Internet Week 2013 T11

midokura ミドクラ

SDN 再入門

～便利なSDN? 難しいSDN? よくわからないSDN?～

2013年11月28日

ミドクラジャパン株式会社 高嶋隆一

本日の流れ

内容	時間	話者
SDN再入門	50分	高嶋
データセンタ編	30分	高嶋
WAN・キャリア編	50分	清水

話者紹介

ミステリス
パートナー1号

当日公開！

ミステリス
パートナー2号

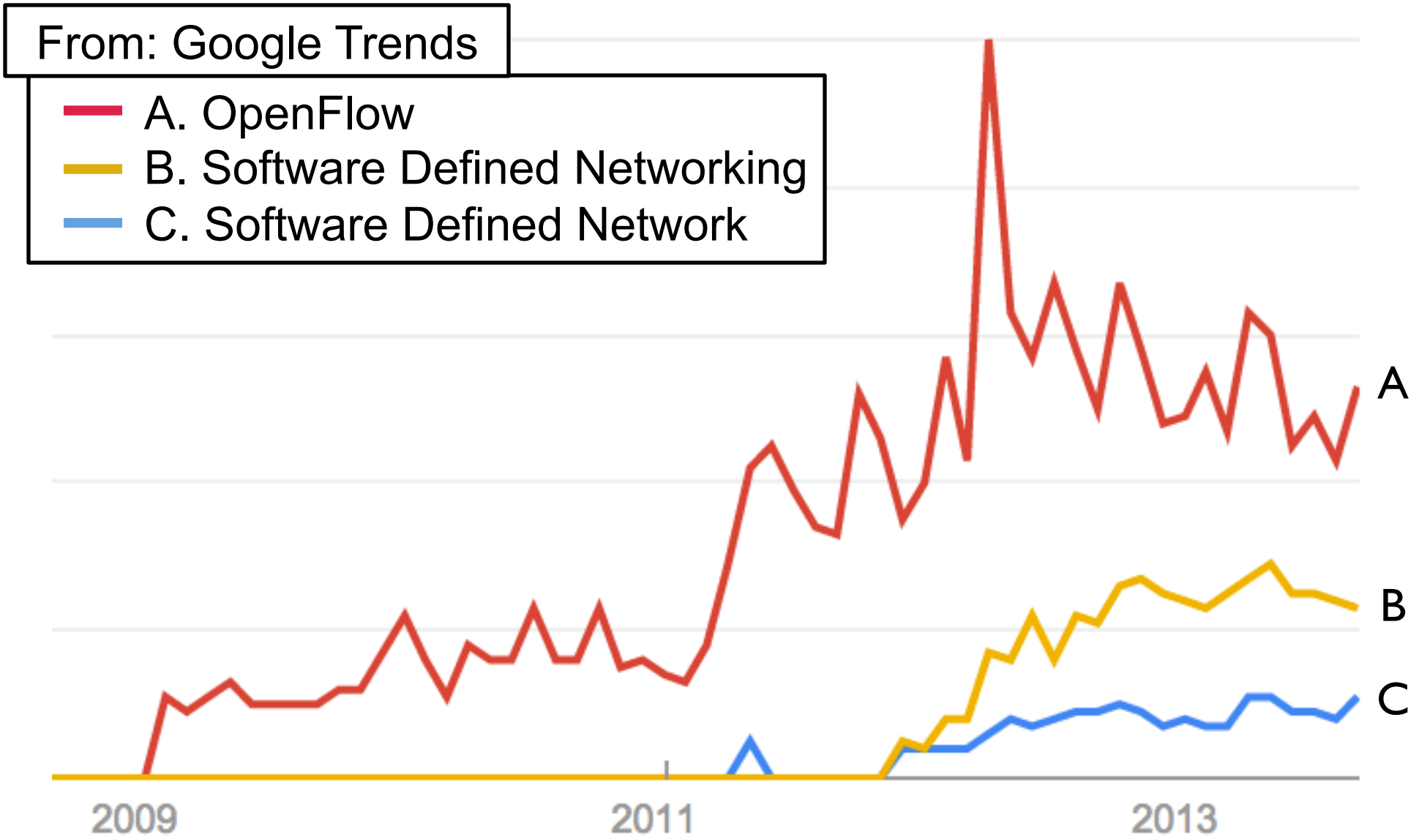
当日公開！

SDN再入門

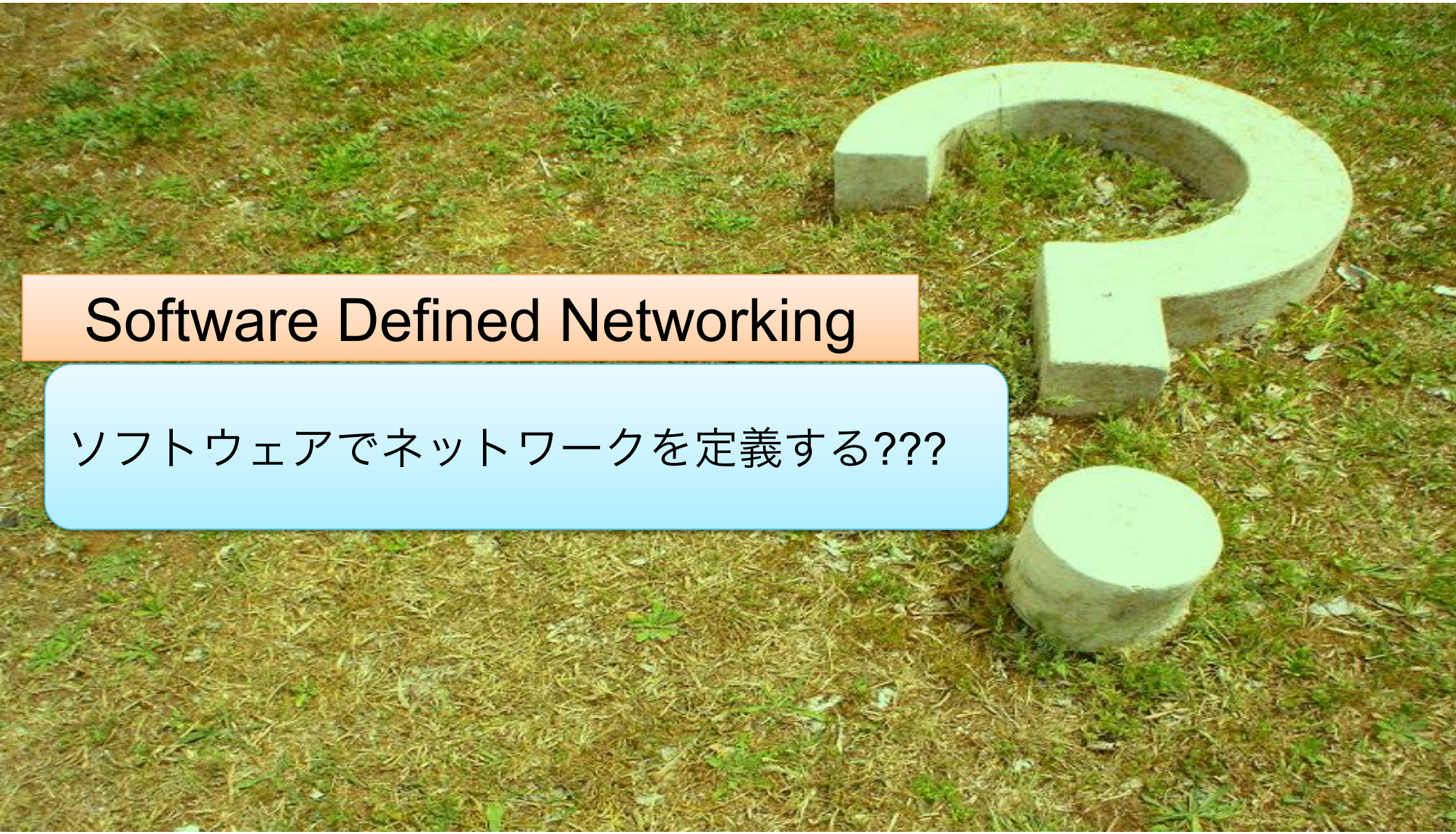
By 高嶋

Photo Credit: Micky Aldridge via flickr cc

2012を境に注目を集めている“SDN”



What's "SDN"?



Software Defined Networking

ソフトウェアでネットワークを定義する???

Web 上で見る様々な“SDN”の定義

SDN is a new approach to networking in which network control is decoupled from the data forwarding function and is directly programmable.

From: https://www.opennetworking.org/about/onf-overview

ネットワークの構成、機能、性能などをソフトウェアの操作だけで動的に設定、変更できるネットワーク、あるいはそのためのコンセプトを指す

From: http://www.atmarkit.co.jp/ait/articles/1304/08/news098.html

Web 上で見る様々な“SDN”の定義 cont.

ソフトウェアによって仮想的なネットワークを作り上げる技術全般を言います。 SDNを用いると、物理的に接続されたネットワーク上で、 別途仮想的なネットワークを構築するといったようなことが可能になります。

From: https://www.nic.ad.jp/ja/basics/terms/sdn.html

SDNとは、ネットワークをソフトウェアで動的に
～中略～
そこで、従来、個々のネットワーク機器が1台ずつで行ってきたネットワーク制御とデータ転送処理を分離し、汎用サーバ側のソフトウェアでデータ転送処理のみを行う機器を動的に制御することで、通信を柔軟に効率よく、安全に行えるようにすることを目指して考えられたのがSDNです。

From: http://jpn.nec.com/sdn/about_sdn.html?

Web 上で見る様々な“SDN”の定義 cont.

共通項

- ✓ “ソフトウェアで”
- ✓ “動的に変更”

その他のキーワード

- ✓ コントロールプレーン、データプレーン分離
- ✓ 自動化
- ✓ 機能の追加
- ✓ 仮想化
- ✓ 汎用ハードウェア

Web 上で見る様々な“SDN”の定義 cont.

ポイント

- ✓ “SDN” という固有名詞の標準技術は存在しない
- ✓ ソフトウェアでネットワークに対して動的制御を行う仕組みをなべて “SDN” と呼んでいる

Photo Credit: @Doug88888 via Compfight cc

“SDN”を分類してみよう

Photo Credit: 5letterdesign via Compfight cc

無数の“SDN”ベンダ

infinera

vmware

midokura

BROCADE

IBM

JUNIPER NETWORKS

PLUMgrid

PICA8 OPEN NETWORKING

CISCO

Wakame Enlightened Enterprise Computing.

big switch networks

hp

nuagenetworks

Stratosphere

分類I. “適用領域”と”物理vs仮想”

当日公開！

分類II. “新興vs既存”と “HardwareとSoftware”

当日公開！

分類III.“利用技術”

当日公開！

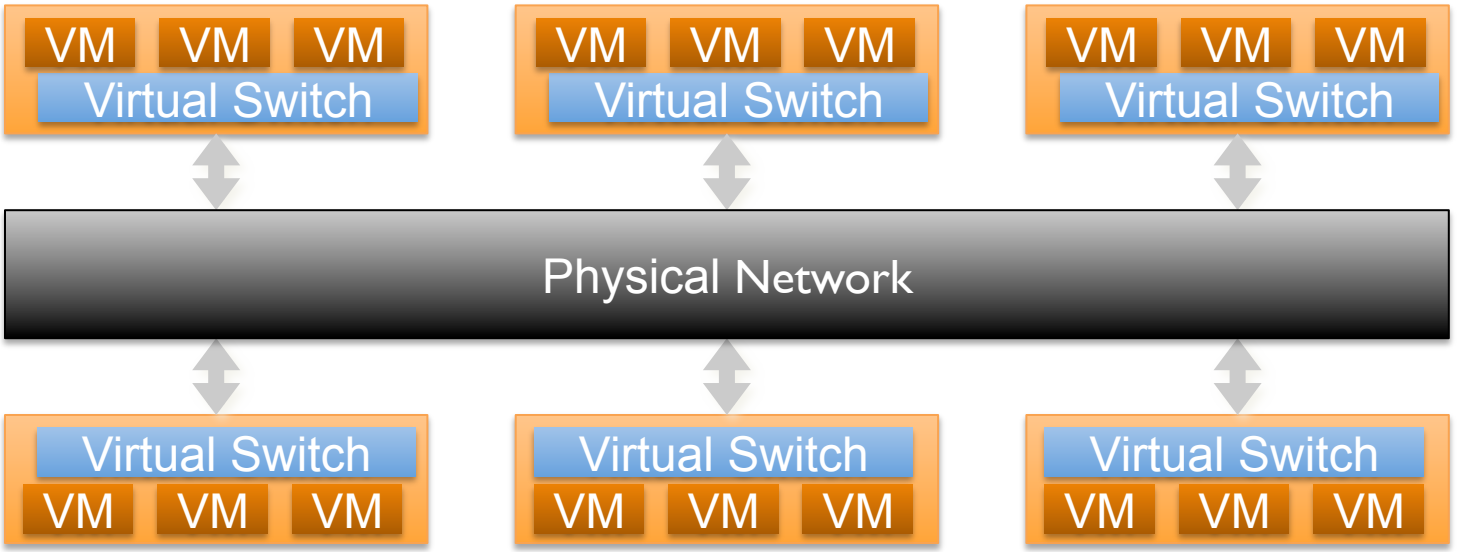
様々な構成技術

Photo Credit: See-ming Lee 李思明 SML via Compfight cc

仮想スイッチ

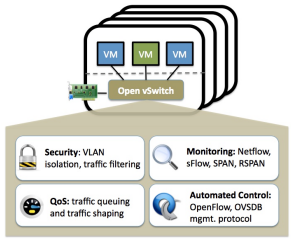
仮想スイッチとは

- ✓ ソフトウェアで動作するスイッチ
- ✓ 仮想化のホストOSで動作し、VMと外部ネットワークを接続する用途が多い
- ✓ 単純なスイッチング以外にもトンネル化、VLAN の追加削除等のヘッダ操作もサポート



仮想スイッチ

代表的な実装例



Open vSwitch

- ✓ Linux 上で動作
- ✓ Data path となる kernel module と、Control plane となる application から構成
- ✓ OpenFlow を使うことも、直接Data pathをプログラミングする事も可能

CISCO

Nexus 1000v

- ✓ ESXi, Linux 上で動作
- ✓ 基本的には Cisco UCS 上での動作を想定
  (IAサーバなら動作は可能…な筈)

vmware

vswitch, vDS

- ✓ VMWare ESXi/Infrastracture 上で動作
- ✓ 単純なスイッチ機能がメイン

仮想スイッチ

Open vSwitch

✓ Linux 上で動作し、Open source で開発されている為、商用利用への応用も盛んで、”SDN 製品” にもしばしば利用される

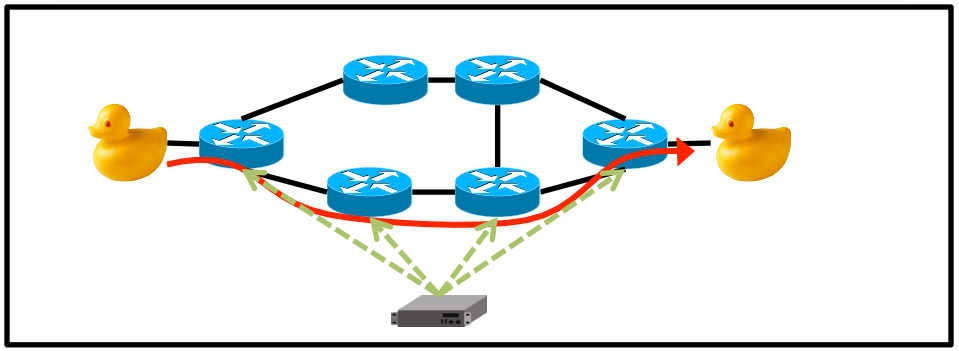
応用例

PICA8 OPEN NETWORKING 汎用FPGAを用いた新興機器メーカ	✓ スイッチのベースOSに Linux を採用しており、データパスプログラミングに Open vSwitch が利用可能 ✓ OVS の機能により最新の OpenFlow や他のトンネルプロトコル等に一早く対応可能
midokura クラウドネットワークスタック	✓ KVM 等の Linux をホストOSとしているものは、そのまま動作可能 ✓ OpenFlow 以外にも、NETLINK 経由で直接プログラミングする事により、より柔軟な拡張が可能

Hop-by-hop と Overlay

Hop-by-hop

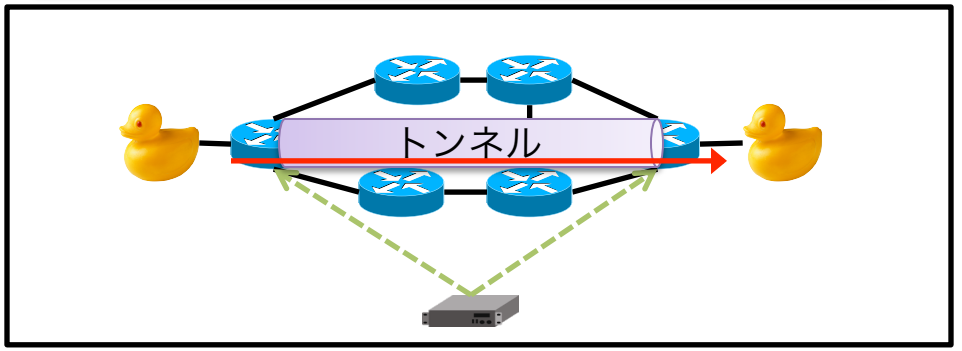
✓ 経由する機器を全て設定



○ きめ細かな制御が可能
× 設定量が膨大

Overlay

✓ 通信元と宛先を収容する装置のみを設定
✓ 装置間は何らかのトンネルプロトコルで接続し、従来のルーティング等により到達性を確保

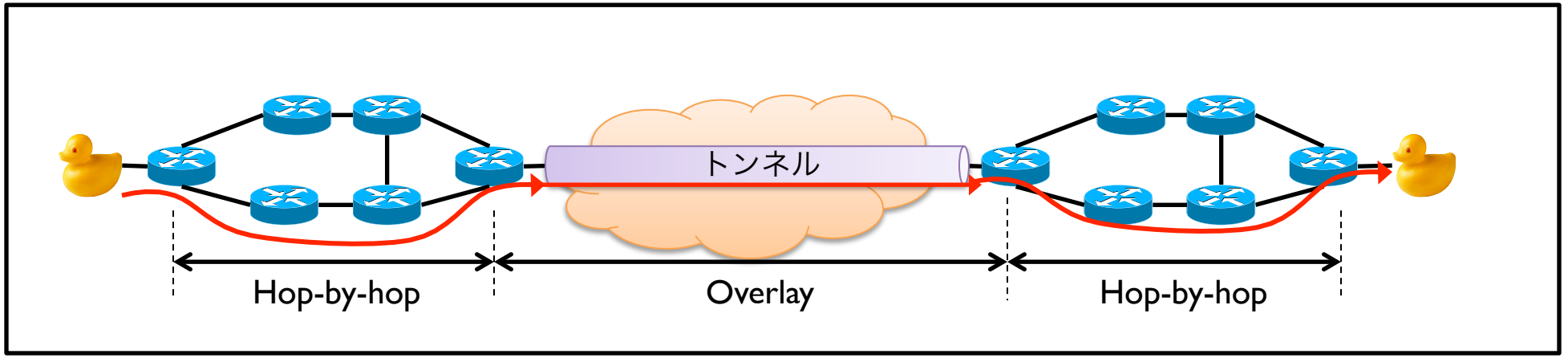


× トンネル区間は複雑な制御は困難
× L2 を上位で動作させる場合は複雑な制御が必要
○ 設定量が現実的な量で納まる

Hop-by-hop と Overlay

併用

- ✓ 制御するインタフェースとしてトンネルインタフェース等を指定できれば、Hop-by-hop と Overlay の差は扱うインタフェースの種別の違いであり、併用は可能
- ✓ 但し、制御はそれだけ複雑に…

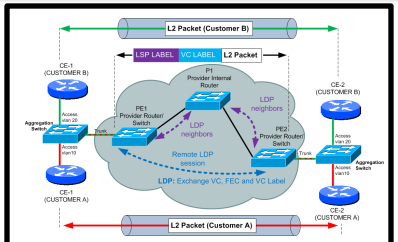


Edge Overlay

Edge overlay

- エンド収容装置 (Edge) がトンネルを張る Overlay

応用例



VPLS

- ✓ 回線収容装置が MPLS LSP を張り、顧客の L2 回線を Overlay として提供

midokura

クラウド
ネットワーク
スタック

- ✓ 仮想化ホストOS上の仮想スイッチが何らかのトンネルを張り、VM-to-VM の通信を Overlay として提供
  (Server-side Edge Overlay)

Overlay で用いられるトンネル技術

GRE	✓ 所謂 GRE トンネリング ○ 枯れており実装も多い × L2 の考慮は上位アプリケーションで考慮する必要がある
NVGRE	✓ L2 フレームを GRE で通す為の拡張、GRE Tunnel ID を分割 ✓ 24bitsの Tenant ID を持ち、VLAN空間12bitsより広大 ○ GREに見える為、既存のフレームワークの改変が不要 × Multicast/Broadcast が Tenant IDとリンクしていない
VXLAN	✓ L2 フレームを UDP で通すトンネリング ✓ 24bits のテナント識別子を持ち、VLAN空間12btisより広大 ○ UDP ヘッダにより、経由ノードで細かな制御が可能 × MAC学習に標準では IP Multicast を使う為、実装が困難※

※VMWare NSX では IP Multicast による動的学習ではなく
独自機構によりコントローラから配信される

OpenStack と Neutron

OpenStack

- ✓ Open Source なクラウドマネジメントシテム
- ✓ 多数のコンポーネントの API による疎結合により実装
- ✓ Network は Neutron というコンポーネント群によって制御され、API も公開されている

Computing

OpenStack Compute (Nova)

OpenStack Image service (Glance)

Networking

OpenStack Networking (Neutron)

Storing

OpenStack Object Storage (Swift)

OpenStack Block Storage (Cinder)

OpenStack と Neutron

Neutron API

- ✓ OpenStack に準拠する為、多くのメーカ、ベンダが Neutron API と自社 API を変換する Plugin を実装
- ✓ データセンタ・クラウド向け Network API のデファクトスタンダードの候補として注目されている

制御対象

- ✓ L2/L3といった基本的な機能だけではなく、高レイヤのサービスにも拡張

L2 separation	LBaaS	Service Insertion
L3 separation	VPNaaS	QoS
Security Group	FWaaS	Etc, Etc …

NFV = Network Function Virtualization

用語としてのNFV

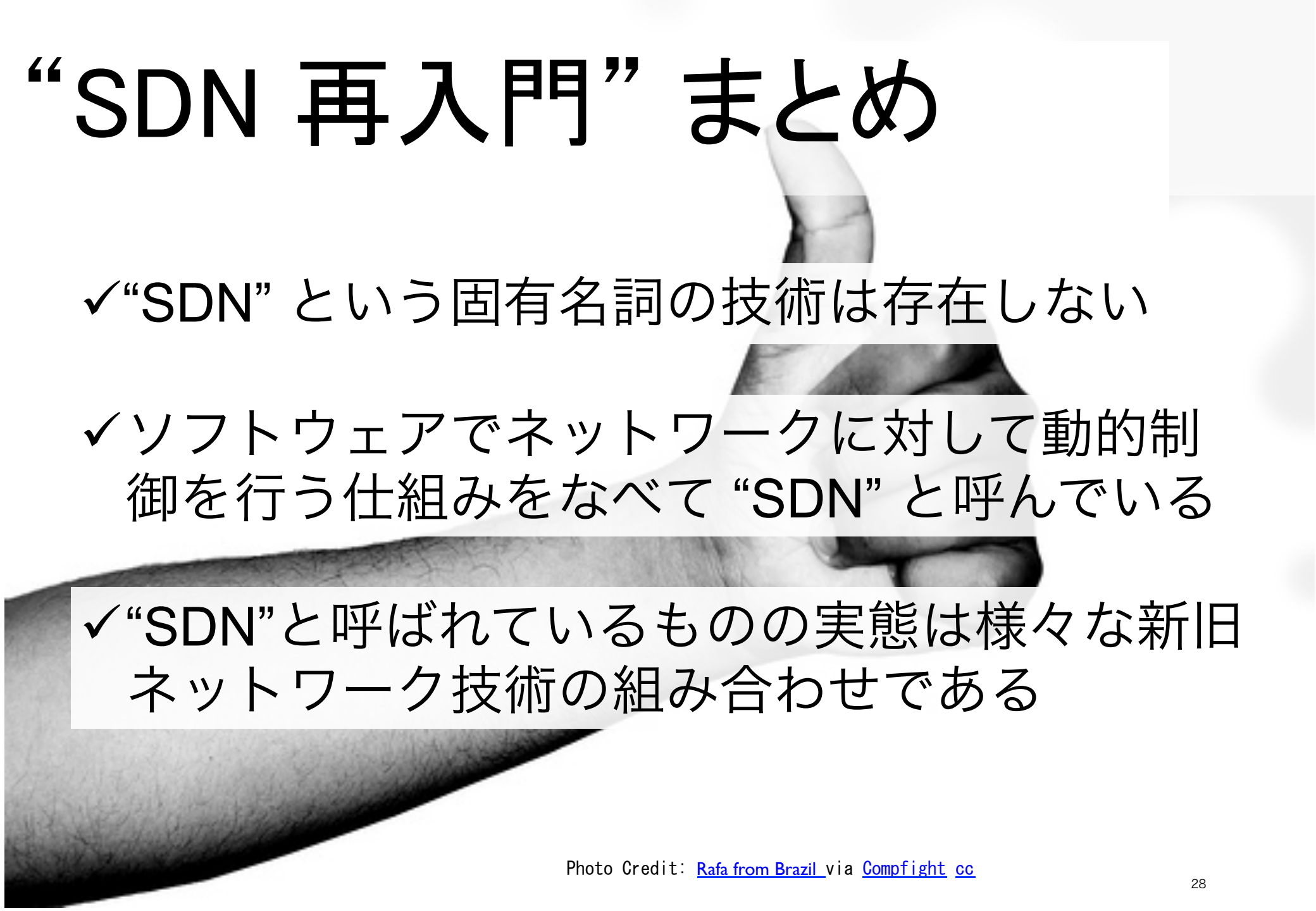
- ✓ LBaaS, VPNaaS, FWaaS 等、従来専用機器が機能を提供していたネットワークサービスを汎用サーバ上で実現する事を呼ぶ広義の NFV と、標準規格としての狭義の NFV が存在する

仕様としてのNFV

- ✓ 通信事業者の標準化団体である ETSI (European Telecommunications Standards Institute) により、規格化を検討
- ✓ 2013年10月14日に初版の仕様が公開

http://www.etsi.org/technologies-clusters/technologies/nfv

“SDN 再入門” まとめ



- ✓“SDN” という固有名詞の技術は存在しない
- ✓ソフトウェアでネットワークに対して動的制御を行う仕組みをなべて “SDN” と呼んでいる
- ✓“SDN”と呼ばれているものの実態は様々な新旧ネットワーク技術の組み合わせである

Photo Credit: Rafa from Brazil via Compfight cc

データセンタ編

By 高嶋

Why “SDN” in Data Center ?

Requests from DC Network

サービス

AWS like な IaaS の構築

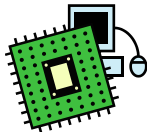
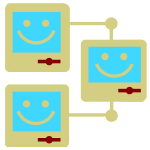
コスト

Network 機器の White box 化

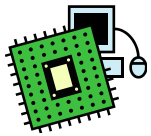
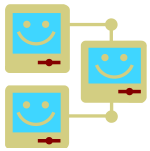
Network 機器の White Box 化の要求

背景

コスト推移

サーバ 	10数年前と比較すると普及クラスのものの価格は数分の一
ネットワーク 	10数年前と比較してもそれほど変わらない

Why ?

サーバ 	IA アーキテクチャへの収斂が進み、汎用化
ネットワーク 	未だに数社のメジャープレーヤがプロプライエタリな実装により寡占している

特に導入台数の多いデータセンタで問題に

White box ?

ホワイトボックス（英語:White Box）とは、特定のブランドを持たないノーブランドパソコンや、卸売業者や販売店、ソリューションプロバイダーなどが自社のブランドをつけて販売するプライベートブランドパソコンやショップブランドパソコンのことである。

From: http://ja.wikipedia.org/wiki/ホワイトボックス_(パソコン)

- ✓ 従来、計算機そのものもハードウェア、OS、アプリケーションは全て作り込みをされたセットであったものが、現在では汎用アーキテクチャの組み合わせにより実現
- ✓ OS、アプリケーションの選択も自由

汎用化が進めば進む程、低廉な White box が出現

Network 機器の White box 化

チップベンダの減少	➢ どこのメーカも同じチップを使用
汎用OSの採用	➢ ベースとなる制御は Linux 等の Open Source based のものが増加
ODM/OEMベンダの共通化	➢ 比較的安価な BOX スイッチ等はどこのメーカも同じODM/OEMベンダに製造委託

共通の OS や API があれば、
計算機と同じ様に低廉な White box 化したネットワーク機器が使えるのではないかという期待

これも、SDN ??

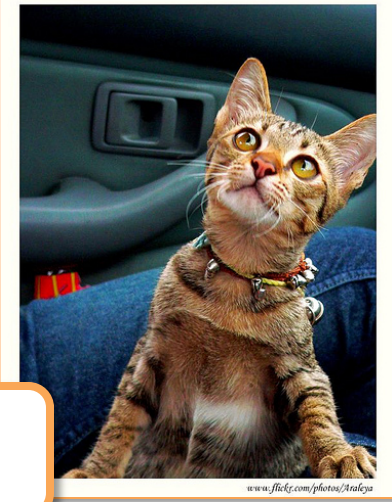


Photo Credit: Araleya via Compfight cc

今日現在の状況

API の例

OpenFlow	✓ パケット処理の API から始まっている ○ 共通化 X 非常に低レイヤを見るフレームワークやライブラリに相当するAPIの為、ユーザが利用するにはアプリケーションやOSに相当する部分を開発しなくてはいけない

実装例

PICA8 OPEN NETWORKING	✓ 大手ベンダと同じチップ、Linux と XORP＊, Open vSwitch を用いた廉価な Box スイッチを提供 ＊ http://www.xorp.org/
big switch networks	✓ 大手ベンダと同じチップを使用した廉価な Box スイッチに対し、OpenFlow を制御のキーにした OS を提供

今日現在の状況 cont.

Open Compute Projectの場合

■OCP

➢Facebook 等が主導する“高効率で低コストのデータセンターを追求するワールドワイドなエコシステムを推進する” Project

■Network Component in OCP

➢Quanta の様な ODMメーカが汎用チップを使ったToR スイッチを提供を開始

Quanta QCT Named OCP Solution Provider and Launches New OCP Product Lineup, Rackgo X

Quanta QCT, an active member of the Open Compute Project (OCP) and design innovator for datacenter gear based upon OCP specifications, has achieved OCP Solution Provider status and is launching a full lineup of OCP server, storage and network gear under the Rackgo X product name.

San Francisco, CA (PRWEB) October 24, 2013

Tweet Like +1 Share EMAIL

Quanta QCT, an active member of the Open Compute Project (OCP) and design innovator for datacenter gear based upon OCP specifications has achieved OCP Solution Provider status and is launching a full lineup of OCP server, storage and network gear under the Rackgo X product name.

Quanta QCT, a leading manufacturer of server, storage and network products for both public and private cloud datacenters, today began accepting orders for its Rackgo X product series.

The Rackgo X product line includes two server solutions, models F03A and F03C. The lineup also includes one storage JBOD (model JBR) and one 10G SFP+ switch (model T3048-LY2). A server motherboard option, the Windmill F03, is also available. The designs for all products in the Rackgo X line will be contributed back to the Open Compute Foundation so everyone in the industry can consume and build upon them.

Quanta worked with Facebook early in the evolution of what would eventually become the Open Compute Project, providing design, engineering and manufacturing support to Facebook as it developed its first servers. The Rackgo X lineup is inspired by OCP's

F03A Open Rack server

“...companies like Quanta, are exactly the kind of thing that is helping the OCP community make datacenter technologies more innovative, more efficient, and more sustainable.”

White box 化に向けた課題

ハードウェアアーキテクチャの共通化

- チップベンダの寡占化は進んでいるが、アーキテクチャの共通化は IA サーバほど進んでいない

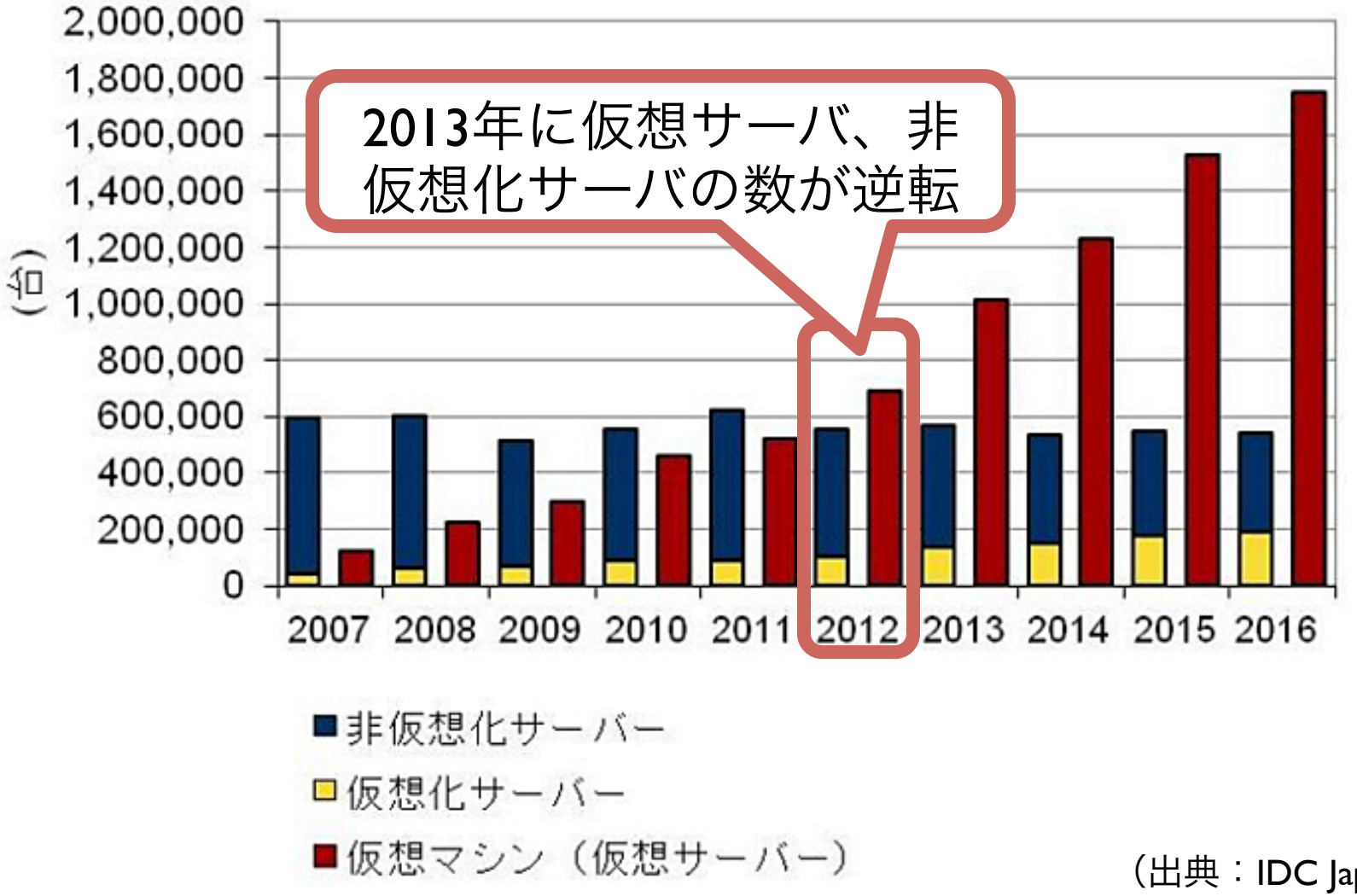
OS、APIのテンプレート化、成熟化

- コンシューマ向けには Windows 8, サーバ向けには UNIX や Windows server といった様な典型的な選択肢がまだ用意されるほど市場が成熟していない

AWS like な IaaS の構築の要求

国内サーバ市場の動向

国内サーバ市場の動向 2007年～2016年



（出典：IDC Japan）

パブリッククラウドの台頭

Cloud meets Enterprise IT !

- ✓ オンデマンドデプロイメント
- ✓ 従量課金

amazon web services™ +

Photo Credit: Stratfordcollege via Compfight cc

パブリッククラウドでの課題

All data on Public Cloud ?

- ✓ 情報漏洩のリスク
- ✓ 国を跨いだ場合の法的リスク

Photo Credit: KarmenRose via Compfight cc

プライベートクラウド？

固定費の増大

- ✓ 結局、”所有”“占有”
- ✓ AWS like なシステムの構築の必要性

Photo Credit: kevin dooley via Compfight cc

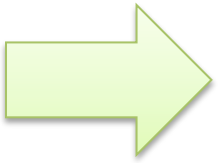
IaaS クラウド ~= AWS like なシステムの要求

IaaSクラウドの要求

ユーザの必要とする「計算機資源」を
「必要な時に」「必要なだけ」
提供すること

How?

ストレージ &
ネットワークが問題



オーケストレーション

計算機資源のスケールアウト

仮想化・クラウド化がネットワークに持たらす課題

自動化

- ✓ CPU・メモリは仮想化、自動化されているが
  ネットワークは？

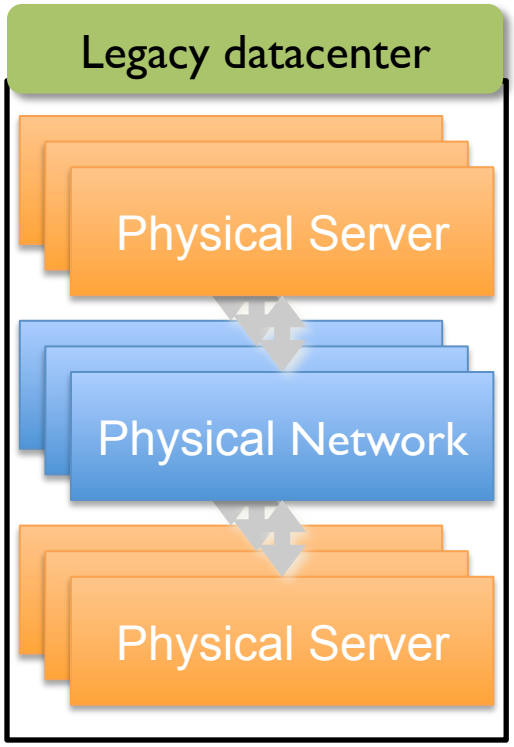
パフォーマンス

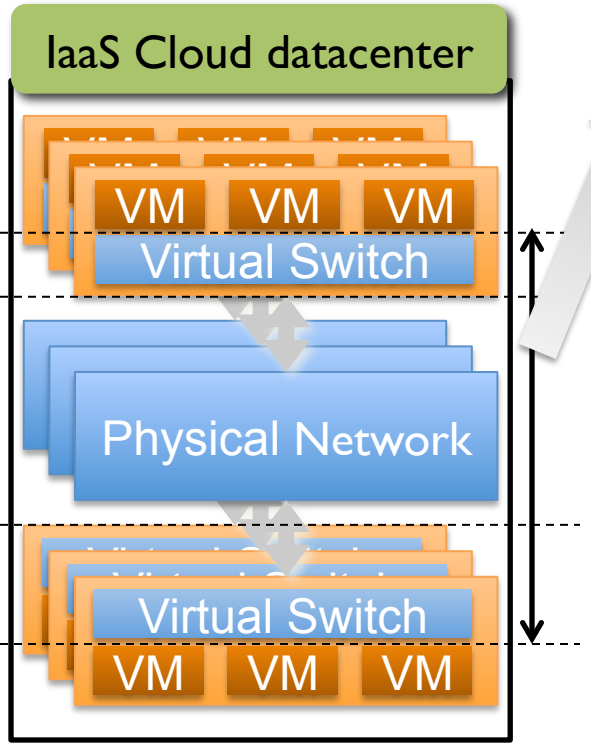
- ✓ VM多重収容によるポート当たりの帯域使用率の激増
- ✓ East-to-West トラフィックの爆発
- ✓ ボトルネックとなる仮想ルータVM

拡張性

- ✓ ネットワーク収容数 → VLAN 4,096 の壁
- ✓ VM収容数 → MAC テーブルの増大
- ✓ L2ネットワーク → STP の限界

データセンタネットワークの変遷と新技術





Virtual Network への要求

- Virtual Network の自動構成
- VM-to-VM の Traffic Engineering
- VM数に対応するスケーラビリティ

Server-side Edge-Overlay

midokura

vmware

PLUMgrid

Stratosphere

Physical Network への要求

- 広帯域でPort数が多い物理 Network
- Port-to-Port の Traffic Engineering

Datacenter-Optimized L2

BROCADE VCS TECHNOLOGY

OFABRIC

Fabric Path CISCO

Software-Controlled Switch

OpenFlow

NEC

big switch networks

Data Center Optimized L2

多分、SDNじゃない

目的

✓ East-to-West トラフィック増に対応できる
広帯域なフラットL2の構築

代表技術

✓ TRILL, SPB に代表されるL2マルチパスによるSTPの排除

特徴

○ 何といってもL2は楽
○ L3の運用性・拡張性を
L2に適用

× 完全新規装置の導入
× プロプライエタリ実装

※ 参考: http://www.slideshare.net/ryuichitakashima3/enog-trilltakasima01

Software-Controlled Switch

多分、SDN

目的

- ✓ 最短経路転送以外のトラフィックエンジニアリング
- ✓ 外部からの操作によるネットワーク設定の自動化

代表技術

- ✓ OpenFlow を制御に利用するスイッチ・コントローラ

特徴

○ 自由度の高さ

× コントローラの作り込みの作業工数
× 別の制御の機構が必須

※ 参考: http://www.janog.gr.jp/meeting/janog29/program/openflow.html

Server-side Edge-Overlay

多分、SDN

目的

- ✓ 物理ネットワークと仮想ネットワーク資源の切り離し
- ✓ VM-to-VM に特化した制御

代表技術

- ✓ エッジオーバレイ + α、OpenFlow + α 等様々

特徴

- ○ 完全自動化
- ○ 仮想化の要求に対応した拡張性

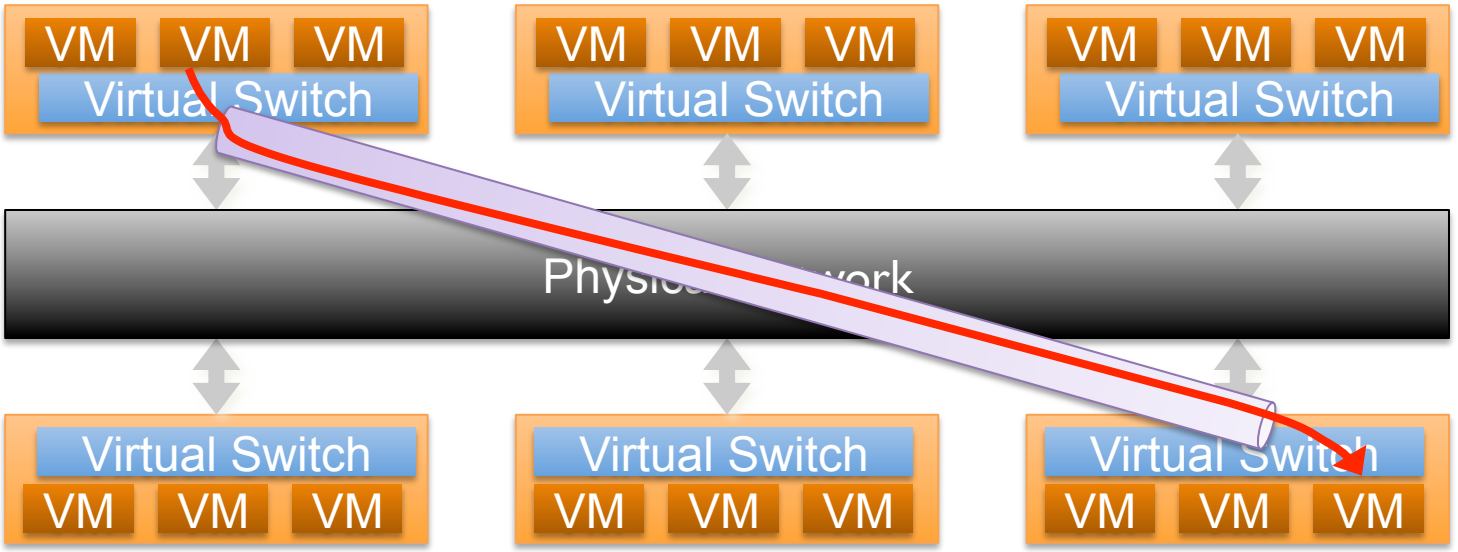
- × 用途が現状 IaaS に限定される
- × プロプライエタリ実装

Server-side Edge Overlay 実装例 / MidoNet の場合

https://commons.wikimedia.org/wiki/File:Plastic_tubing.jpg

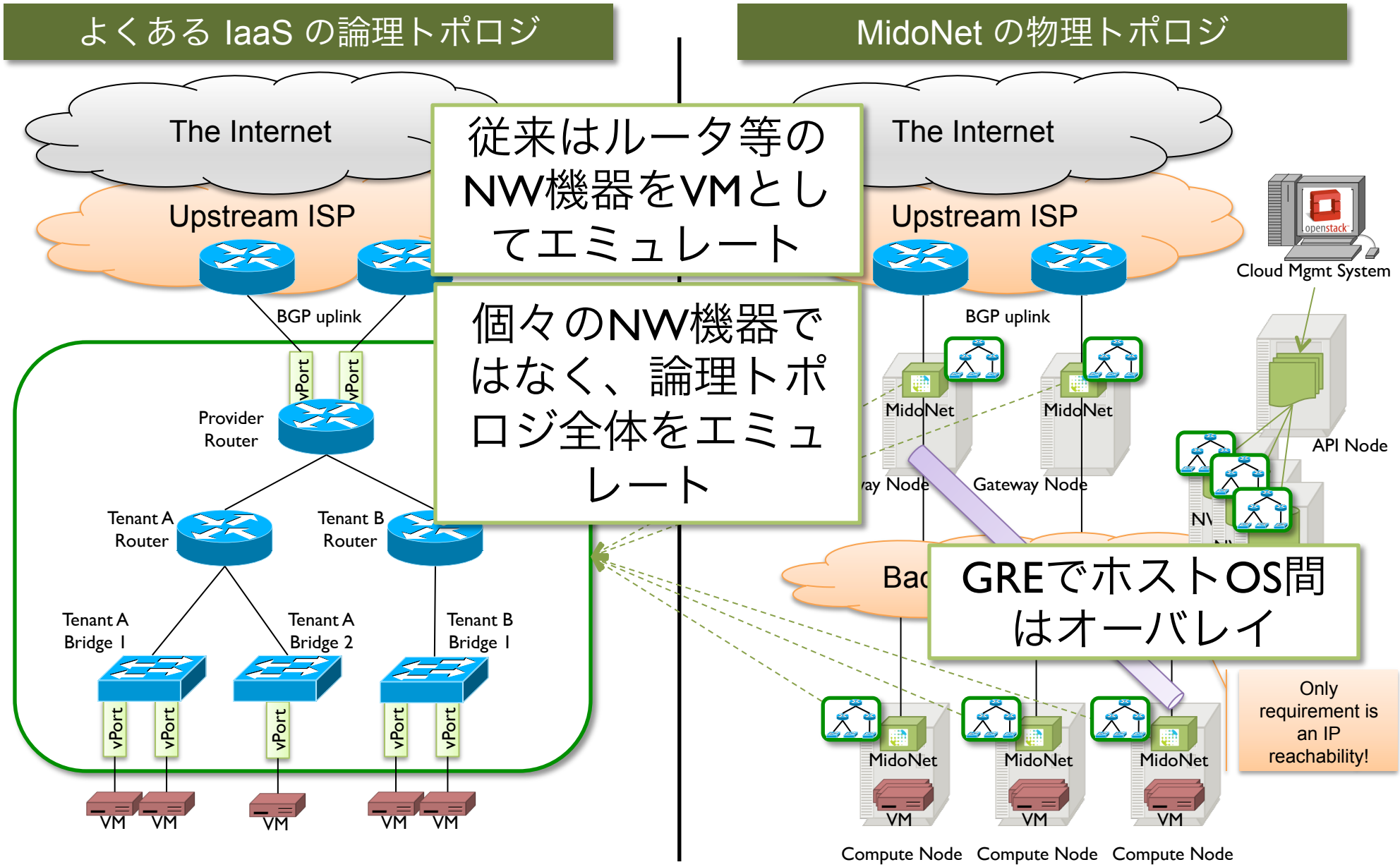
Server-side Edge overlay

概念図



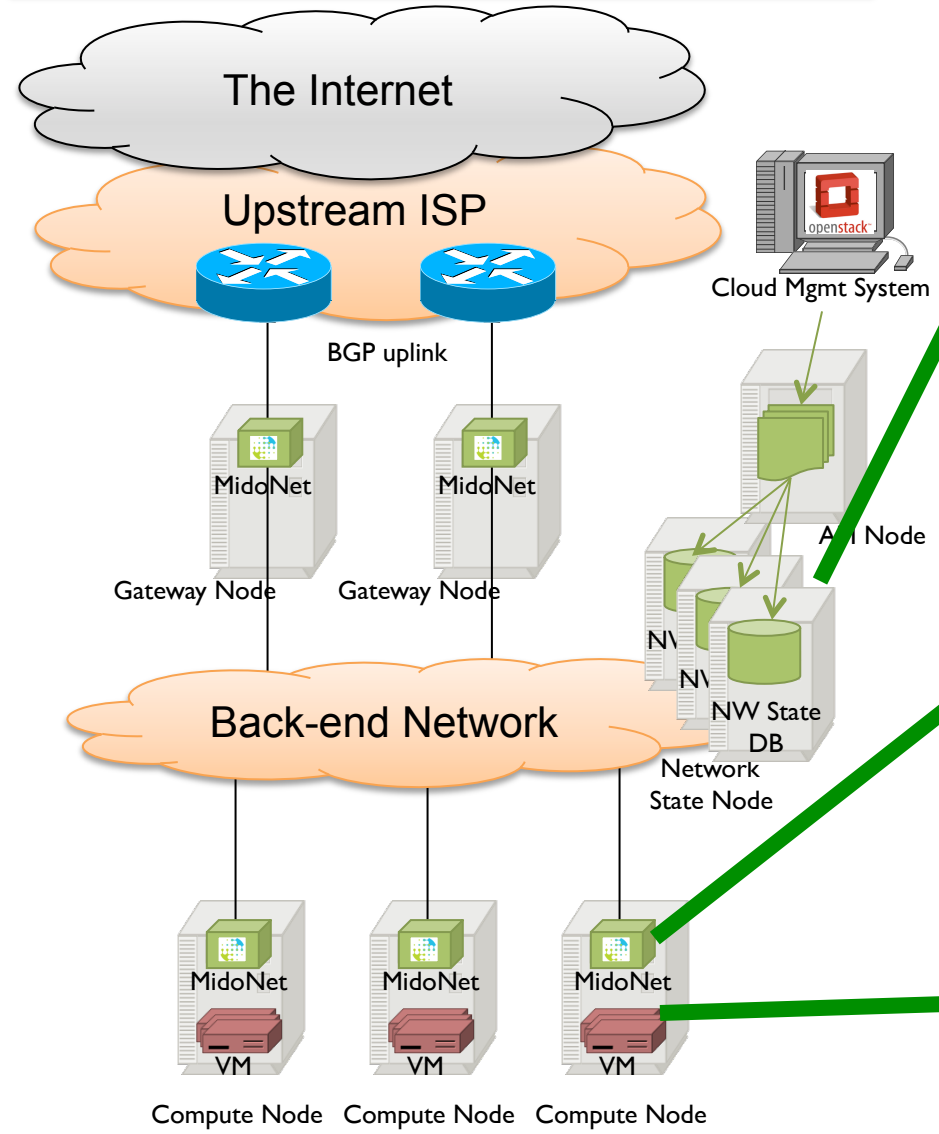
- GRE, NVGRE, VXLAN 等、IP based なトンネリングプロトコルを利用
- その為、Undelay となる Physical Network は IP Reachability さえあれば良い

MidoNet の場合 / 概要



MidoNet の場合 / 実装

MidoNet のコンポーネント



NSDB

Zookeeper, Cassandra.
トポロジ情報の保持、IP-MAC table、接続ホスト情報等の全体情報を持つ「コントローラ」ではなく「データベース」。プッシュ配信を極力行わない

Agent

ホストOSで動作するプロセス。
NSDBからオンデマンドで必要な情報をダウンロードしトポロジエミュレーションを実施。
結果を Data path にプログラミング。

Data path

ホストOS上のOVS kernel module

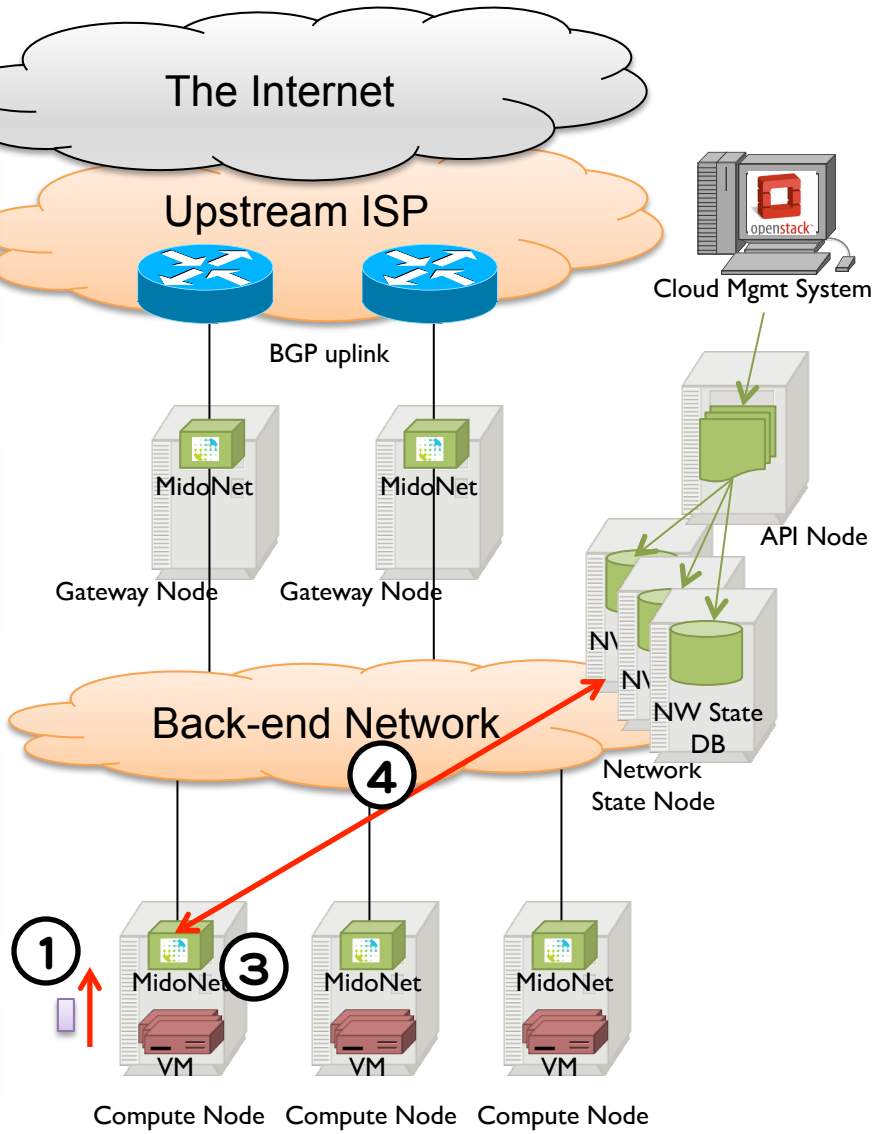
How does it work ? / For 1st packet

④ MidoNet Agent downloads information that is necessary for topology emulation of the packet.

③ MidoNet Agent checks local topology data. If it doesn't have enough data for it, then ④

② If OVS doesn't has an entry matches with the packet, then ③

① VM sends packet and OVS receives it.

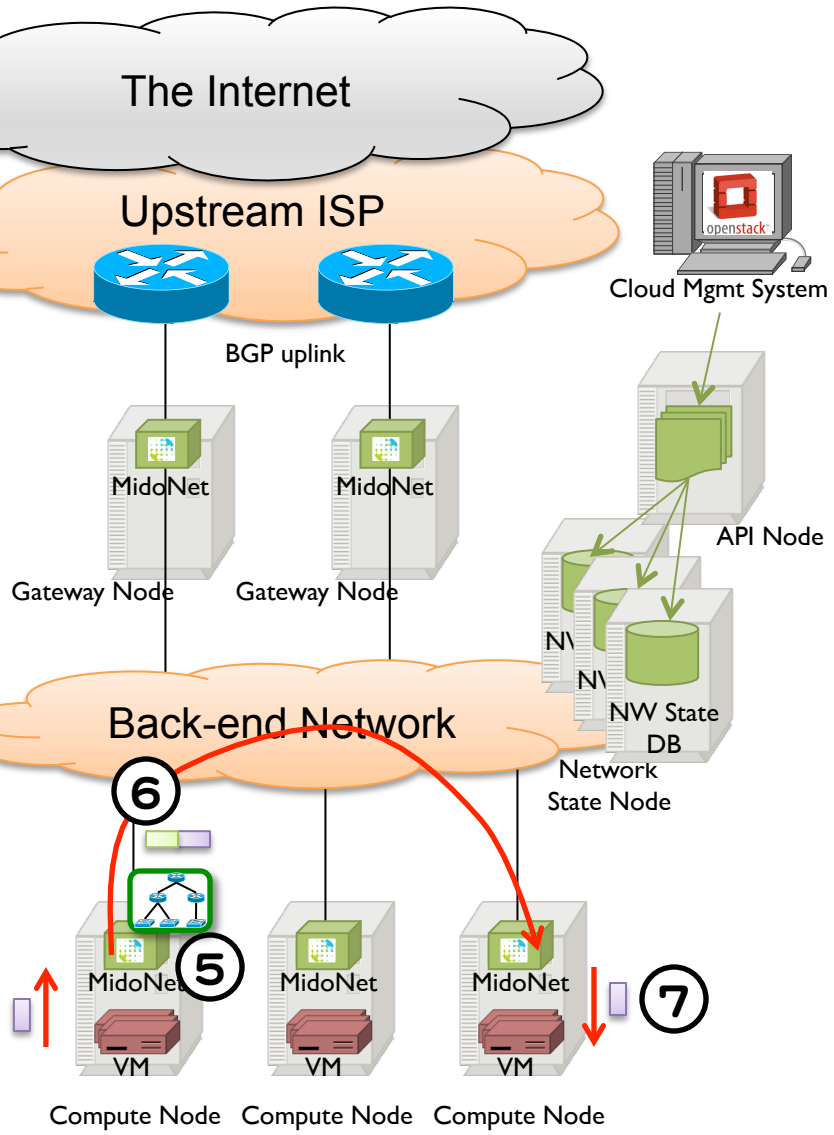


How does it work ? / For 1st packet

⑦ Remote OVS decaps GRE and forwards it to destination VM.

⑥ Local OVS modifies header, encapsulates GRE and forwards it.

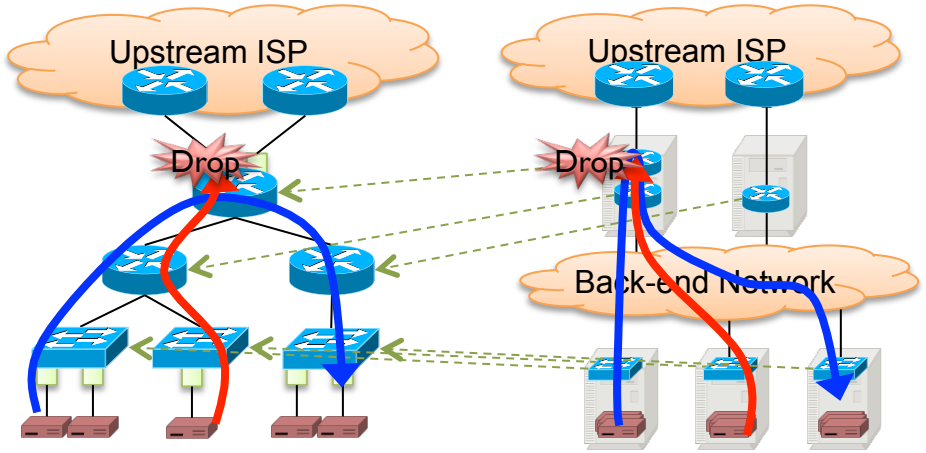
⑤ MidoNet Agent simulates the topology and programs local OVS data-path according to the result.



何がうれしいの？

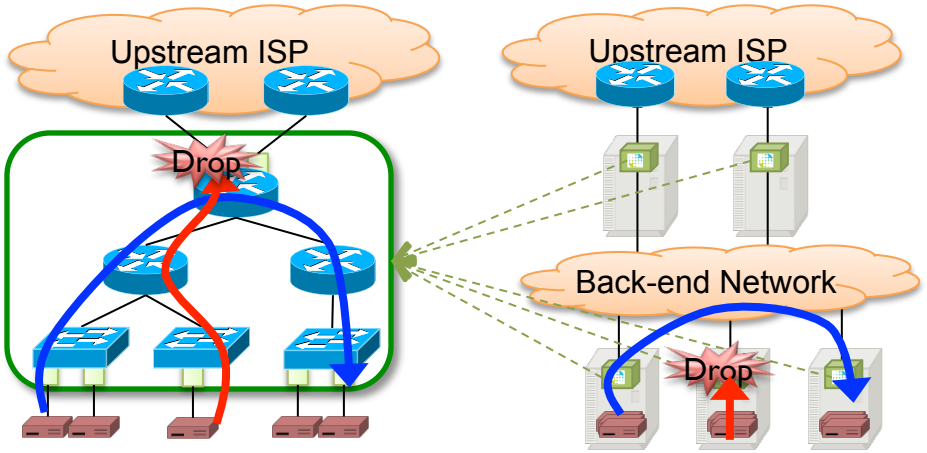
1. East-West トラフィックの最適化

従来のVirtual Router



Network機器をVMとしてエミュレーション

MidoNet

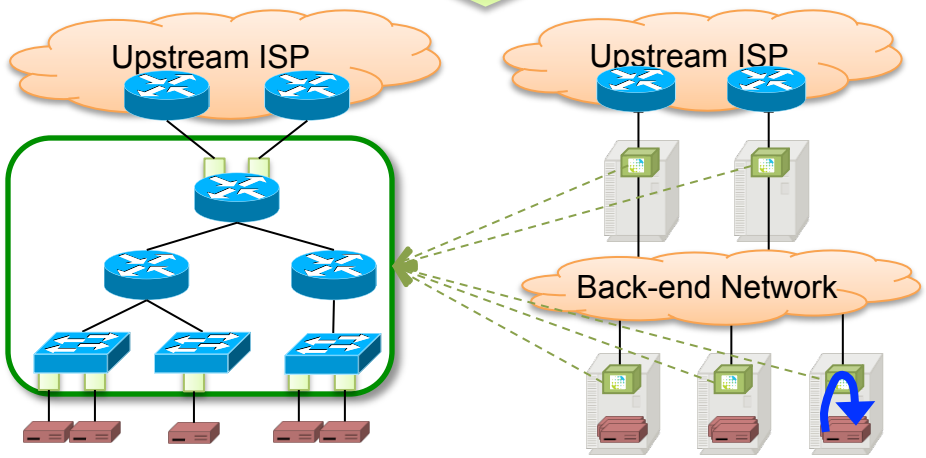


Network Topology全体をエミュレーション

Ingressでのトポロジエミュレートにより、目的地となるホストに直接転送する為、”行って来い” が発生しない

2. ボトルネックとなるルータVMの排除

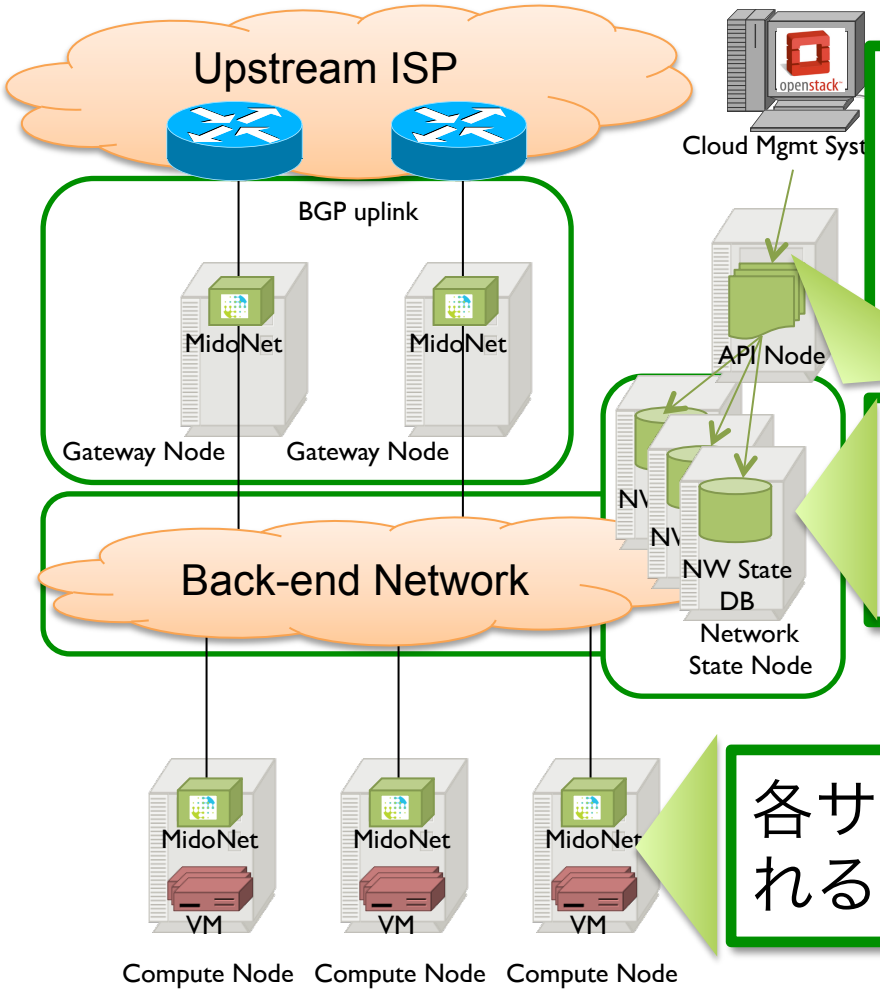
トポロジエミュレーションによりエッジで分散処理する為、ルータVMが存在しない



各MidoNet Agent が IP-MAC対応表を持ち、同一ホスト内VMからのARPに代理応答

VMではなく、ホストOSでのフォワーディング処理

3. スケーラビリティ



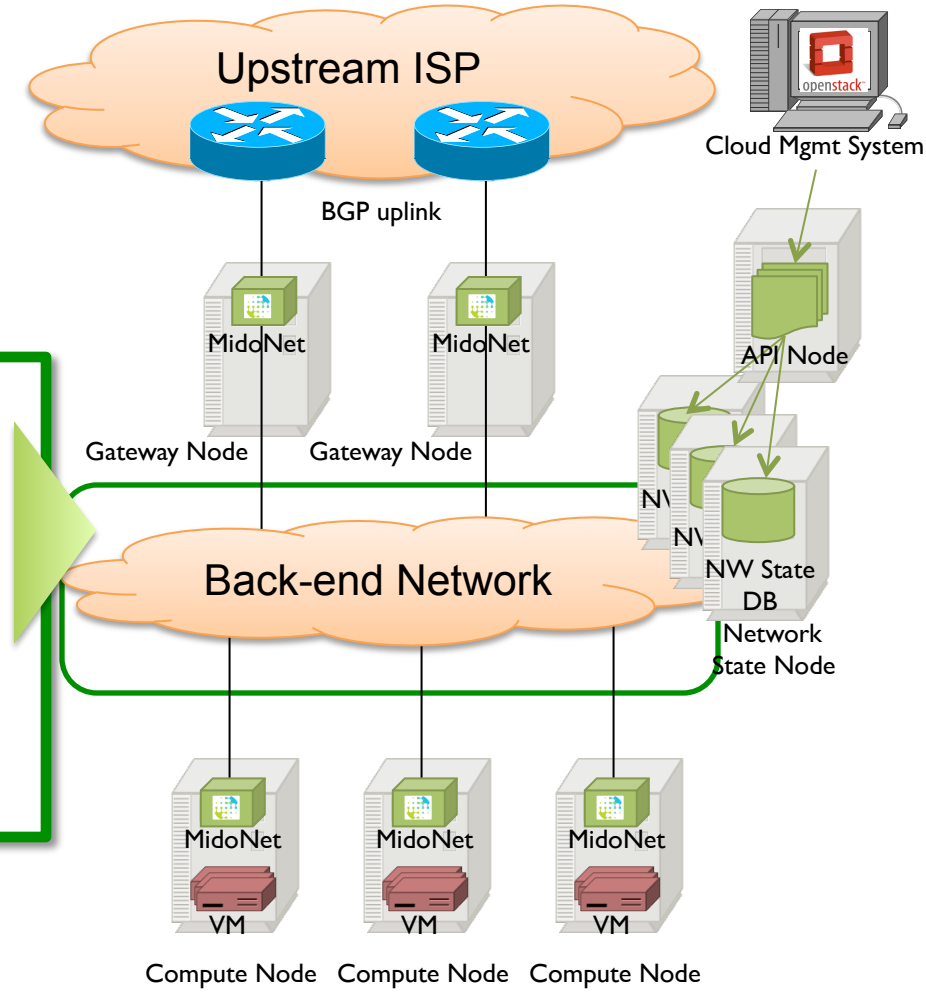
VLANを利用しないトポロジエミュレーションの為、4,096 の壁に制限されずスケールアウトが可能

全体のMAC table は分散DB上に格納される為、スケールアウト可能

各サーバのデータパスにプログラミングされるものは収容するVMの通信のみ

4. Underlay Network の簡易性

各 MidoNet Agent 間通信は GRE でカプセル化
→ IP Reachability さえあればよい
→ IGP を用いた L3 スケールアウトが利用可能



“データセンタ編”まとめ

✓データセンタネットワークに求められる
”自動化” “パフォーマンス” “拡張性”
を満たす為、物理ネットワーク拡張技術に加えて
“Server-side Edge overlay” が登場してきた

✓ネットワーク機器の低廉化の手段として、White Box 化が求められている。その実装として OpenFlow を始めとする SDN と呼ばれる技術が用いられる場合もある

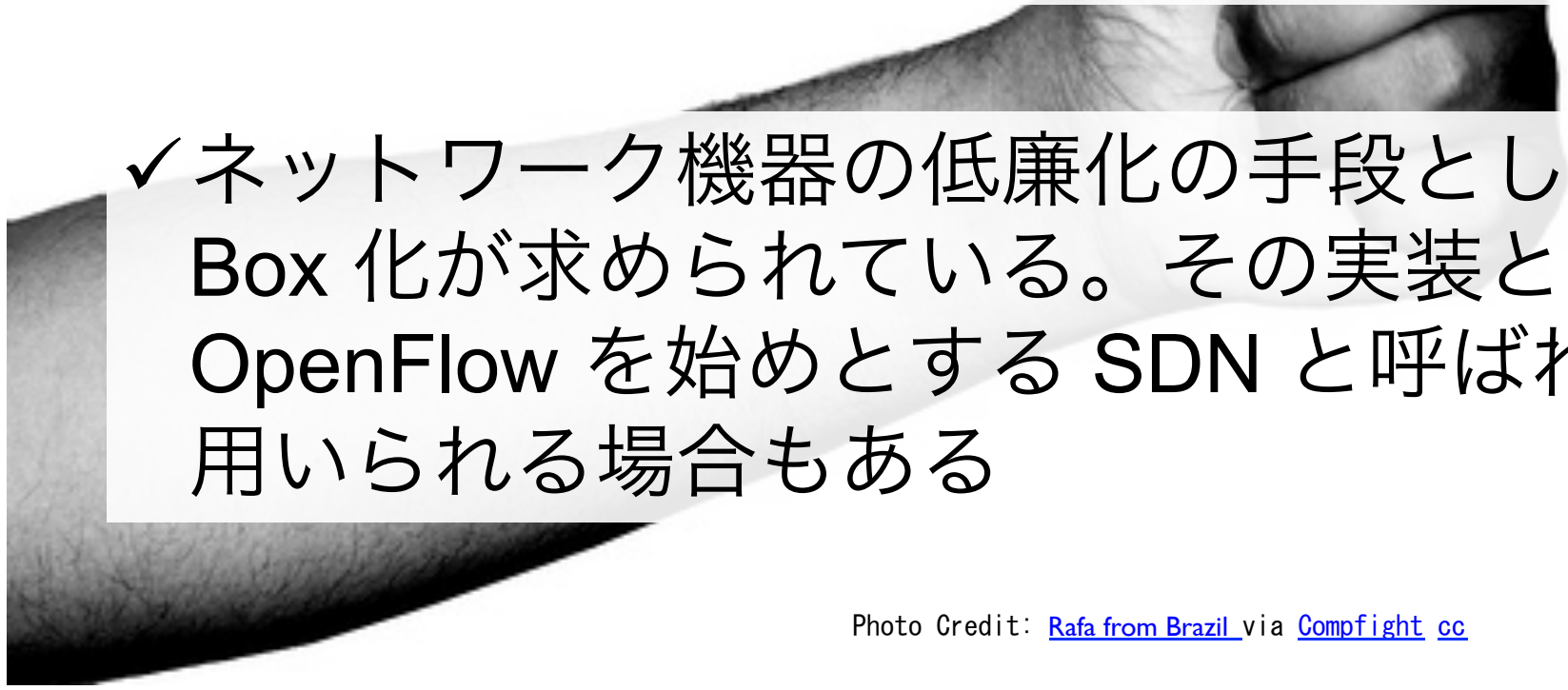


Photo Credit: Rafa from Brazil via Compfight cc

WAN・キャリア編

By 清水

Thank you!