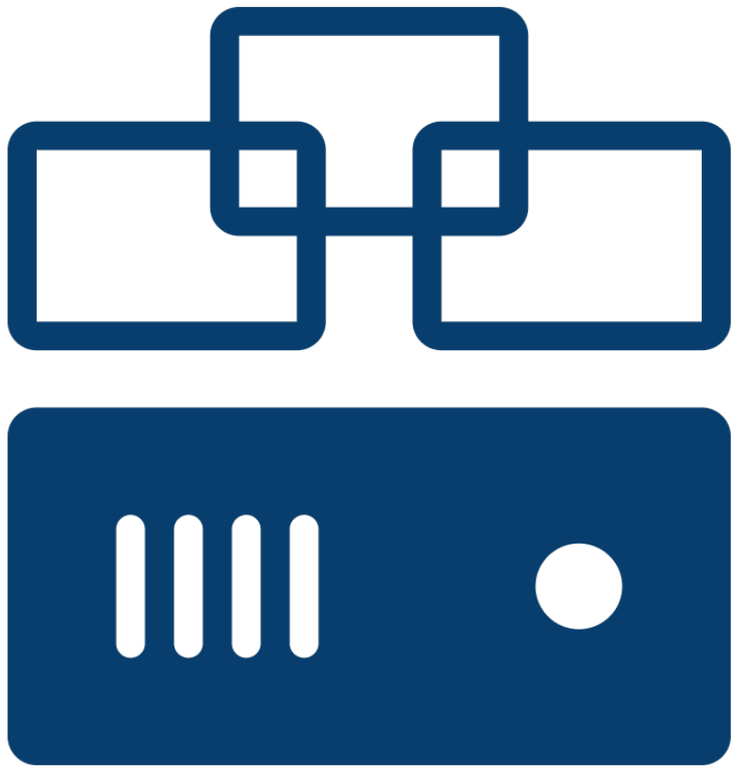
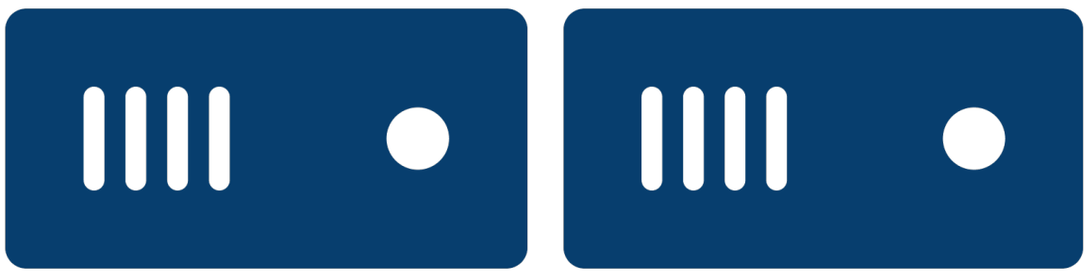
SAKURA internet

さくらのVPSのローカル接続の実装



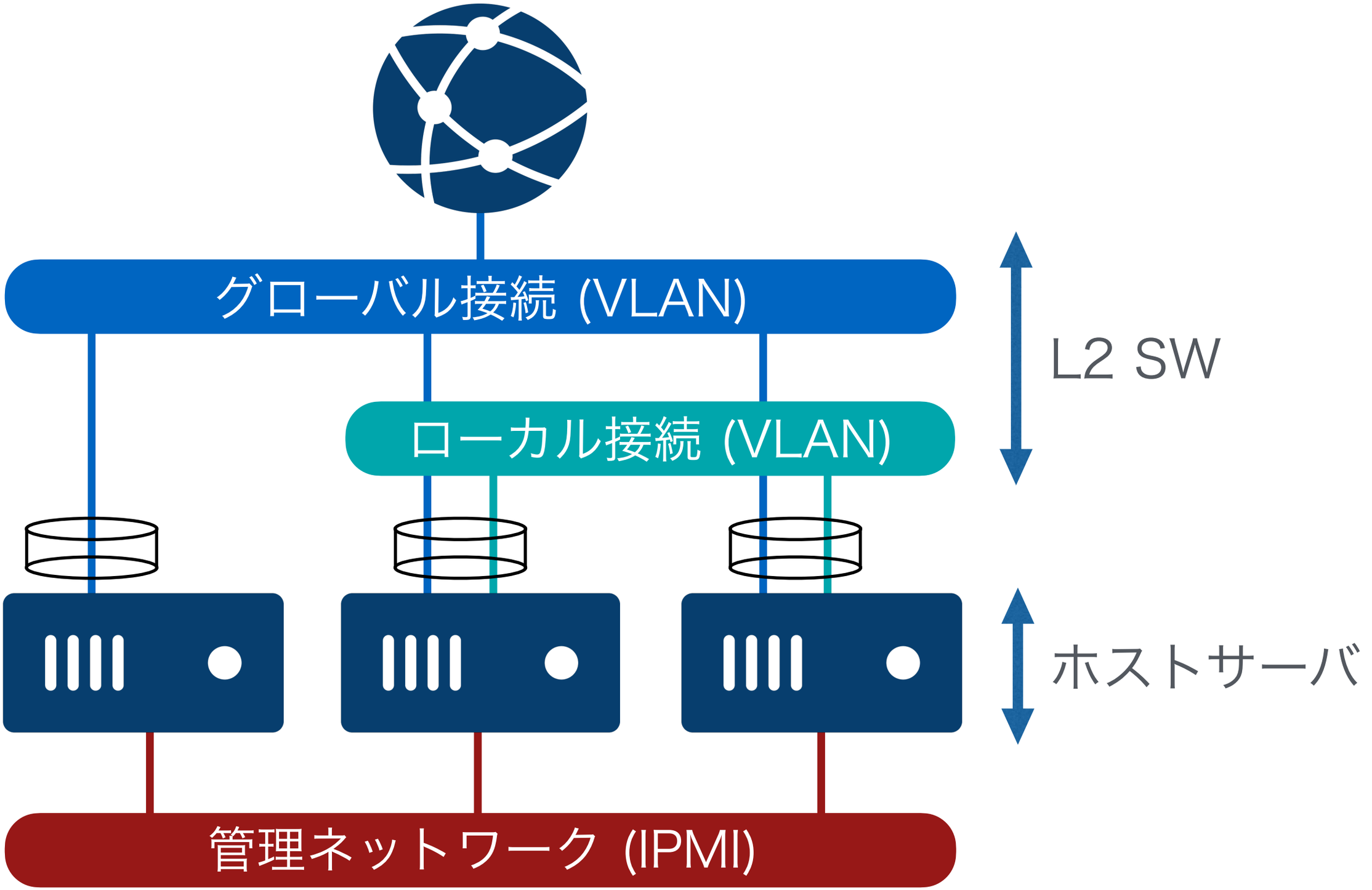
SAKURA internet

さくらのVPS ベアメタルプラン

VPS	ベアメタル
ホストサーバー上に仮想化基盤を構築 仮想化基盤上でVMを作成して提供	ホストサーバをそのまま提供 仮想化基盤は無し
	

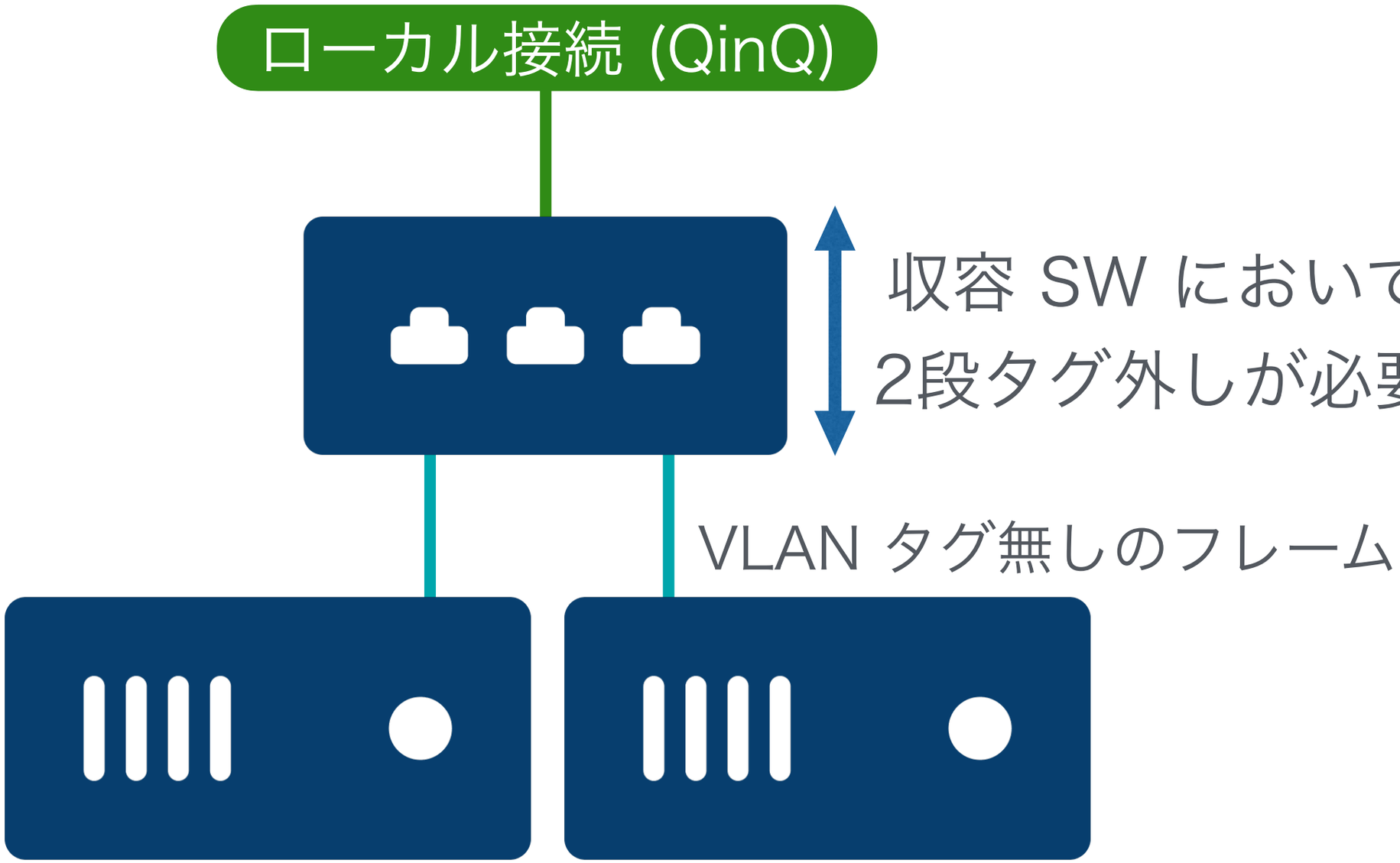
SAKURA internet

さくらのVPS ベアメタルプランのローカル接続の実装

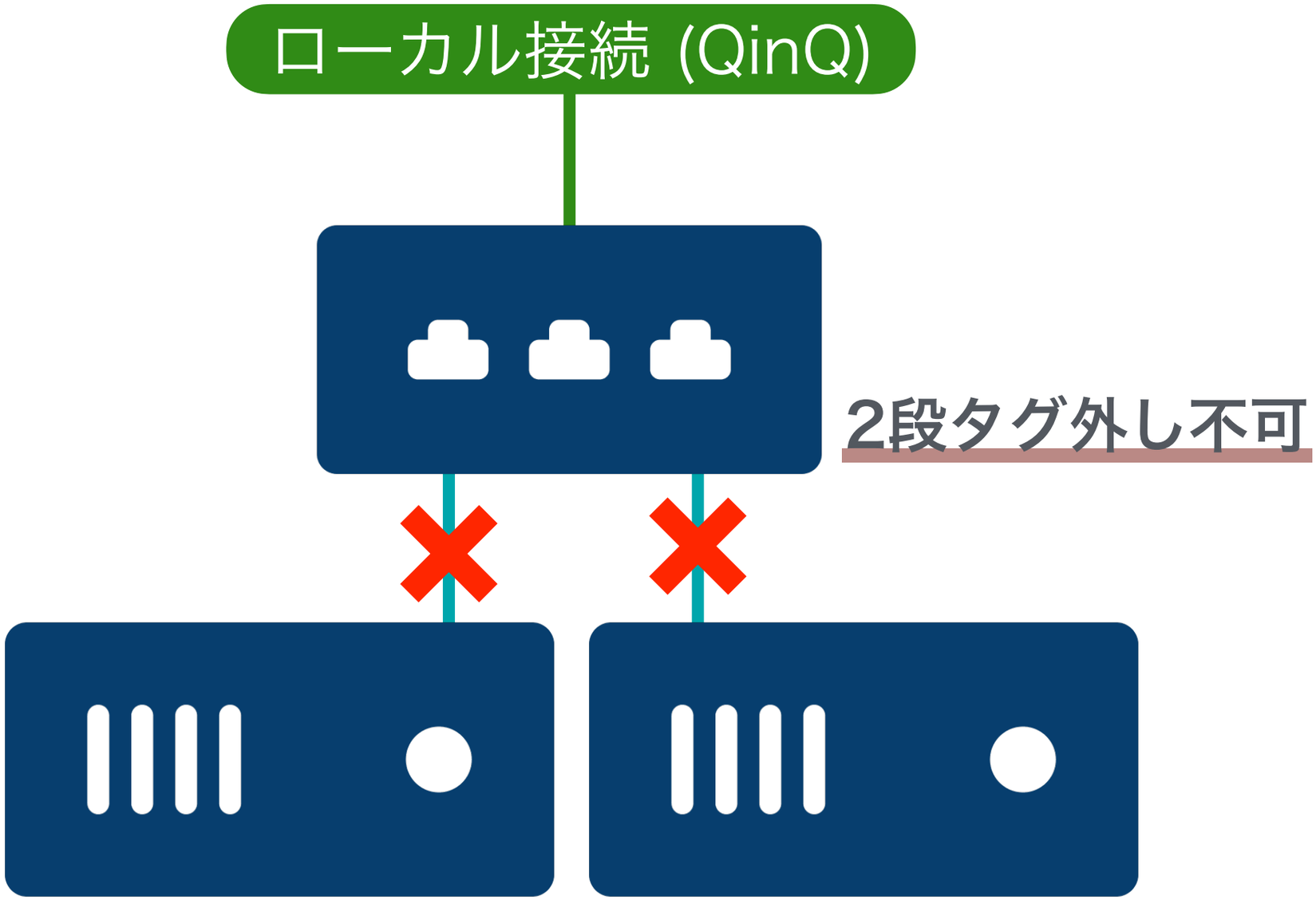


SAKURA internet

さくらのVPS ベアメタルプランのローカル接続の実装

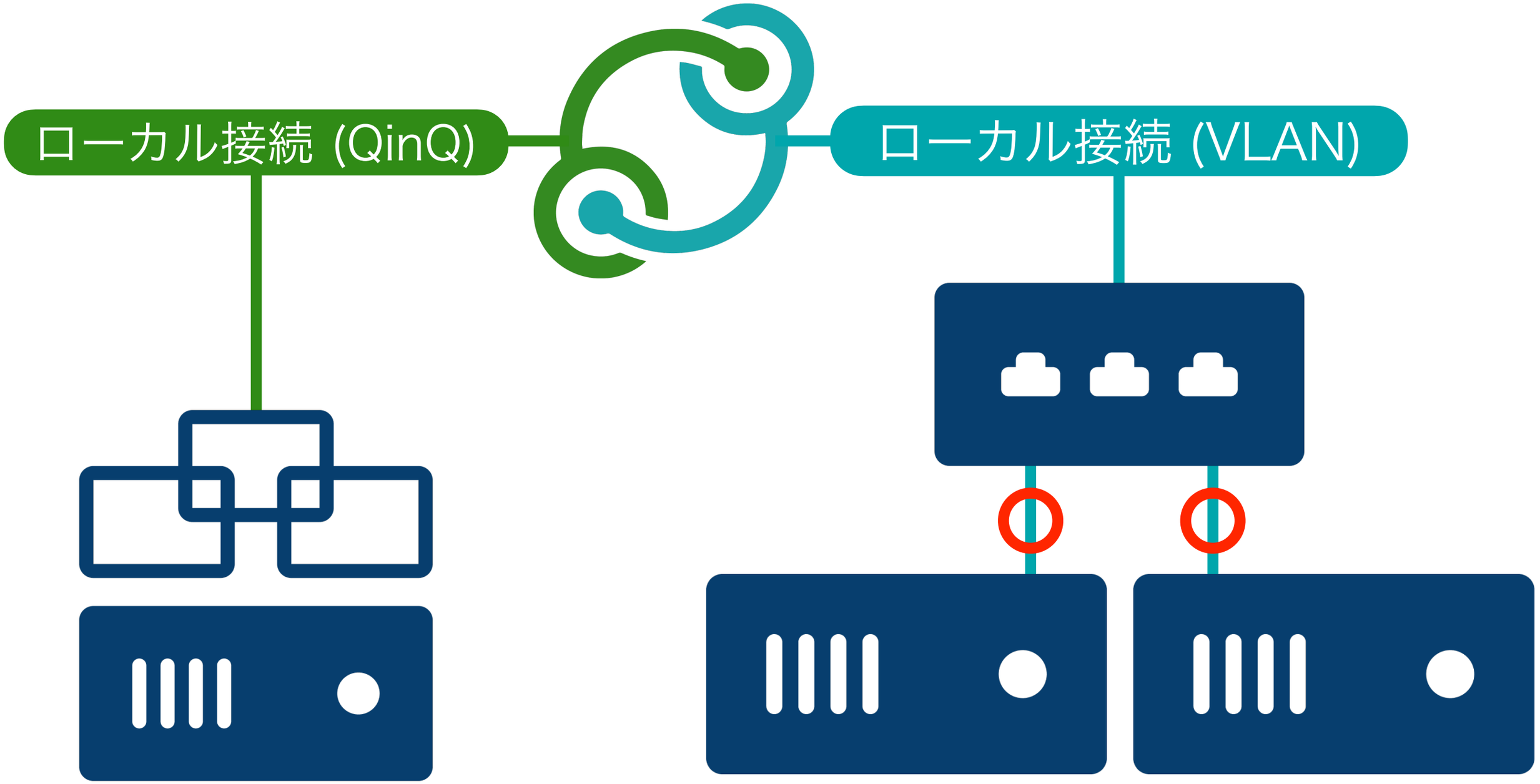


さくらのVPS ベアメタルプランのローカル接続の実装



SAKURA internet

さくらのVPS ベアメタルプランのローカル接続の実装

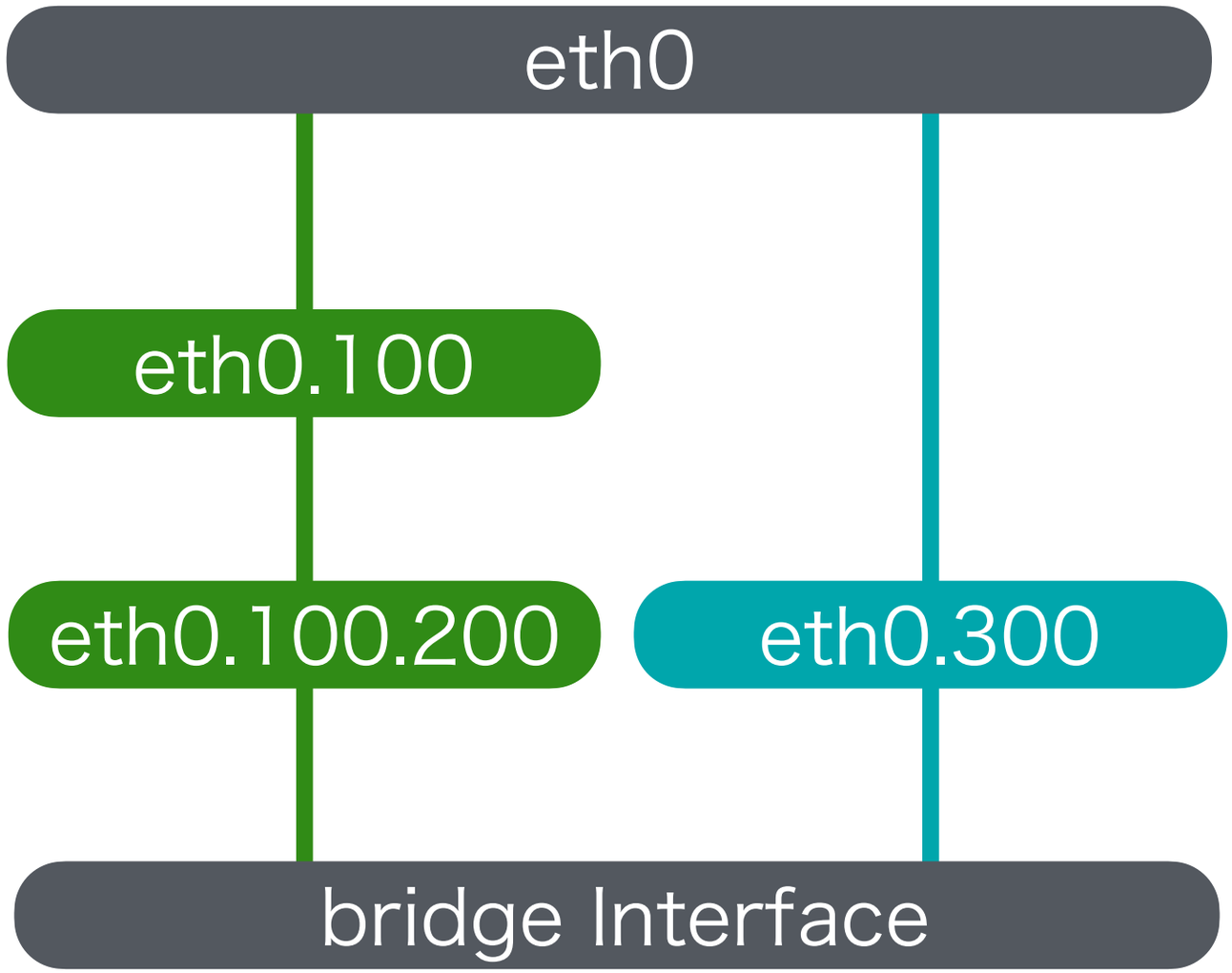


SAKURA internet

変換サーバ version1 の実装

変換サーバ version1 の実装

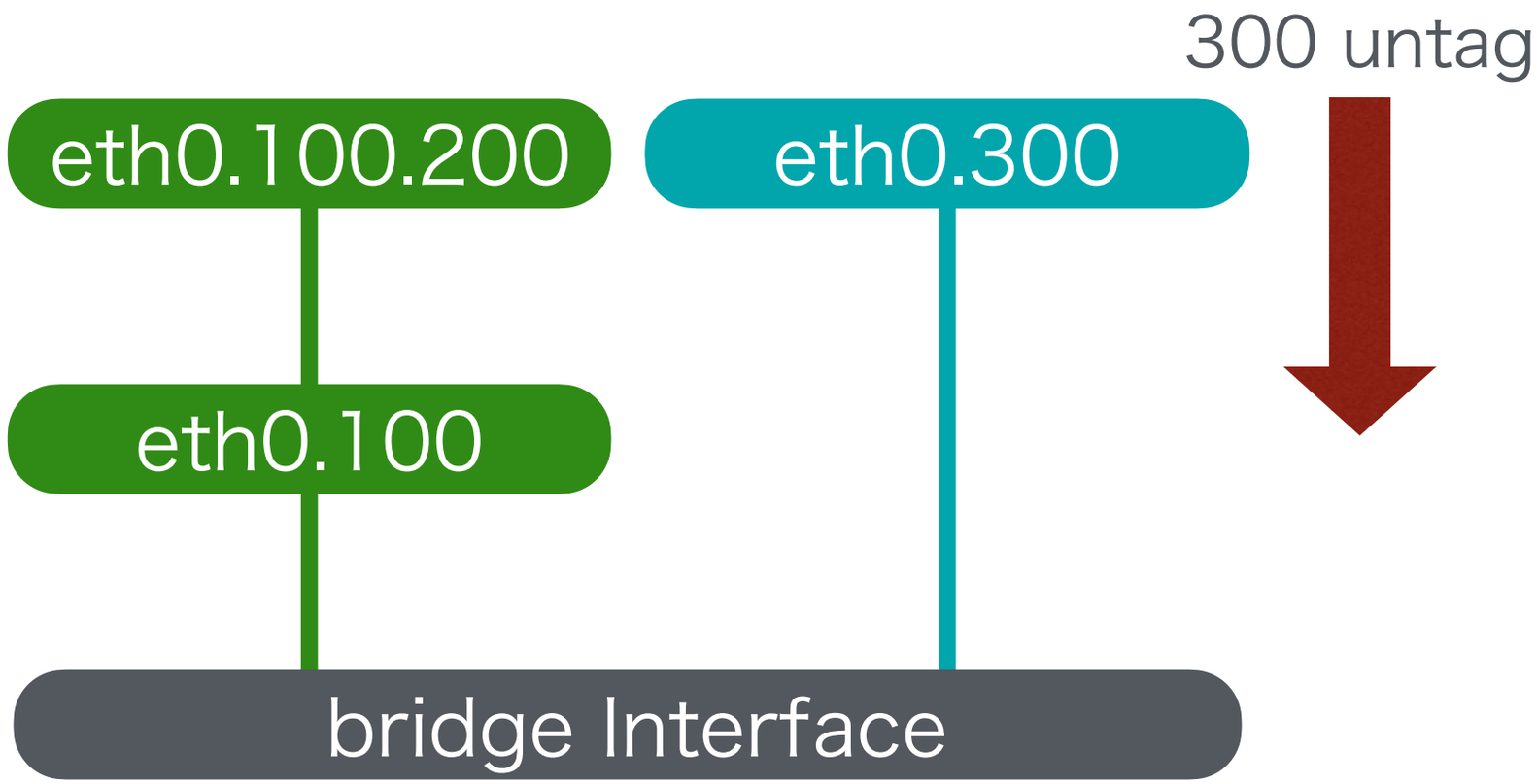
Linux Config



SAKURA internet

変換サーバ version1 の実装

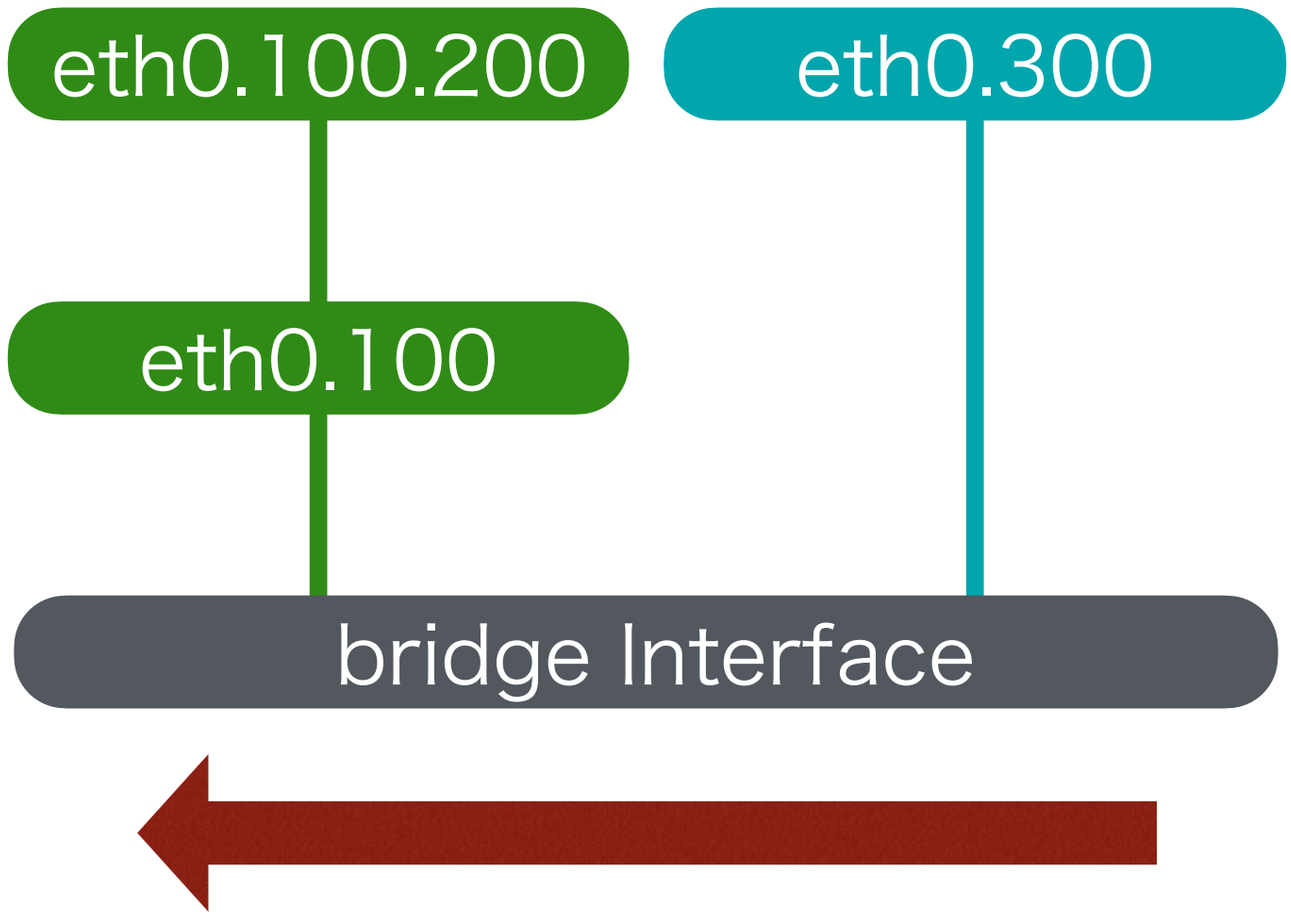
Flow image (VLAN:300 —> QinQ:100.200)



SAKURA internet

変換サーバ version1 の実装

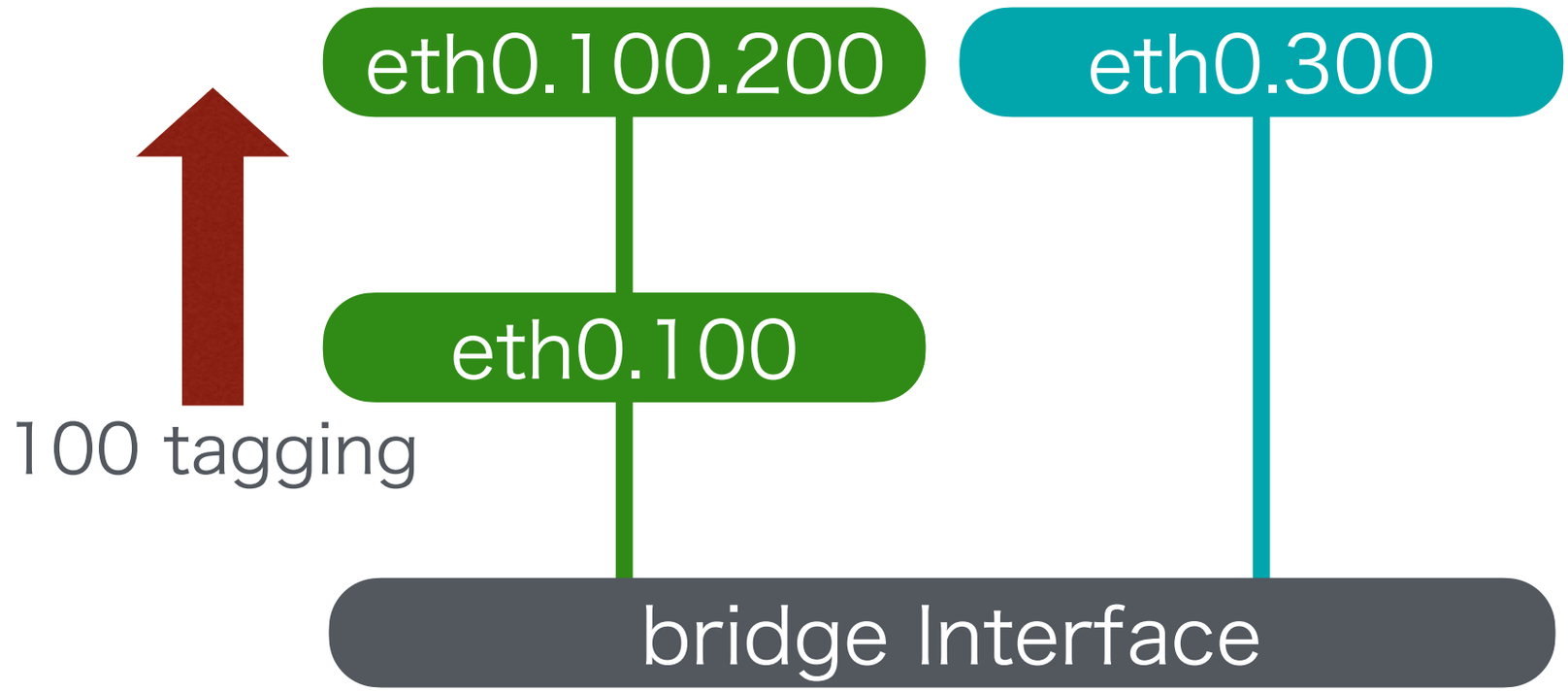
Flow image (VLAN:300 —> QinQ:100.200)



SAKURA internet

変換サーバ version1 の実装

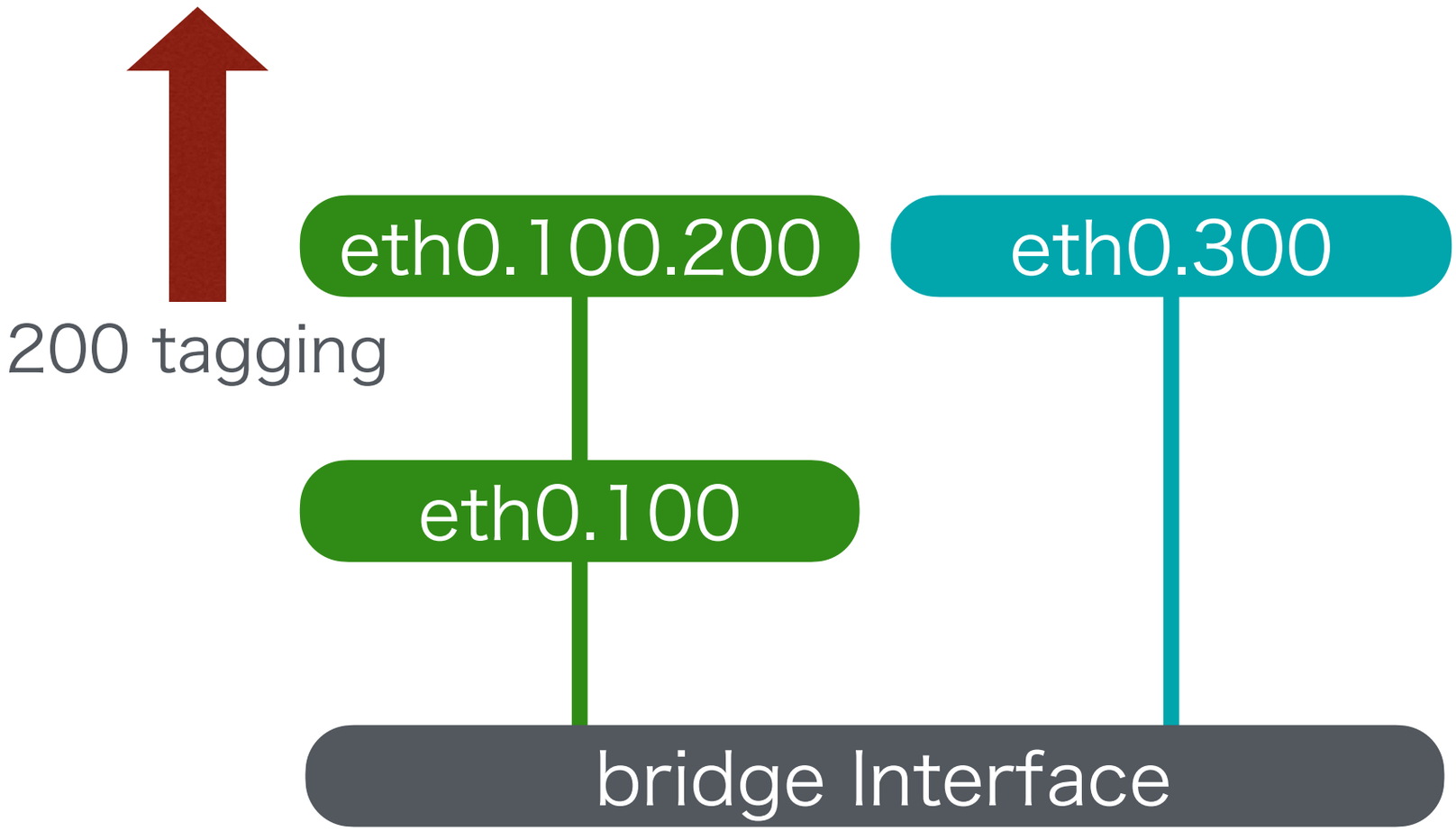
Flow image (VLAN:300 —> QinQ:100.200)



SAKURA internet

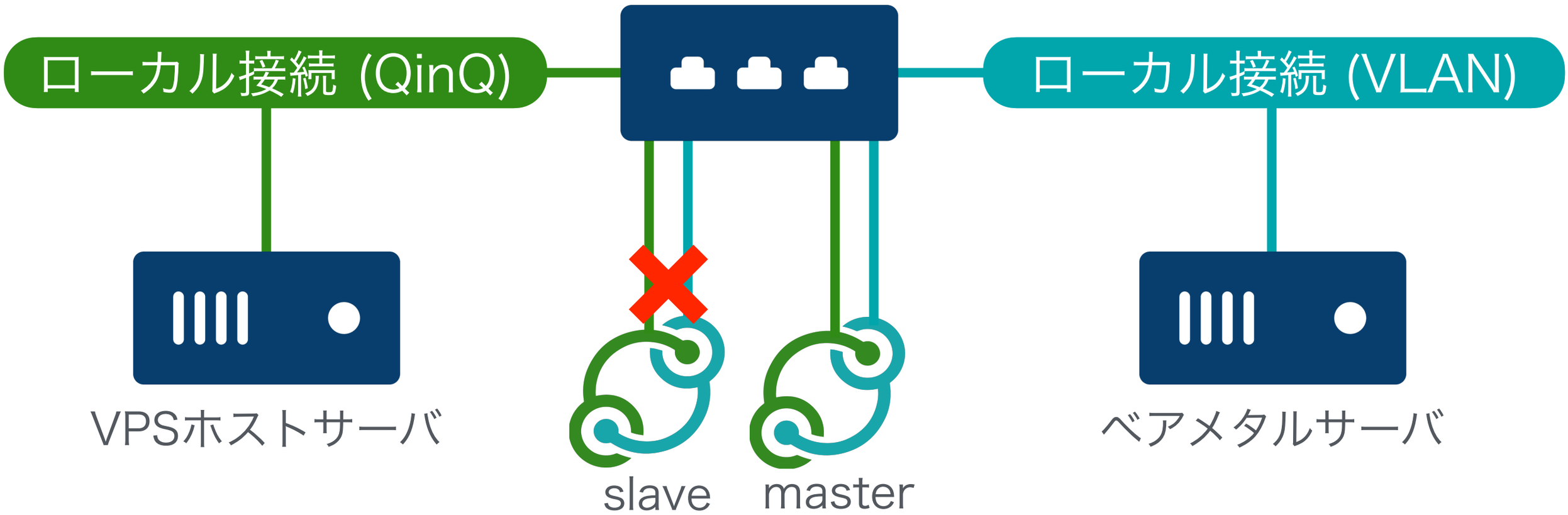
変換サーバ version1 の実装

Flow image (VLAN:300 —> QinQ:100.200)



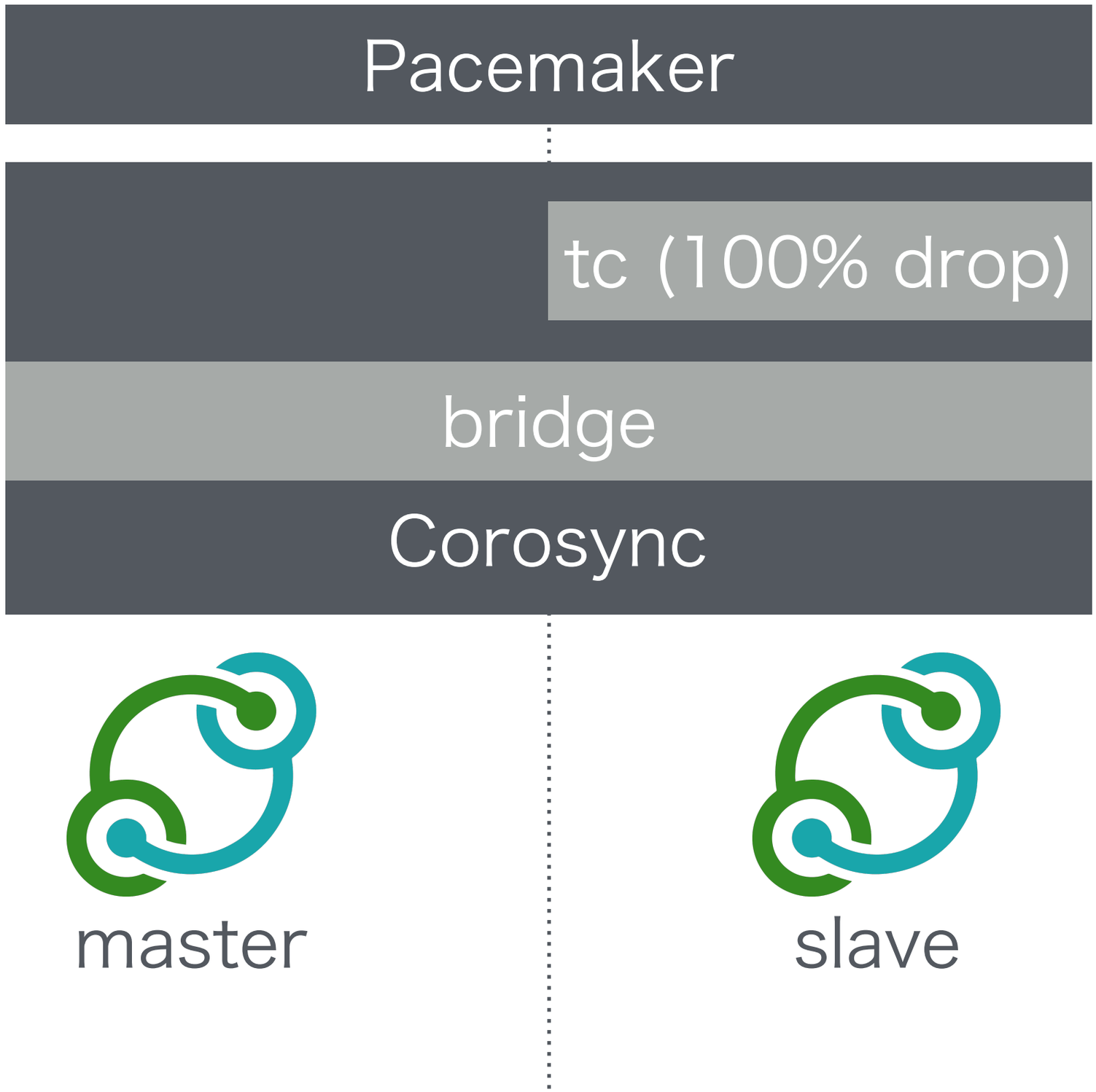
SAKURA internet

変換サーバ version1 の実装



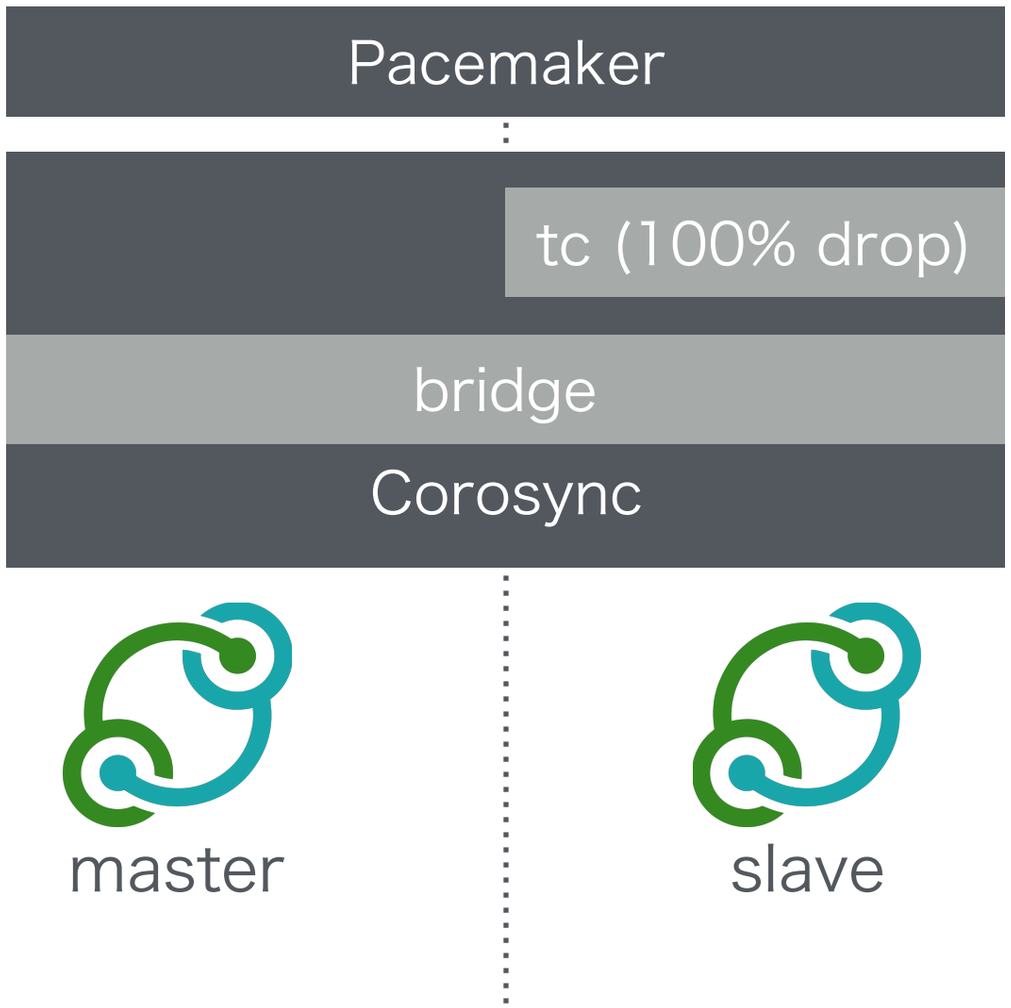
変換サーバ version1 冗長化構成

変換サーバ version1 冗長化構成



SAKURA internet

変換サーバ version1 冗長化構成



- Pacemaker/Corosync を利用
- Corosync プラグインを自前で実装
  - bridge plugin
  - tc plugin

- bridge plugin
  - フレーム変換用 sub interface を作成する
  - master/slave 両方で有効にすることで障害時の切り替えを高速化する
- tc plugin
  - slave において変換用 interface に対して tc コマンドで受信フレームを全て drop
  - slave が master に昇格する場合、tc のルールが削除され、トラフィックが流れる

SAKURA internet

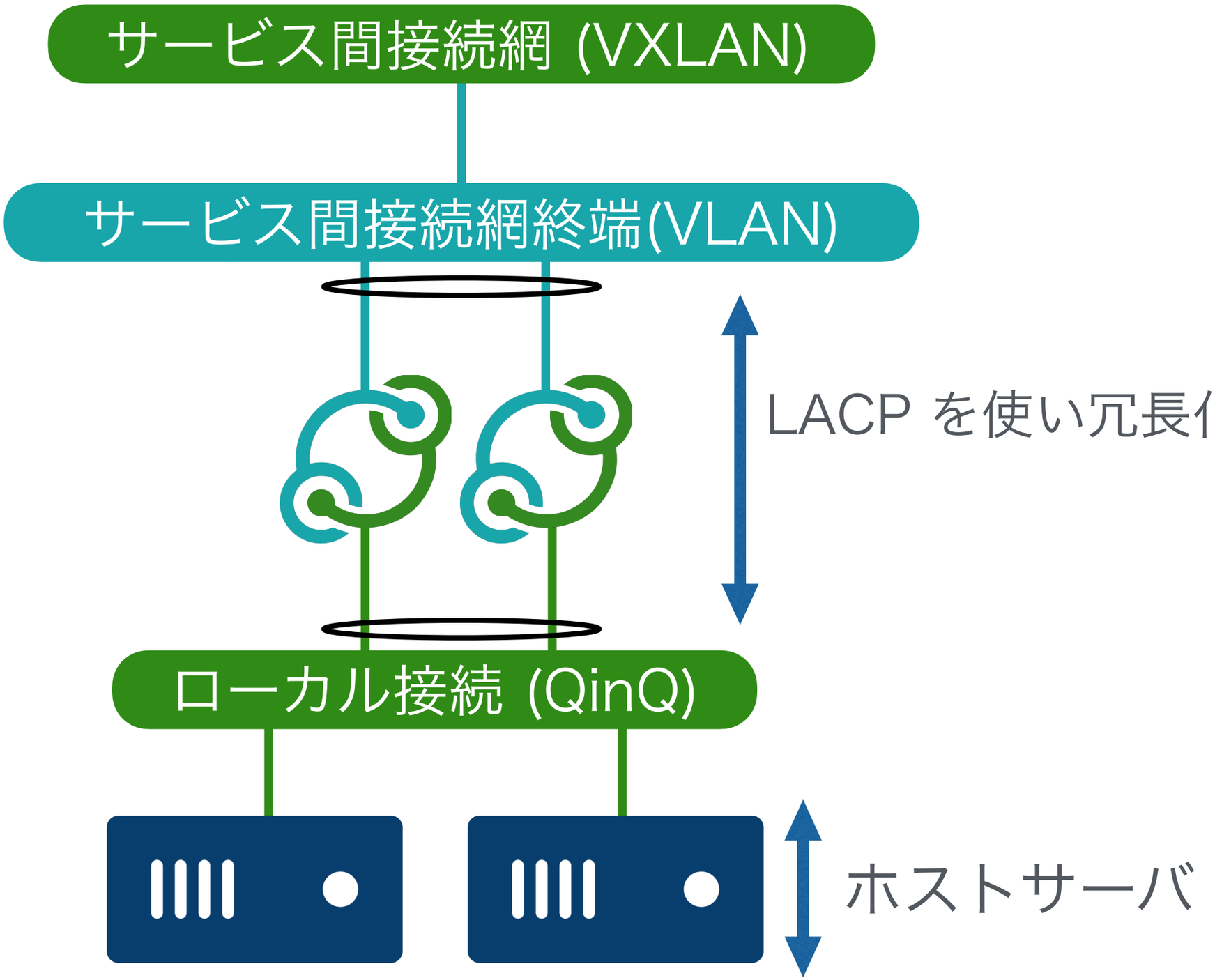
QinQ VLAN 変換サーバ version2

QinQ VLAN 変換サーバ version2

- さくらのVPSのサービス間接続対応のために開発
- サービス間接続は VLAN の島を VXLAN 網で繋ぐ仕組みのため QInQ のままでは使えないことが判明
- そこで、QinQ フレームを変換サーバで VLAN に直してから VXLAN 網に収容することでこの問題を解決
- 現在2ペア4台が本番稼働中
- そしてこの頃から社内でこのシステムを伊東バコと呼ぶように…

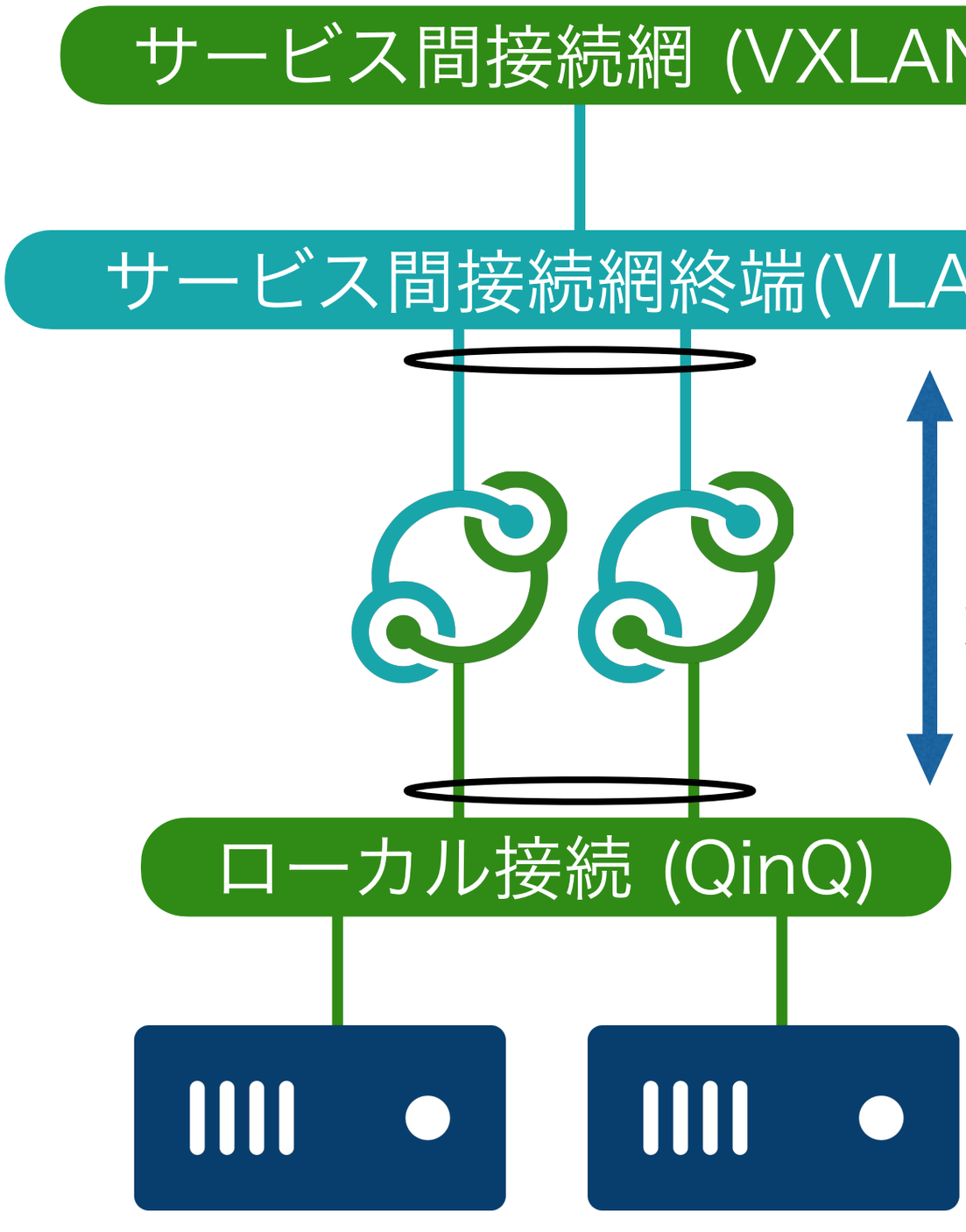
変換サーバ version2 の実装と冗長化

変換サーバ version2 の実装と冗長化



SAKURA internet

変換サーバ version2 の実装と冗長化



LACP を使い冗長化は両端の SW 任せ

が、できませんでした

ホストサーバ

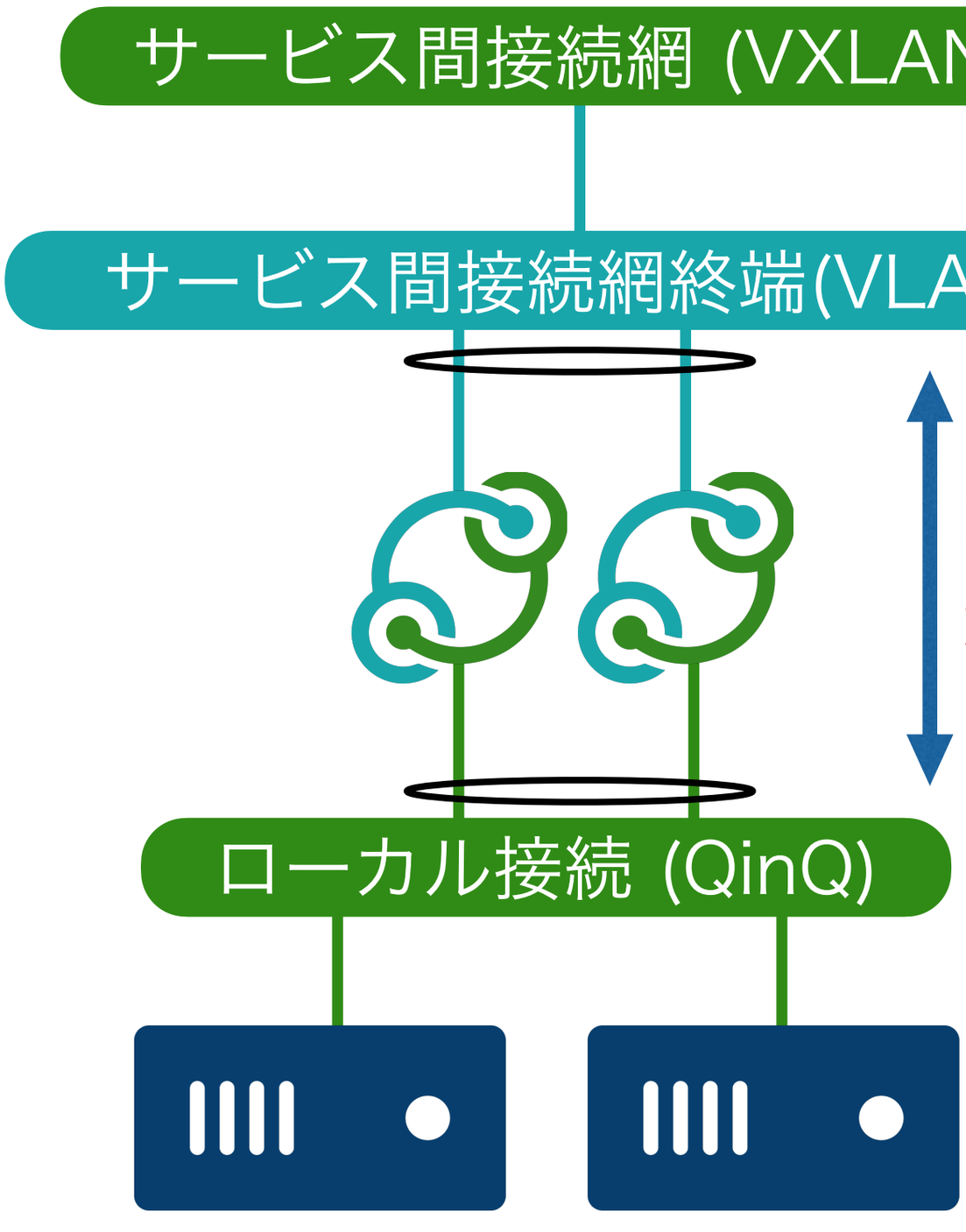
SAKURA internet

変換サーバ version2 の実装と冗長化

- version1 の Linux bridge 方式では LACP を
  通すことができない様子
- LACP を通すために Open vSwich で変換 bridge を実装
- 両端の SW からはまるで QinQ と VLAN が変換される
  不思議な光ファイバーのように見える
- これにより Corosync/Pacemaker による複雑な
  冗長化構成から脱却

SAKURA internet

変換サーバ version2 の実装と冗長化



LACP を使い冗長化は両端の SW 任せ

が、できました！

ホストサーバ

SAKURA internet

まとめ

- サーバのスペックが年々上がり、ある程度のトラフィックなら
  サーバで運べるようになってきたことを実感
  - 10G interface で long packet は 7~8Gbps 程度の性能
  - short packet はめっぽう弱く 500~1Gbps 程度の性能
- 工夫することで高価な製品の一部の機能を安価に実装できた
  - トラフィックも売上も増えたら製品を買いましょう、という
    選択肢を持てるようになった
- ただし自前で実装するので、運用まで考えた設計・実装を
  心がけないとひどい目に合うので要注意

会場への質問と議論

- この仕組みを使ってみたいという方はいらっしゃいますか？
  その場合どんな NW やサービスで使ってみたいですか？
- ネットワーク機器ベンダーさんから見たアドバイスや
  感想などいただけると嬉しいです
- もっとこうしたらいいんじゃない？などアイディアがあれば
  是非お願いします！

SAKURA internet