OpenStackに最適なMidoNetの最新利用事例とオープンソースコミュニティの紹介

鈴木孝彰
システムエンジニア
ミドクラジャパン株式会社

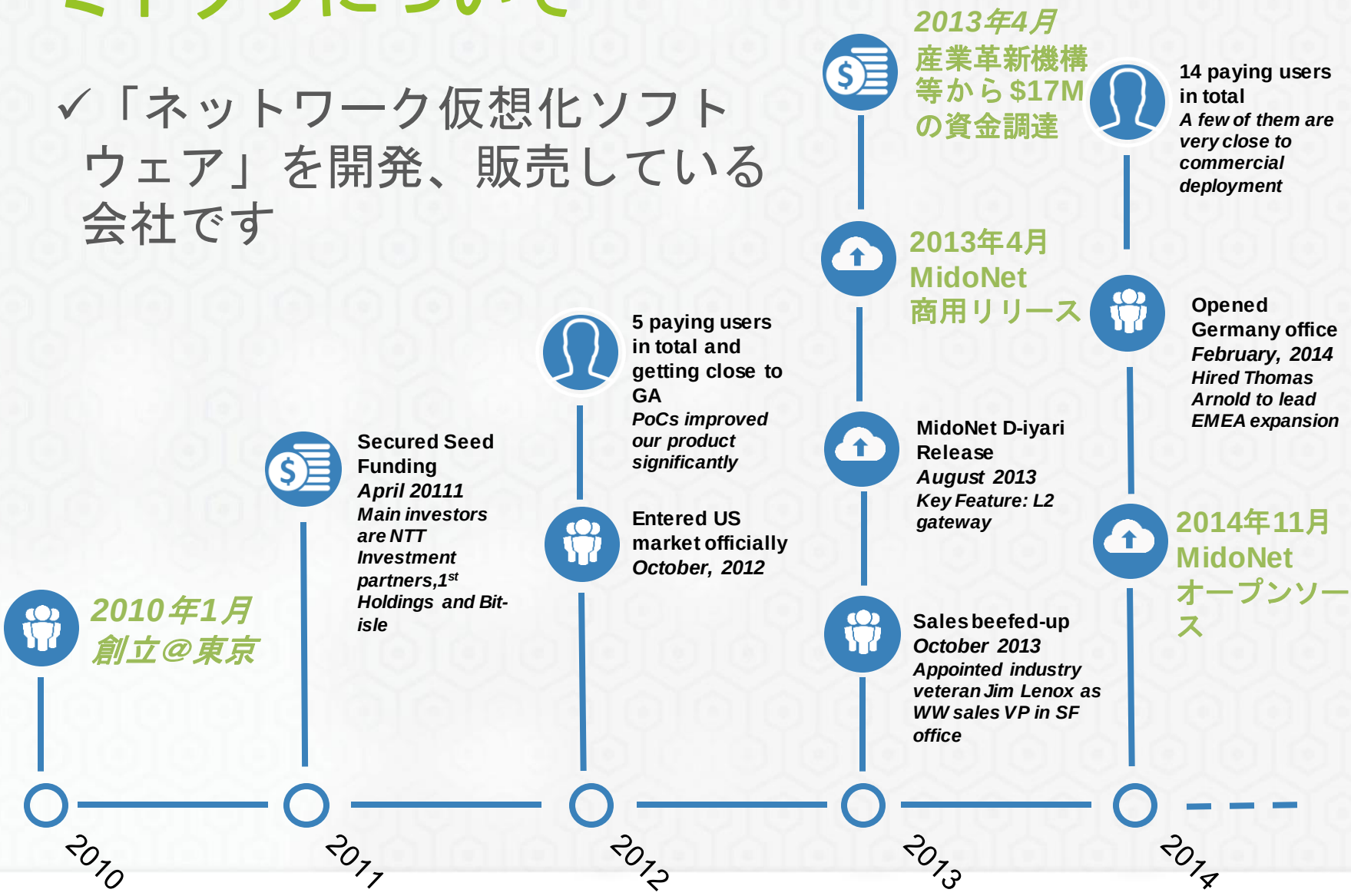
アジェンダ

- ✓ ミドクラ会社紹介
- ✓ MidoNetとは
- ✓ OpenStack + MidoNetの活用事例
- ✓ MidoNetのテクノロジー
- ✓ オープンソースになったMidoNet
- ✓ まとめ

ミドクラって誰？
～ミドクラ会社紹介～

ミドクラについて

✓「ネットワーク仮想化ソフトウェア」を開発、販売している会社です

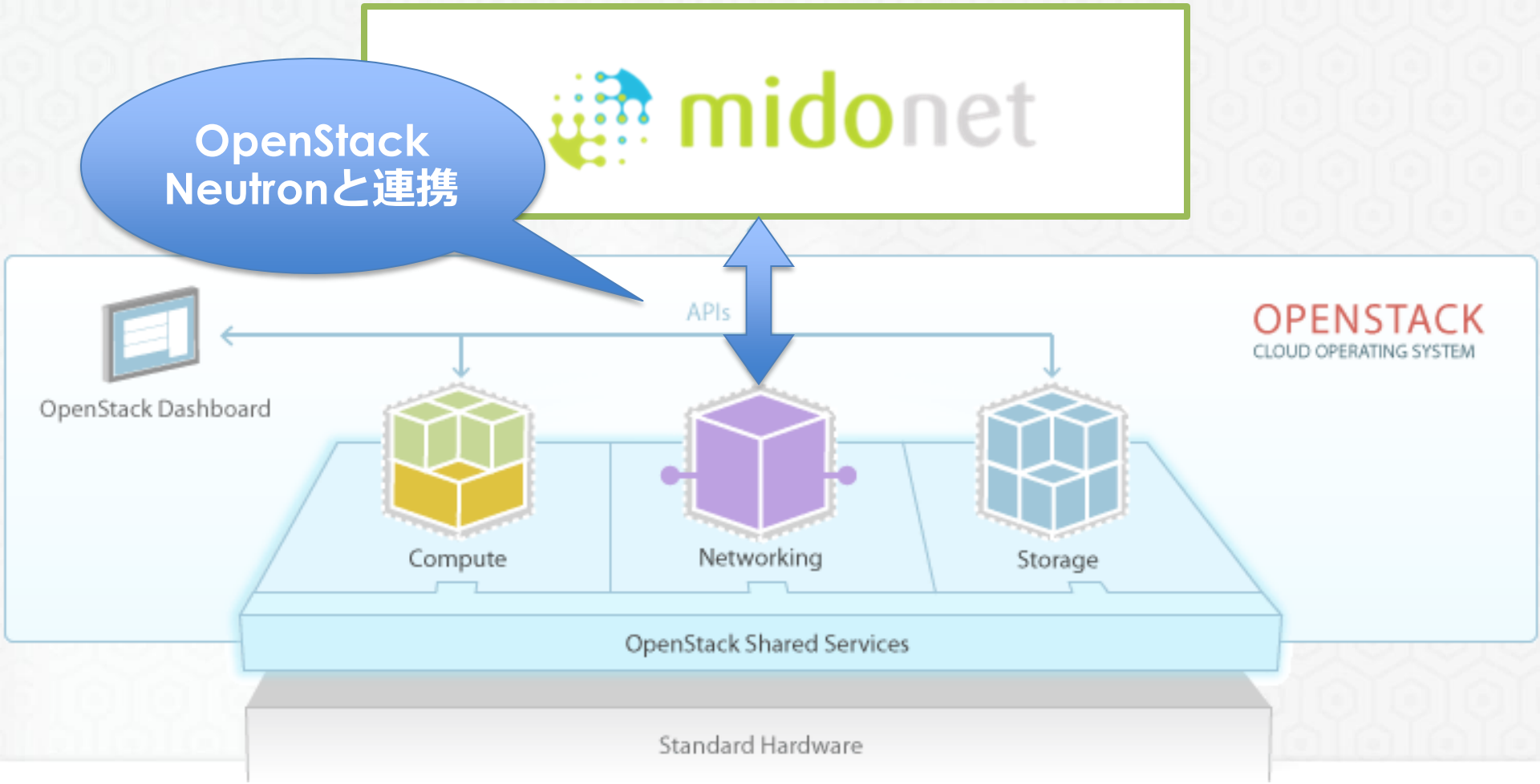


ミドクラについて

- ✓ 東京にも開発拠点を持ち、フロントからダイレクトに開発部隊にエスカレーション
- ✓ タイムゾーンが異なる3拠点に展開し、問題が発生した際も迅速な対応可能

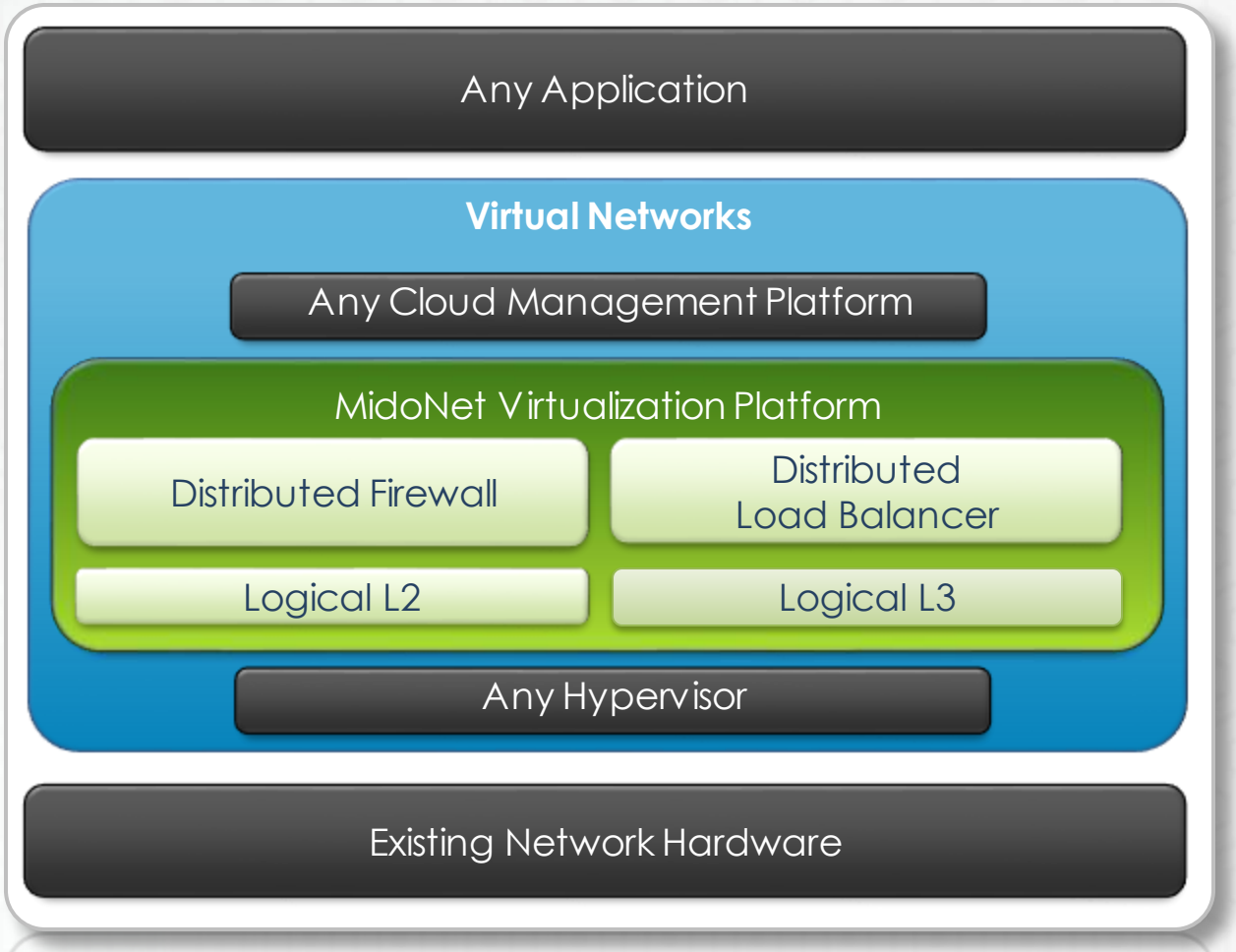
MidoNetとは？

OpenStackに最適な
ネットワーク仮想化ソリューション



source: http://www.openstack.org/software/

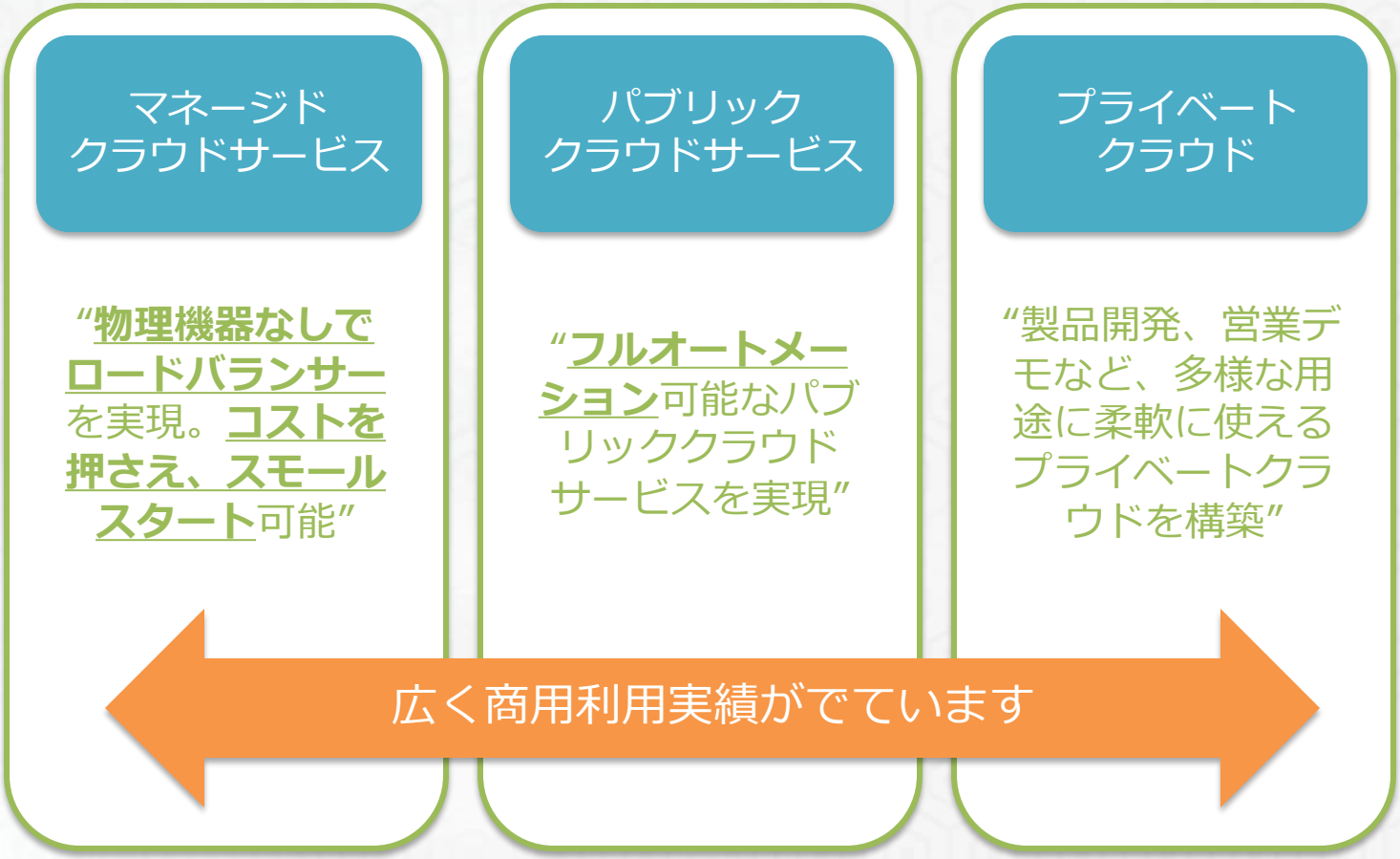
スケーラビリティおよび耐障害性を持つOpenStackネットワークを実現



OpenStackやMidoNetは、商用利用できるの？

～OpenStack + MidoNetの活用事例～

OpenStack + MidoNetの活用事例



マネージドクラウドサービス
KVH様 “プライベートクラウド TypeS”

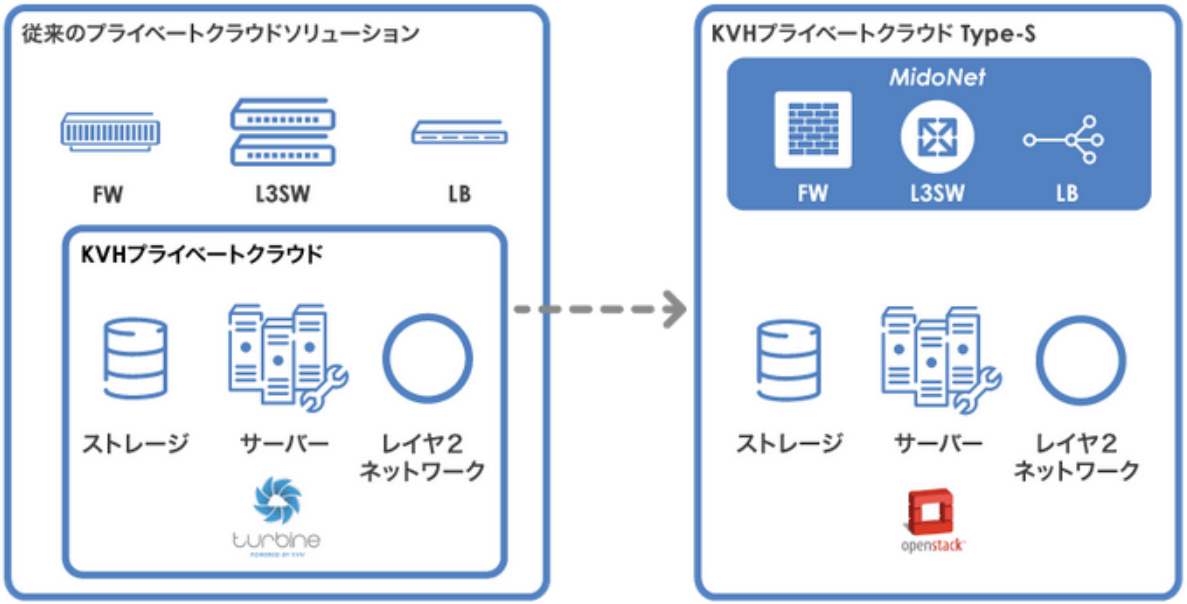
Asia's Information Delivery Platform

サービス概要

- 高価なネットワーク機器の購入やマニュアル作業による設定無しで、FW, LB, L23/スイッチングなどのネットワーク機能を実現
- 従来サービス比で、20%程度のコスト削減を見込む
- プライベートクラウドを小規模から構築しやすく

MidoNet 採用理由

- OpenStack Neutronとの互換性
- L4LBをはじめとしたネットワークサービスの充実



KVHホームページより抜粋

パブリッククラウドサービス
Zetta IOテクノロジー様 “zetta io”

zetta.io

サービス概要

- ITサービス企業が商用環境で利用できるパブリッククラウドをノルウェーで提供
- 全機能が自動化され、ユーザーによるセルフサービスが可能
- MidoNetを使い、仮想ネットワークサービス（L2, L3, FW, LB)を提供

MidoNet 採用理由

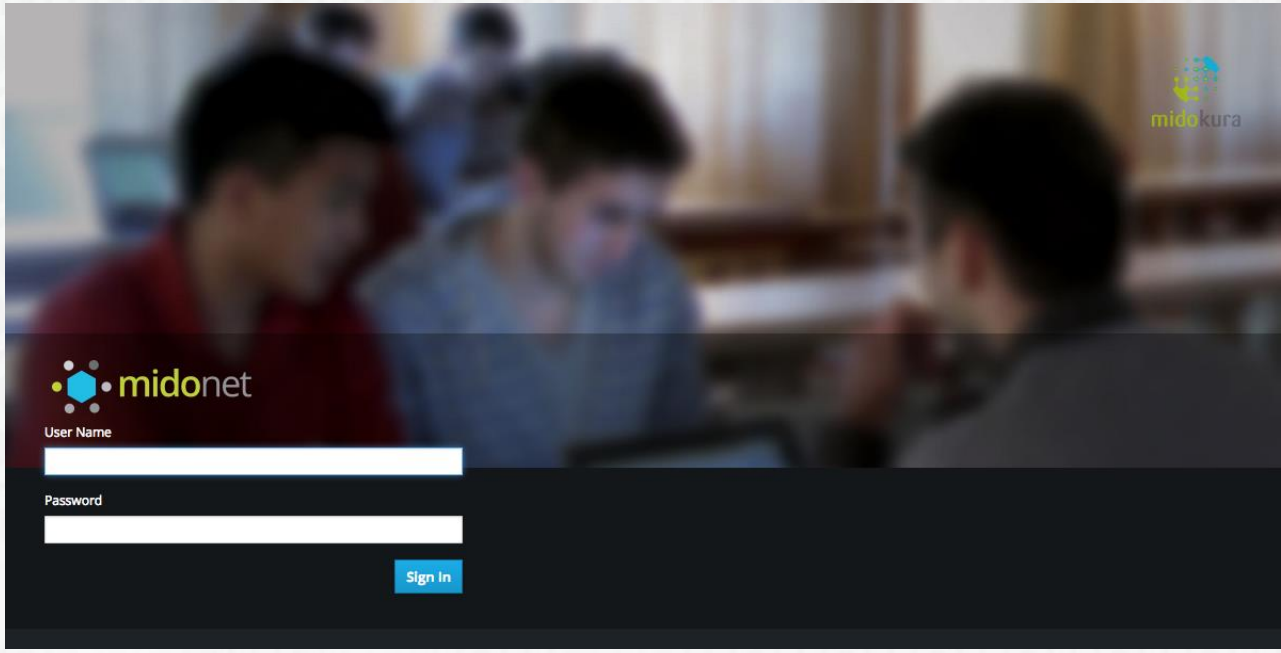
- 「顧客ニーズに合わせた迅速な製品化、安定性、拡張性」というzetta.ioの要件を満たした製品
- PoCを通じ、ネットワーク仮想化ソフトウェアで最も優れていると判断

プライベートクラウド
ミドクラ “MidoCloud”

サービス概要

- ミドクラ社内用の
  プライベートクラウド
- 2014/01から運用実績
- HavanaからIcehouseへの
  アップグレード実績あり
  ダウンタイムなし！

- システム物理構成
  - サーバー：30台
  −物理コア640
  - ストレージ：120TB
  −10Gスイッチ40ポート2台



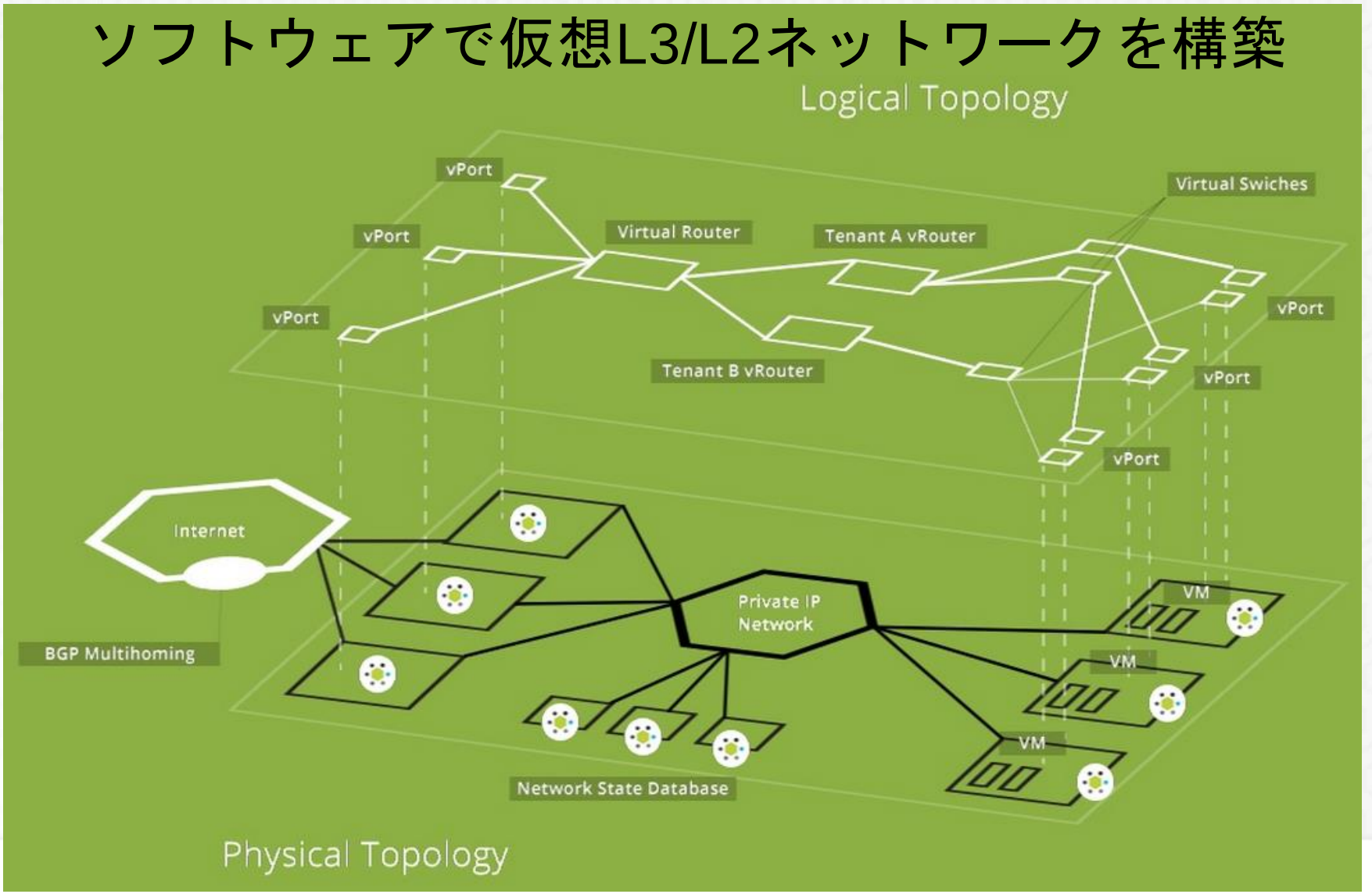
powered by RED HAT® ENTERPRISE LINUX® OPENSTACK® PLATFORM

midonet

MidoNetのテクノロジー

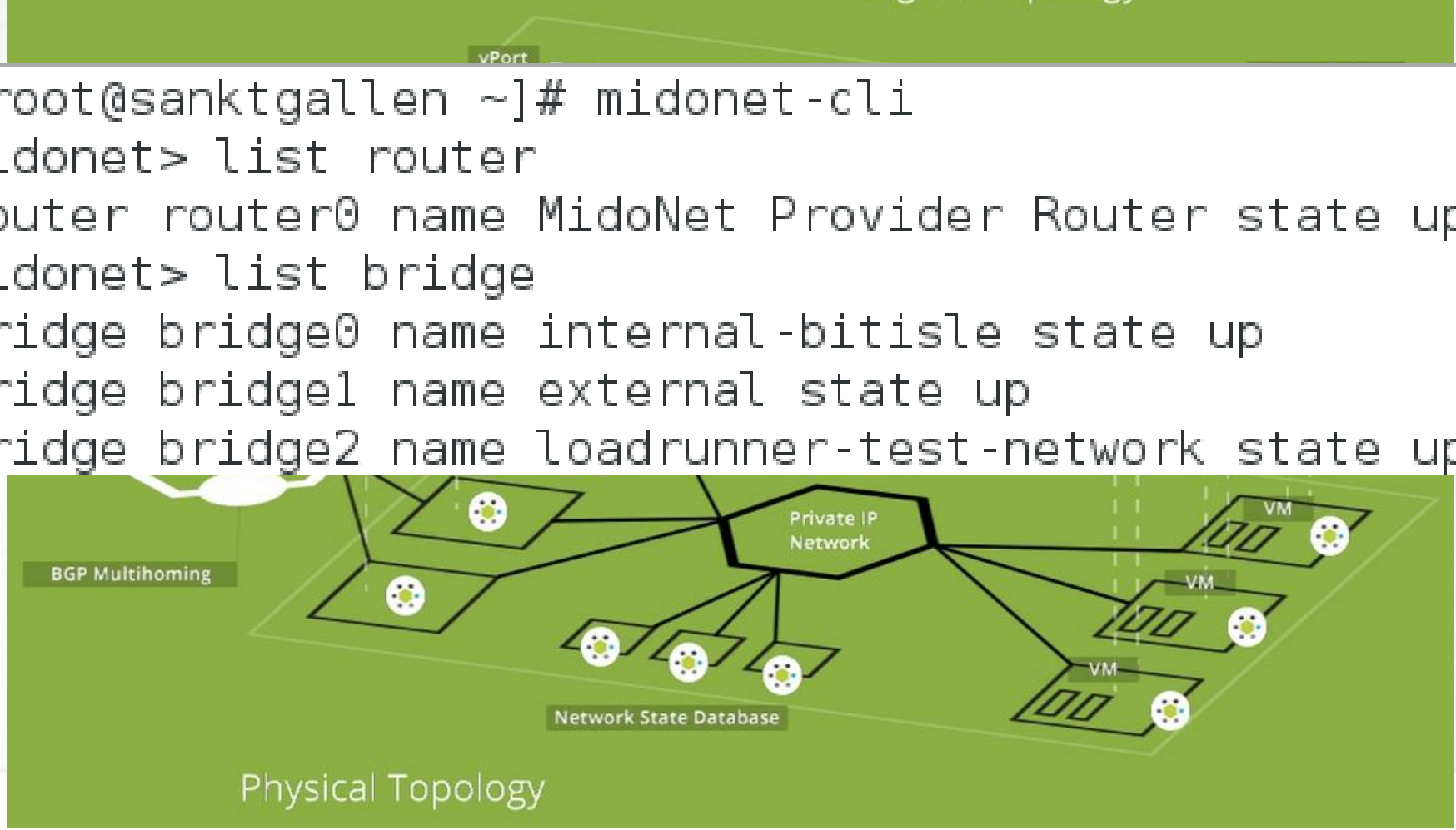
エッジオーバレイ

ソフトウェアで仮想L3/L2ネットワークを構築



エッジオーバレイ

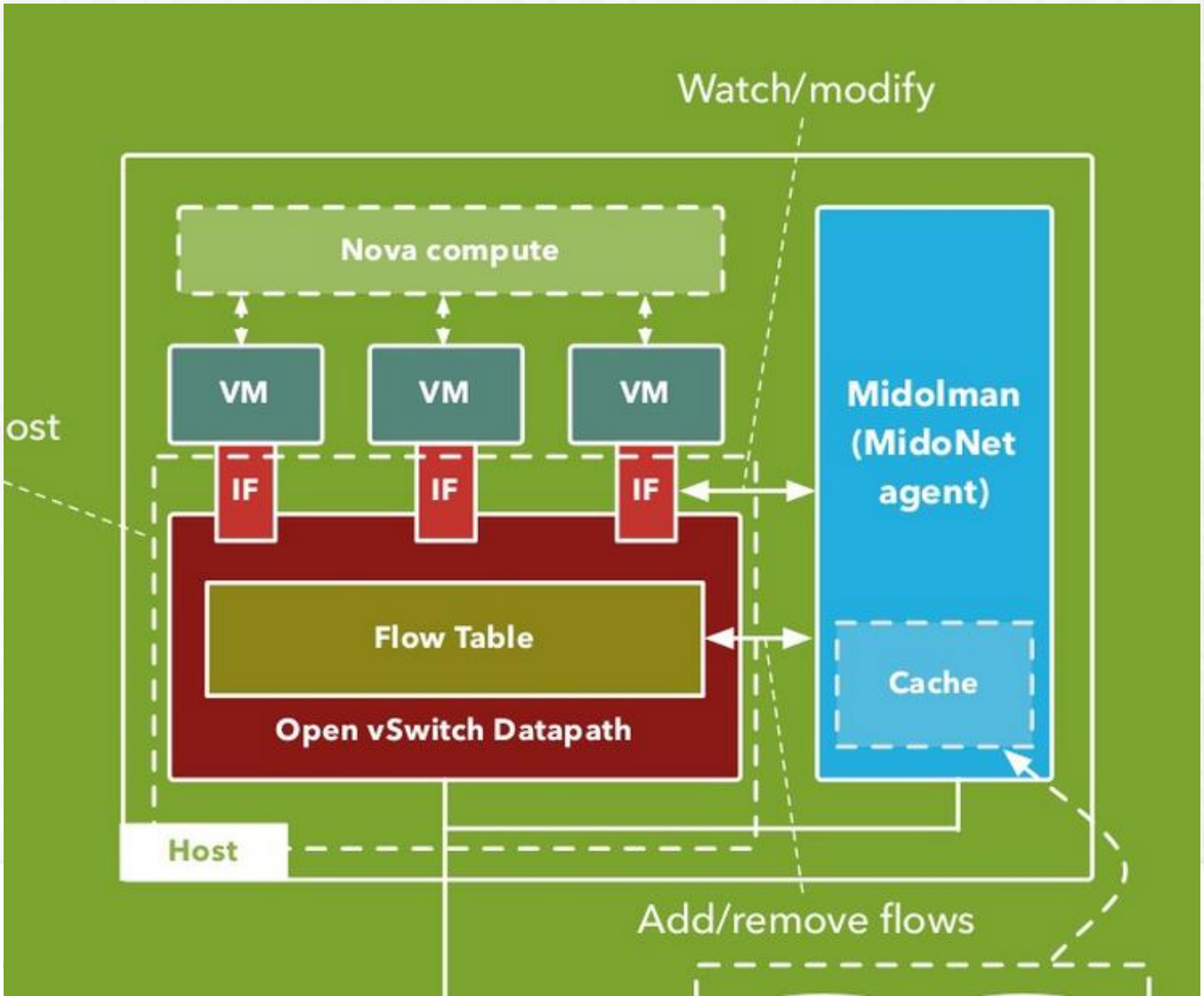
ソフトウェアで仮想L3/L2ネットワークを構築




```
[root@sanktgallen ~]# midonet-cli
midonet> list router
router router0 name MidoNet Provider Router state up
midonet> list bridge
bridge bridge0 name internal-bitisle state up
bridge bridge1 name external state up
bridge bridge2 name loadrunner-test-network state up
```

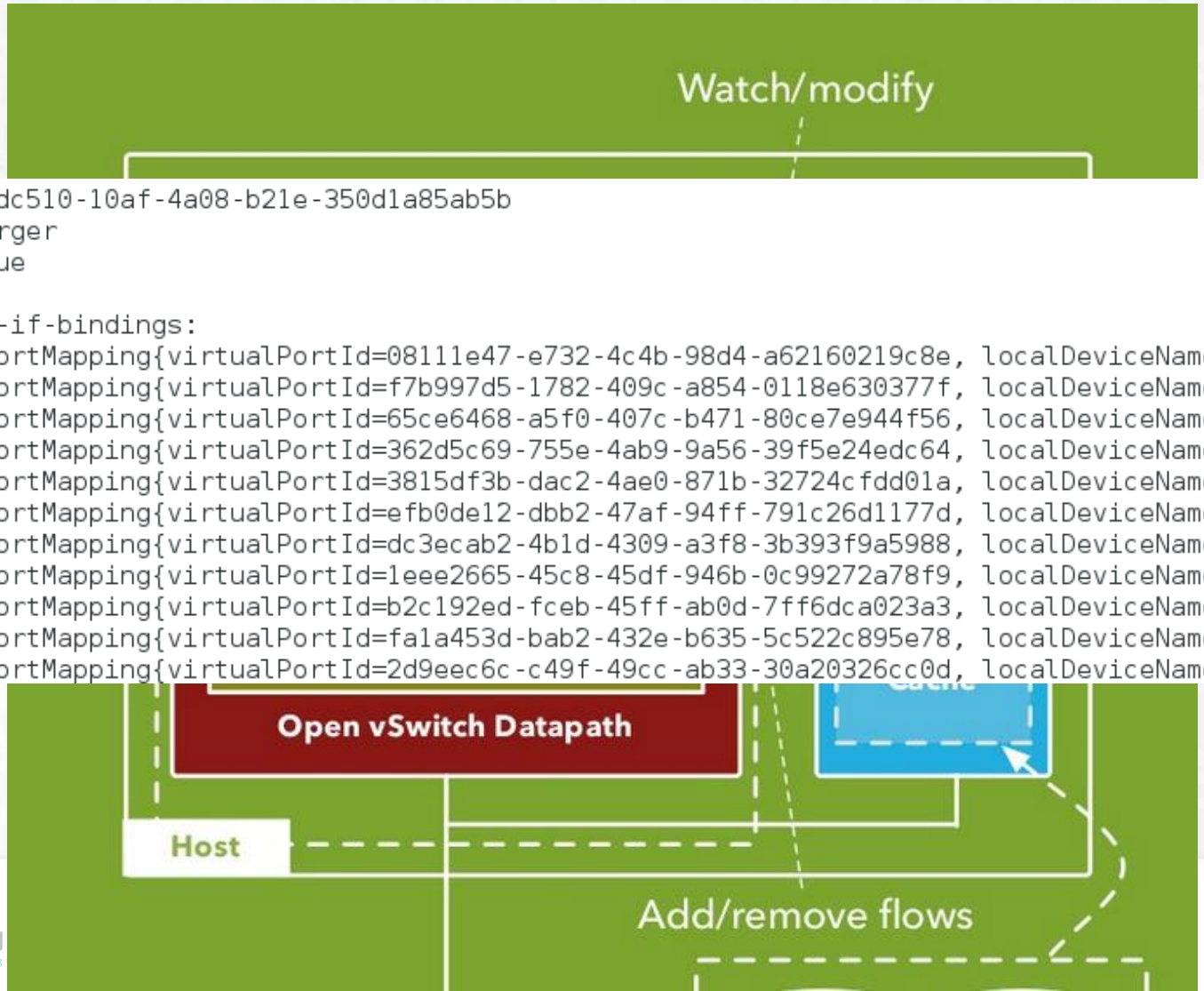
分散アーキテクチャ

- “エッジ”にコントローラ機能をおくことで、無駄なトラフィックが発生しない



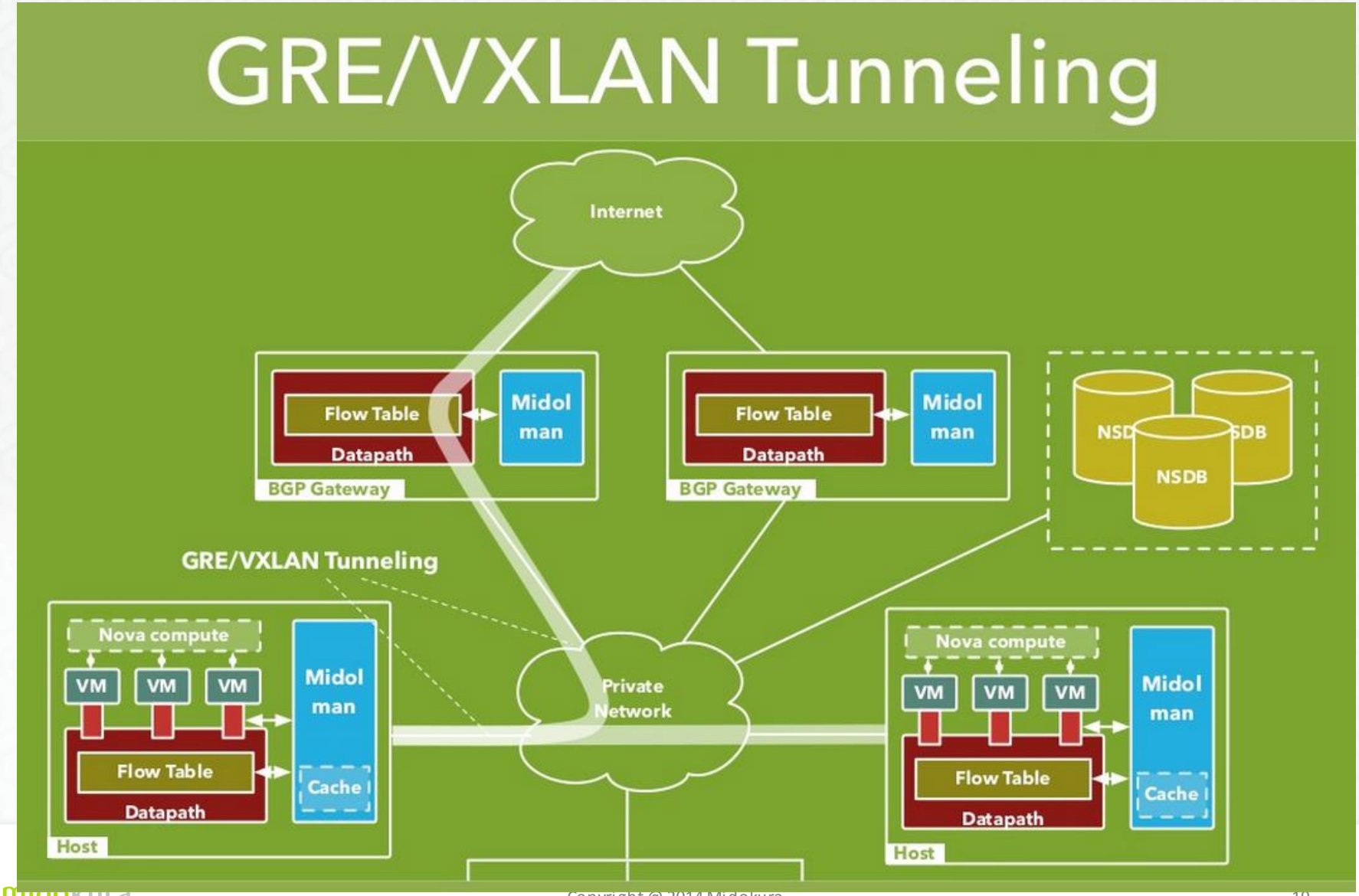
分散アーキテクチャ

- “エッジ”にコントローラ機能をおくことで、無駄なトラフィックが発生しない

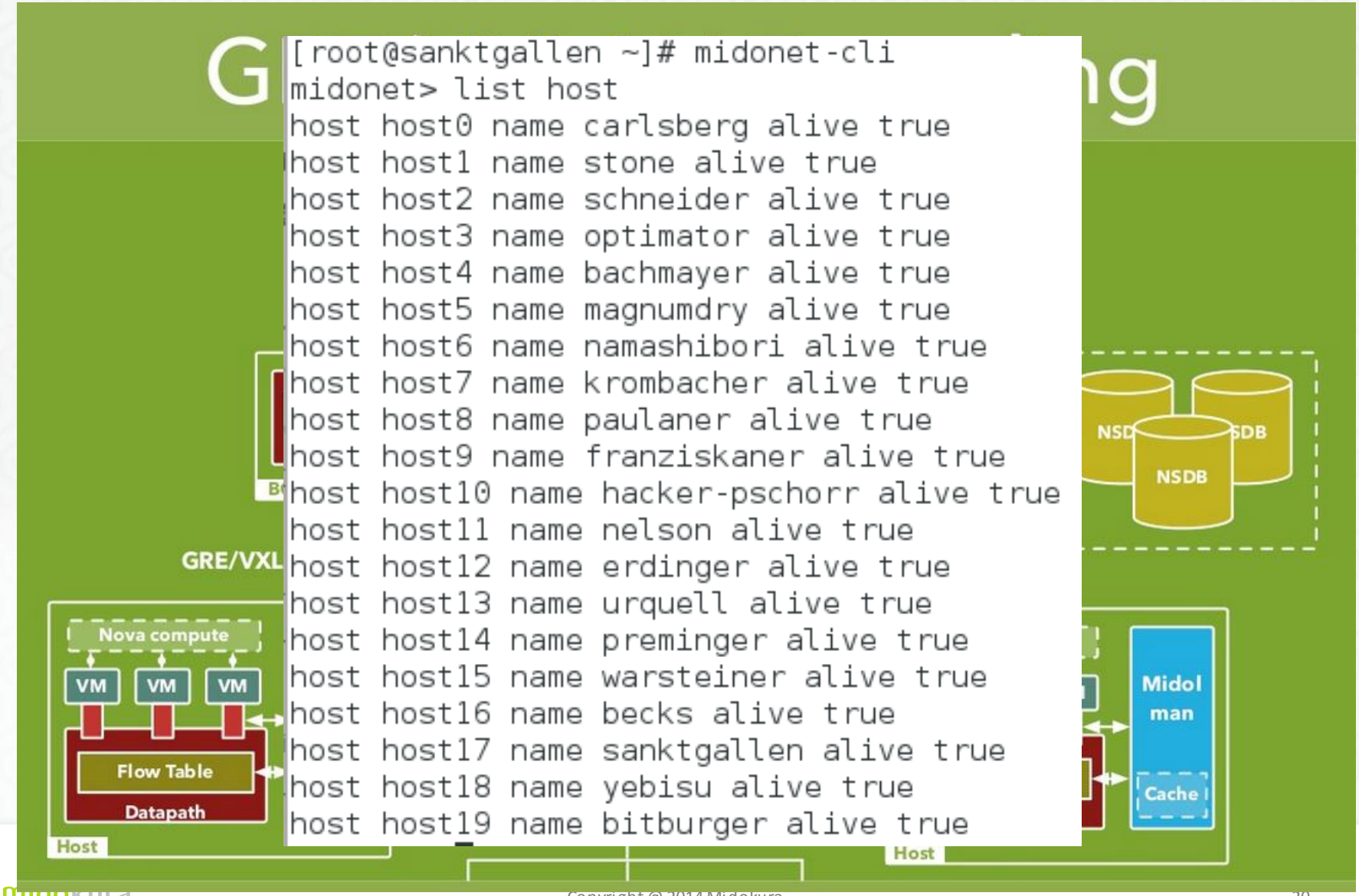



```
Host: id=3cedc510-10af-4a08-b21e-350d1a85ab5b
  name=bitburger
  isAlive=true
  addresses:
  vport-host-if-bindings:
    VirtualPortMapping{virtualPortId=08111e47-e732-4c4b-98d4-a62160219c8e, localDeviceName='tap08111e47-e7'}
    VirtualPortMapping{virtualPortId=f7b997d5-1782-409c-a854-0118e630377f, localDeviceName='tapf7b997d5-17'}
    VirtualPortMapping{virtualPortId=65ce6468-a5f0-407c-b471-80ce7e944f56, localDeviceName='tap65ce6468-a5'}
    VirtualPortMapping{virtualPortId=362d5c69-755e-4ab9-9a56-39f5e24edc64, localDeviceName='tap362d5c69-75'}
    VirtualPortMapping{virtualPortId=3815df3b-dac2-4ae0-871b-32724cfdd01a, localDeviceName='tap3815df3b-da'}
    VirtualPortMapping{virtualPortId=efb0de12-dbb2-47af-94ff-791c26d1177d, localDeviceName='tapefb0de12-db'}
    VirtualPortMapping{virtualPortId=dc3ecab2-4b1d-4309-a3f8-3b393f9a5988, localDeviceName='tapdc3ecab2-4b'}
    VirtualPortMapping{virtualPortId=1eee2665-45c8-45df-946b-0c99272a78f9, localDeviceName='tap1eee2665-45'}
    VirtualPortMapping{virtualPortId=b2c192ed-fceb-45ff-ab0d-7ff6dca023a3, localDeviceName='tapb2c192ed-fc'}
    VirtualPortMapping{virtualPortId=fa1a453d-bab2-432e-b635-5c522c895e78, localDeviceName='tapfa1a453d-ba'}
    VirtualPortMapping{virtualPortId=2d9eec6c-c49f-49cc-ab33-30a20326cc0d, localDeviceName='tap2d9eec6c-c4'}
```

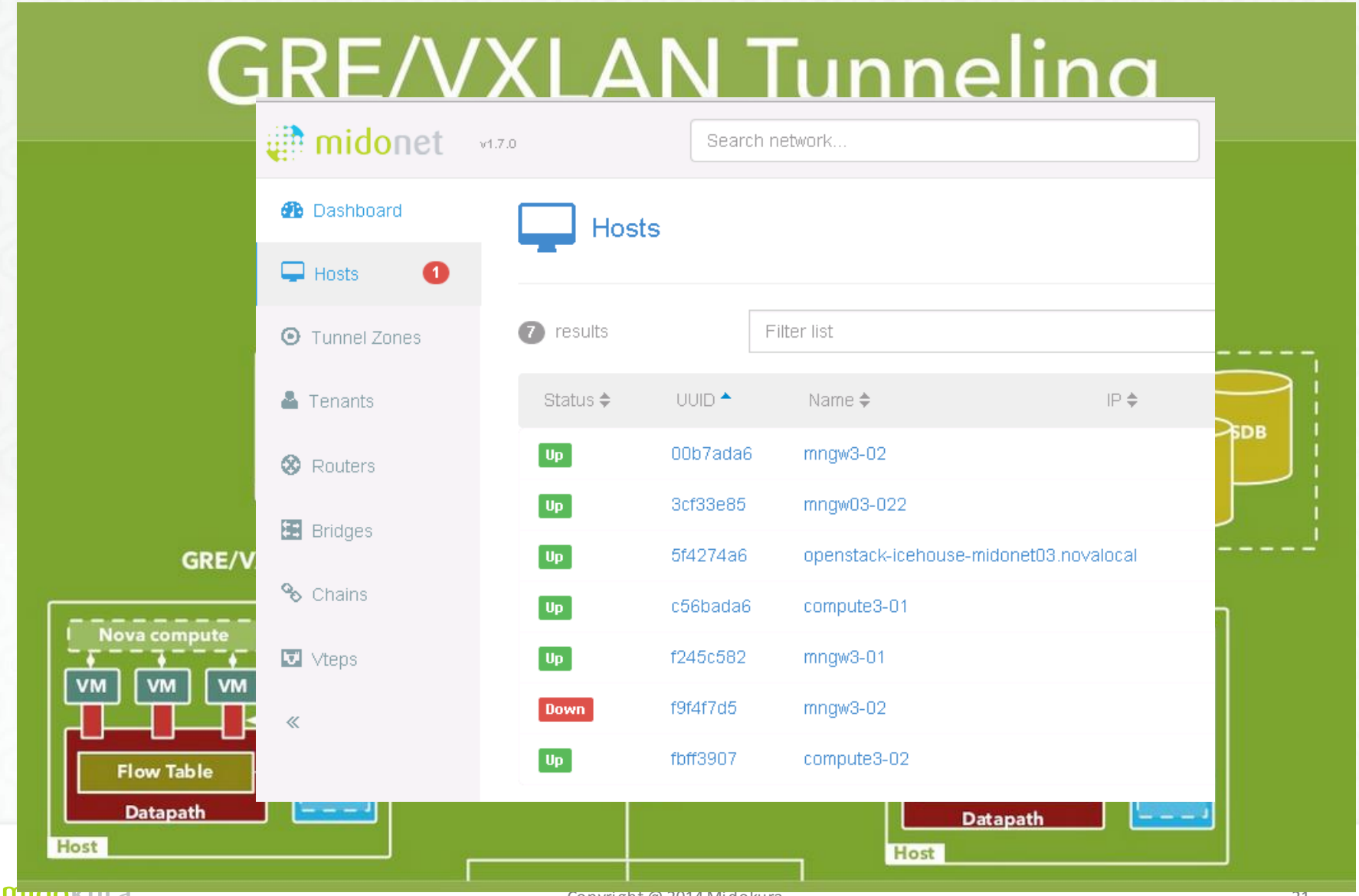
分散アーキテクチャ


GRE/VXLAN Tunneling
Internet
Flow Table
Datapath
Midol man
BGP Gateway
Flow Table
Datapath
Midol man
BGP Gateway
NSDB
NSDB
GRE/VXLAN Tunneling
Nova compute
VM
VM
VM
Midol man
Flow Table
Datapath
Cache
Host
Private Network
Nova compute
VM
VM
VM
Midol man
Flow Table
Datapath
Cache
Host

分散アーキテクチャ

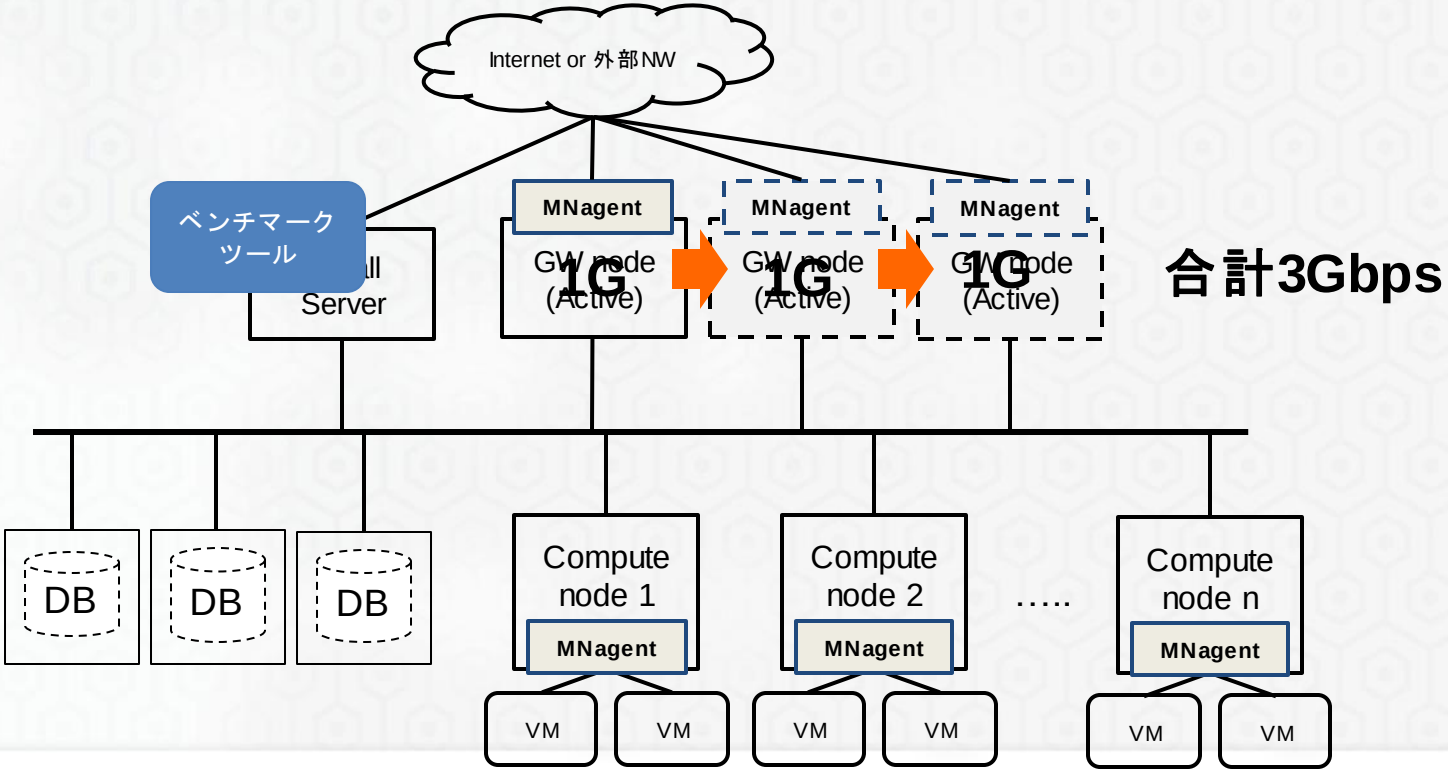


分散アーキテクチャ



スケーラビリティの評価

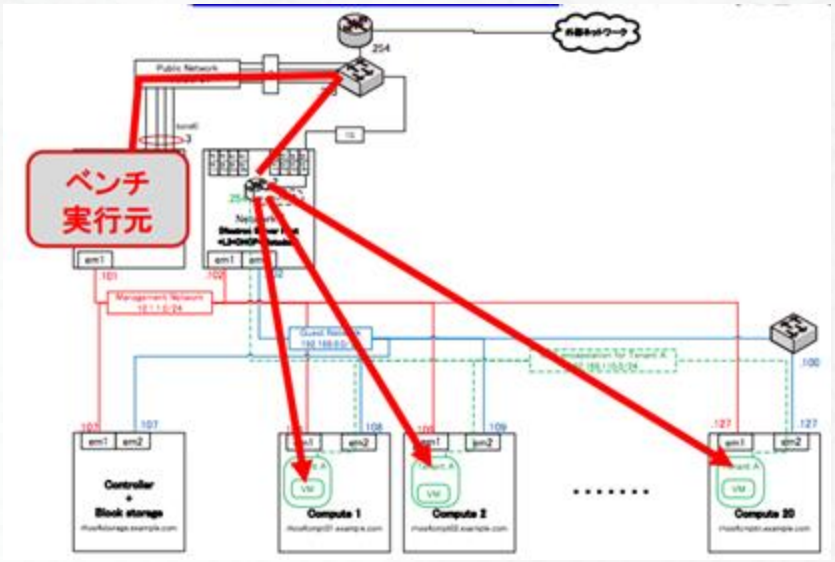
- 外部NWを経由して8VMに同時接続した時のスループットを測定.
- 外部ネットワークを1本、2本、3本と増やした時のスループットを測定し、分散アーキテクチャによるスケーラビリティを確認.



スケーラビリティの評価 (OpenStack標準OpenvSwitch)

- 外部ネットワーク1本

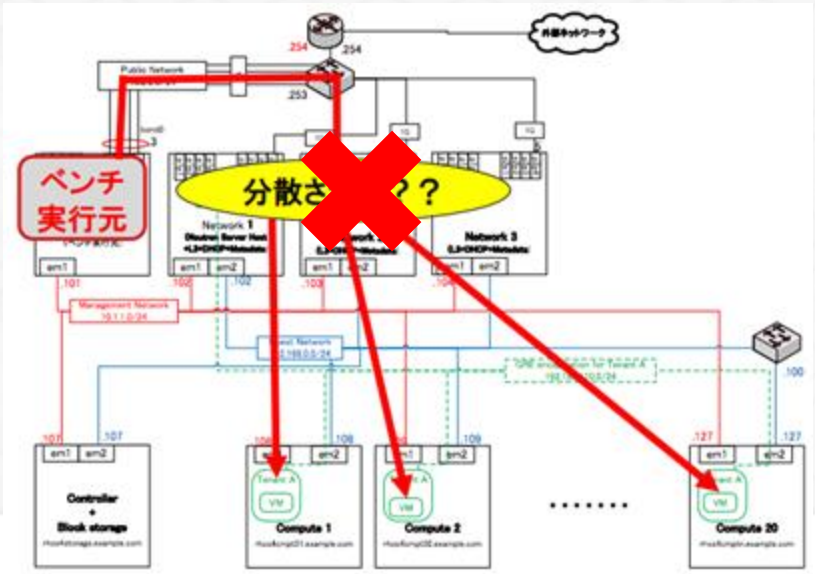
870Mbps



- 外部ネットワーク3本

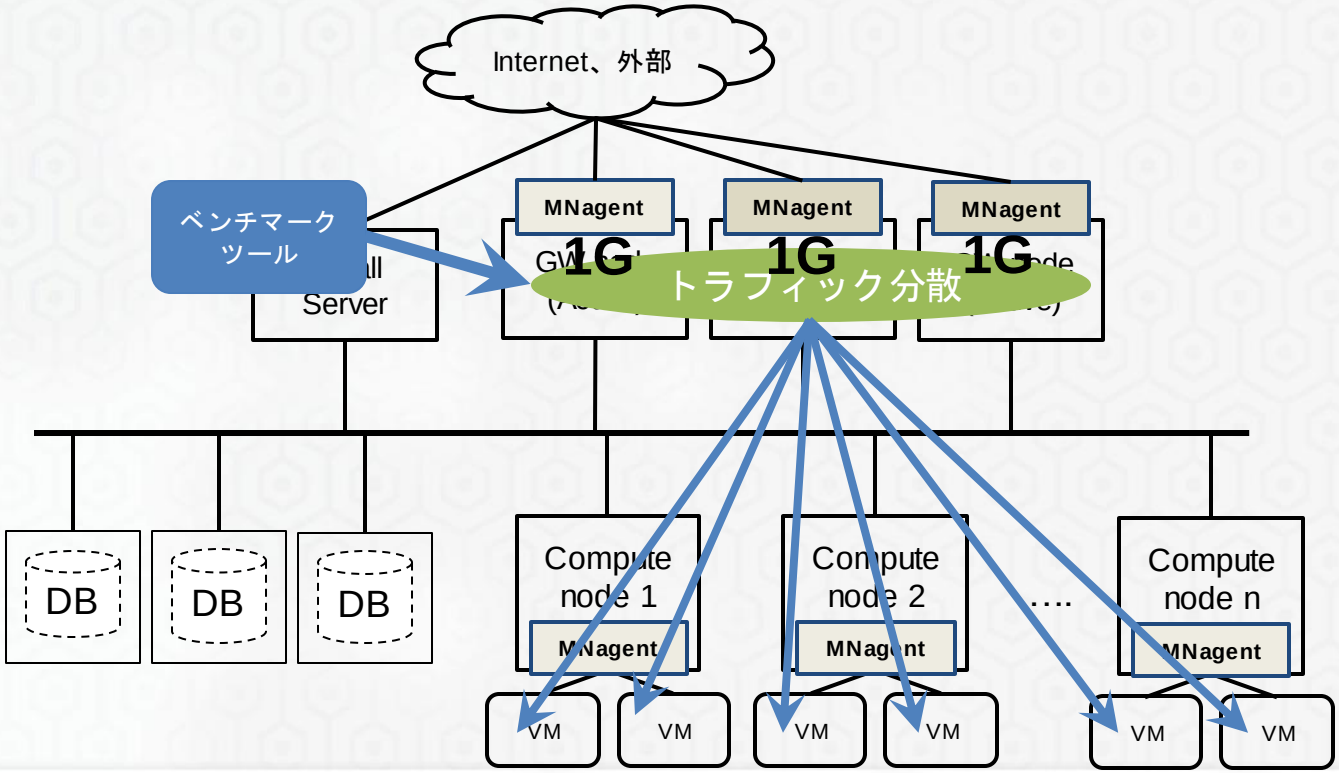
870Mbps

トラフィックを自動的に
分散することはできない



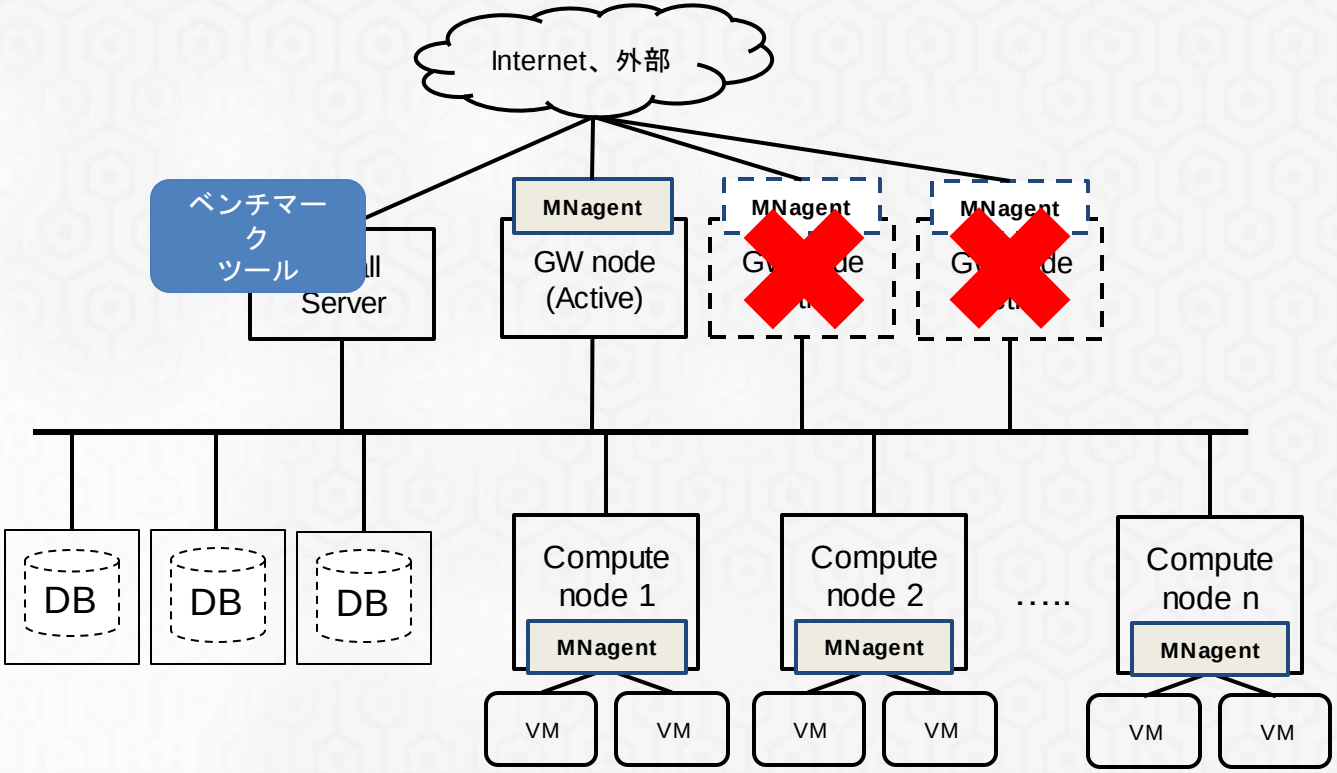
スケーラビリティの評価(MidoNet)

- 3本でトラフィック分散 ”2.6”Gbps
- アップリンク本数は必要に応じて”**自由に増速が可能**”



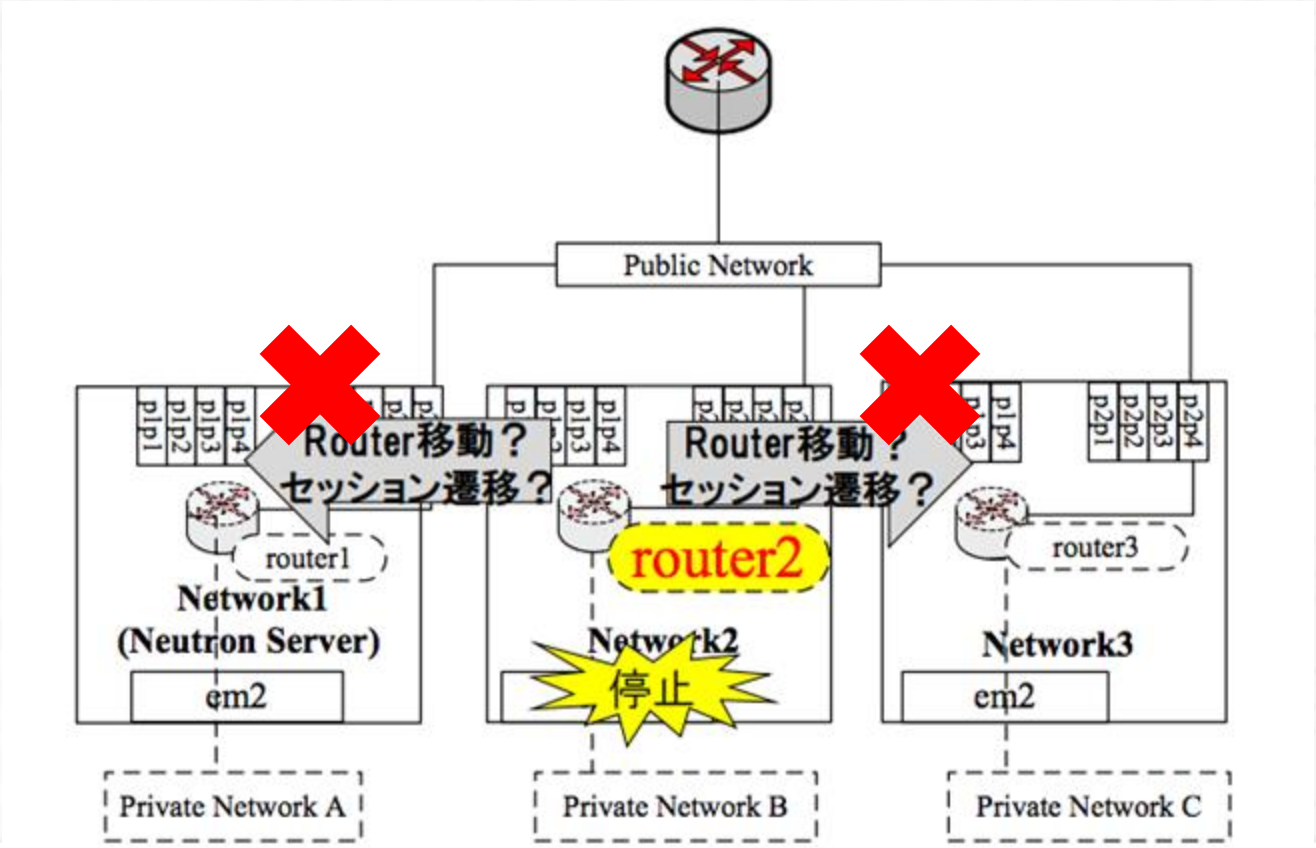
耐障害性の評価

- インターネット回線を順に落とした時のスループットの変化を確認
- インターネット回線を復旧させた時のスループットの変化を確認



耐障害性の評価 (OpenStack標準OpenvSwitch)

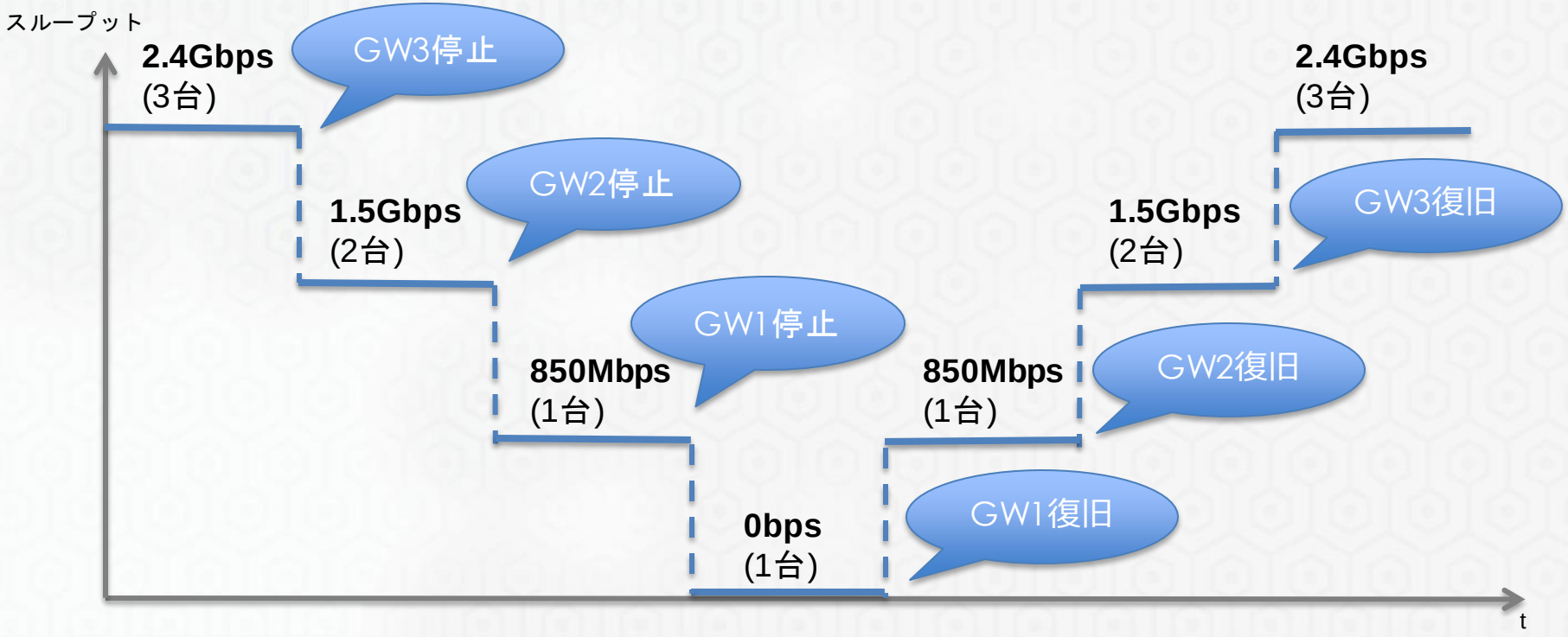
- 別々のルータに稼動しているルータは１つダウンしても自動切替されない



耐障害性の評価 (MidoNet)

- インターネット回線ダウン、**自動でフェイルオーバ**
- インターネット回線アップ、**自動で切り戻る**

帯域測定結果（下記は平均値）



話したい事はたくさんあります！しかしながら

MidoNet is now Open!

- http://www.midonet.org

皆さんの環境で自由に動かせます！

Step 1 : Download Midostack

Clone the midostack repository

```
git clone http://github.com/midonet/midostack
```

Step 2 : Run Midostack

```
cd ./midostack
```

Time to run it, this can take awhile, so grab a coffee and come back later.

```
./midonet_stack.sh
```

開発コミュニティに参加しよう！

ミドクラ トレーニングを提供開始！

2コースを提供中！

MKT101

OpenStackファンダメンタルズ

これからOpenStackを使ってクラウド環境を構築するエンジニアが、OpenStackの基礎的な考え方を理解し、OpenStackの各コンポーネントの機能や関連性といった全体構造について深く理解するための基礎コースです。

MKT102

OpenStackネットワーキングとMidoNet

OpenStackコンポーネントの中でも、最も理解が難しいと言われているネットワーク部分にフォーカスした中級者向けコースです。OpenStack Neutronの基礎、デフォルトのOVSプラグイン、Neutron Plug-inのミドクラのMidoNetの機能、設定方法、など、についてエクササイズを交えながら深く学ぶことができます。

詳しくはミドクラブースにお越しください！

- お問い合わせ先：training@midokura.com -

仮想ネットワークを作ってみたい方、お試し下さい！

まとめ

まとめ

- MidoNetはOpenStackネットワーキングに最適なネットワーク仮想化ソフトウェアです
- OpenStackとMidoNetで、柔軟なプライベートクラウドが、小さい規模から構築できます
- OpenStackおよびネットワーキングを基礎から学びたい方に、トレーニングを提供しています
- MidoNetはオープンソースになったので、自由に触ってみてください！

Mellanoxが提供する OpenStack最新ソリューション

株式会社アルティマ
北島佑樹
2015/2/27

本資料に含まれる測定データは一例であり、測定構成や条件によって変わることがあります。
また、本資料はMellanox Technologies社の公式見解を表すものではありません。
The results in this documents may differ for the configurations or/and conditions.
This documents does not reflect the official views of Mellanox Technologies.

アジェンダ

- Mellanox紹介
  - 会社概要
  - 製品紹介
- OpenStack最新ソリューション
  - ミドクラ x HWオフロード
- まとめ

Mellanox会社概要

NASDAQ® TELAVIV STOCK EXCHANGE Ticker: MLNX ALTIMA

- サーバ・ストレージエリア向け広帯域・低レイテンシーなインターコネクト市場のリーディングプロバイダー
  - InfiniBand：FDR(56Gbps)、EDR(100Gbps)..coming soon
  - Ethernet：10/40/56GbE
  - 共通ハードウェアでInfiniBand / Ethernetをサポート

- 本社・従業員数
  - ヨークニアム（イスラエル）、サニーベール（米国）
  - 全世界で1652人の従業員（2014年9月末時点）

- 安定した財務基盤
  - 2011年度の売上：　$259.3M
  - 2012年度の売上：　$500.8M
  - 2013年度の売上：　$390.9M
  - 2014年度の売上：　$463.6M

Mellanox社　製品ラインアップ

- サーバ・ストレージエリア向け広帯域・低レイテンシーのインターコネクト市場のリーディングプロバイダー

※資料中の図はメラノックス社提供資料です

Ethernetスイッチラインアップ

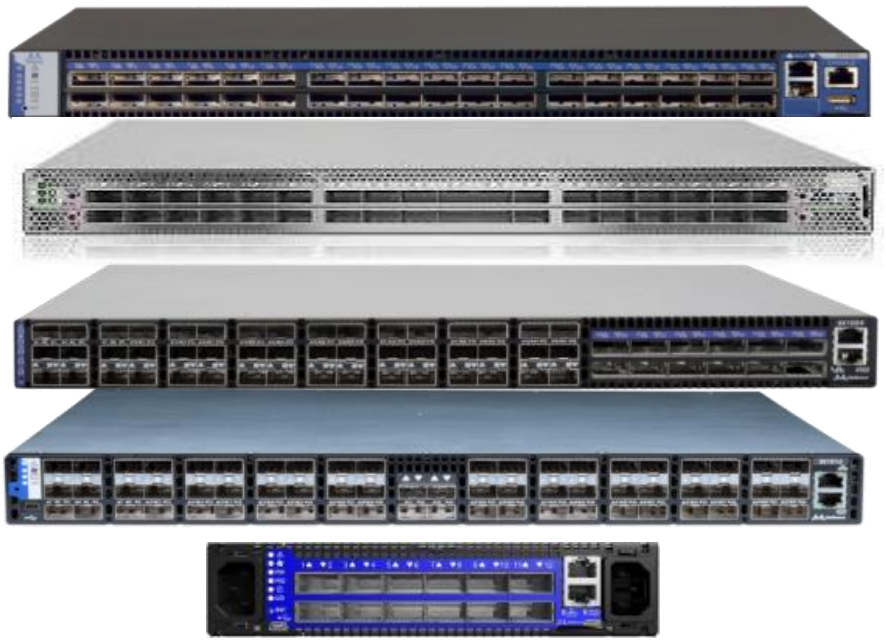
SX1036: 36 X 40/56G
The Ideal 40GbE ToR/Aggregation

SX1710: 36 X 40/56G
Intel Xeon Dual Core base system
NEW

SX1024: 48 X 10G + 12 X 40/56G
Non-blocking 10GbE➔ 40GbE ToR

SX1016: 64 X 10G
Highest density 10GbE ToR

SX1012: 12 X 40/56G(or Up to 48 X 10G)
Ideal storage/Database 10/40GbE Switch



- Highest Capacity in 1RU
  - 4TB Switching Capacity
  - Full non-blocking
- Unique Value Proposition
  - VPI: Ethernet and IB support
  - 10/40/56GbE
  - L2/L3 Feature set
- Latency
  - 220ns L2 latency
  - 330ns L3 latency
- Power (SX1036)
  - 83Watt (full 40GE rate)
  - 2.3W per 40GbE interface

ネットワークカード(NIC)ラインアップ

ALTIMA

ConnectX-3 PRO




型番	MCX311A-XCCT	MCX312B-XCCT	MCX313A-BCCT	MCX314A-BCCT
ポート	Single 10GbE	Dual 10GbE	Single /10/40/56GbE	Dual /10/40/56GbE
コネクタ	SFP+	SFP+	QSFP	QSFP
ケーブル	ダイレクトアタッチカッパー、光ファイバ			
ホストバス	PCIe 3.0			
特長	VXLAN/NVGRE オフロード, RDMA, SR-IOV, 各種オフロード(CheckSUM offload, TCP Segmentaion offload, Stateless offload)			
対応OS	RHEL, SLES, Microsoft Windows Sever, FreeBSD, Ubuntu, VMWare ESXi			

Mellanoxがもたらす サーバ/ストレージネットワークの高速化

ALTIMA

高い性能とオーバーレイオフロードエンジンやRDMAに
対応した業界トップクラスのNICカード

ConnectX-3 PRO

仮想ネットワークにおける
CPU負荷をNICにオフロード

vmware NSX

openstack Neutron

Microsoft

GWIND

PLUMgrid

midokura

ストレージネットワークにおける
データ通信をRDMAにより高速化

EMC²

FUJITSU

DataDirect NETWORKS

hp

Violin MEMORY

Microsoft SMB Direct

NIMBUS DATA

WD Western Digital

TOSHIBA

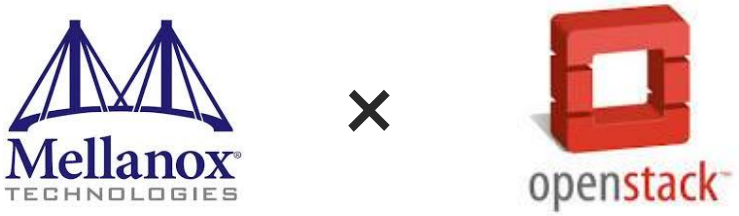
IBM

VIRIDENT

xiv tms

NetApp

MellanoxのOpenStack向けソリューションの紹介



Strong Partnership for OpenStack

ALTIMA

Mellanox TECHNOLOGIES × openstack

OpenStack Distribution

MIRANTIS

CANONICAL ubuntu

redhat LINUX

Network Virtualization

midokura

PLUMgrid

WIND

cumulus

Software Difined Storage

inktank ceph

RED HAT STORAGE redhat

オーバーレイネットワークのHWオフロード

- オーバーレイネットワーク
  - メリット
    - オーバーレイ(Tunneling Protocol)により、拡張性、柔軟性、リソースの稼働率向上を実現
  - 現在の課題
    - VXLAN/NVGREの処理に**CPUリソース**が使われるため、アプリケーション**性能の劣化**、**統合率低下**などのリスクが懸念される

- ConnectX-3 ProのVXLAN/NVGREオフロード
  - オーバーレイネットワーク用HWオフロードエンジンを搭載
  - VXLAN/NVGRE高速化
  - CPUのオーバーヘッドを劇的に削減

オーバーレイネットワークにおける
理想的なネットワーク基盤を実現可能

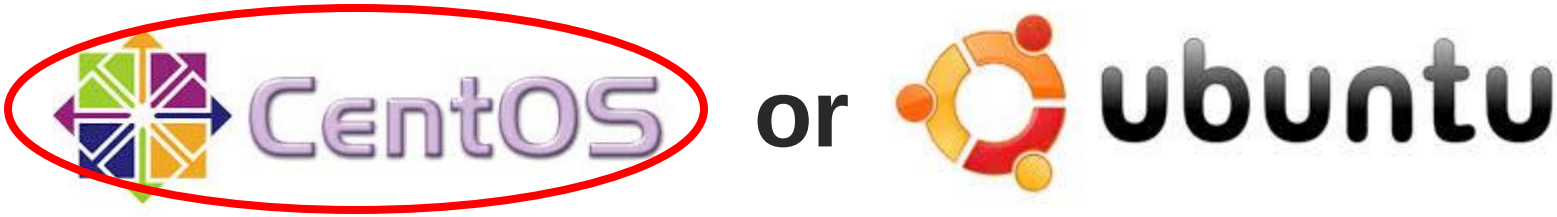
まとめ

- Mellanox社は10/40GE向け、スイッチシステム、NIC製品を提供
  - 業界トップクラスの性能とコスト競争力
- エッジオーバーレイのHWオフロード
  - CPU負荷の低減
  - 高い性能効率と収容率の実現
- 仮想化ネットワークソフトとの連携
  - オーバーレイ型SDNの導入にはHWオフロードは必須！！
  - midonetとのソリューションで柔軟で高速なOpenStack基盤を提供

    ※メラノックスブースでデモ説明対応中

実際どうなの!? VXLANオフロード
～VXLANオフロードの勘所～

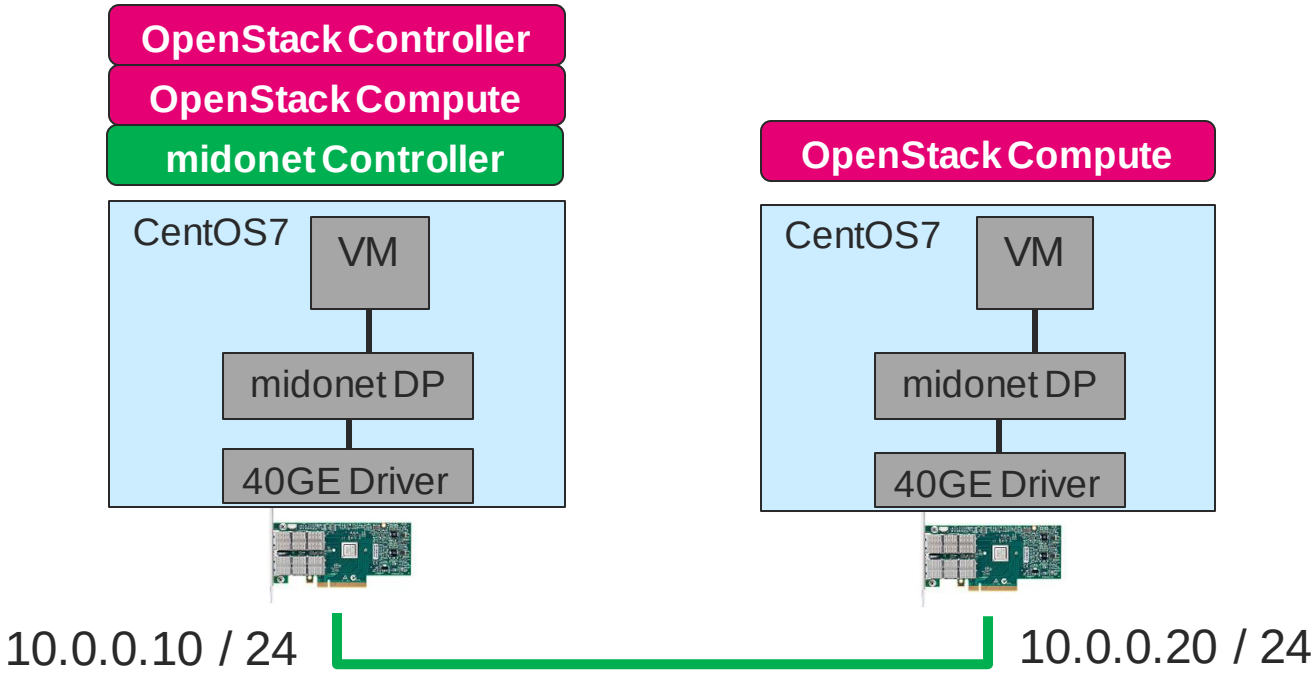
OS選定



- 理由
  - VXLANオフロードの観点ではどちらでもサポートだが、
  - サーバとの相性でUbuntu14.04は断念
  - CentOS7　→　3.10.0-123.el7

評価環境

- CPU : 1CPU(12core) ・・・ Xeon 2.40GHz (Fujitsu RX200S8)
- MEM : 18GB
- OS : CentOS7 (Kernel : 3.10.0-123.el7.x86_64)
- ConnectX-3 Pro (FW:2.33.5000)



VM info
OS:CentOS6.5
vCPU : 4core
vMEM : 4GB

Mellanoxドライバインストール

- RHEL7(CentOS7)ではin-boxでドライバサポート
- 今回は、Mellanoxから提供されるドライバを使用
  - Mellanox OFED v2.4-1
  - http://www.mellanox.com/page/products_dyn?product_family=26

```
# ethtool -i enp4s0
driver: mlx4_en
version: 2.4-1.0.0 (Jan 13 2015)
firmware-version: 2.33.5000
```

OpenStackインストレーション

ALTIMA

RDO

openstack

PackStack or スクラッチ

(or Devstack)

- 理由
  - 評価環境に自由度を持たせるためにはスクラッチから構築がベター
  - OpenStackの構成が理解できる
  - OpenStackコミュニティのマニュアル通りで動きます！
    - OpenStack Juno使用
- 苦労した点
  - Personal Issue

midonetインストレーション

ALTIMA

midostack or スクラッチ

- 理由
  - midostackは複数ノード構成を作れない
    - 最近、複数ノード構築ツールでoriduruができたようです。
    - https://github.com/midonet/orizuru
  - midonetを理解するため

midonetインストレーション

- midonetのお試しであれば"midostack"

Step 1 : Download Midostack

Clone the midostack repository

```
git clone http://github.com/midonet/midostack
```

Step 2 : Run Midostack

```
cd ./midostack
```

Time to run it, this can take awhile, so grab a coffee and come back later.

```
./midonet_stack.sh
```

http://www.midonet.org/#quickstart

midonetの勘所

- マニュアルがRHEL用のため、OpenStackマニュアルとの整合性がない

Documentation

MidoNet Documentation in Online (HTML) and Downloadable PDF format can be found below.

Quick Start Guide: Ubuntu 14.04 / Juno

Quick Start Guide: RHEL 7 / Juno (RDO)

Quick Start Guide: Ubuntu 14.04 / Icehouse

Quick Start Guide: RHEL 7 / Icehouse

Operation Guide

Reference Architecture

REST API

Developer Documentation (GitHub)

midonetの勘所

ALTIMA

- ネットワーク、インスタンス作成時、Security Groupエラーが発生

原因

- Neutronデフォルトプラグインが有効(L2,L3,OVS etc…)
- OpenStackマニュアルのコピペによるHuman Error…
- midonetプラグインを正確に入力しましょう

/etc/neutron/neutron.conf

```
[DEFAULT]
...
core_plugin = midonet.neutron.plugin.MidonetPluginV2
```

midonetの勘所

ALTIMA

- midonetサービスが立ち上がらない

  原因

  - NSDBノードのZookeeperがエラー
  - 1ノードで立ち上げようとしていたためNG
  - Zookeeperは3ノード構成がデフォルト

  ※今回はソースコードからインストールし1ノードで構成することで回避

ZooKeeper Installation

1. **Install ZooKeeper packages**

```
# yum install zookeeper zkdump
```

2. **Configure ZooKeeper**

   a. **Common Configuration**

   Edit the `/etc/zookeeper/zoo.cfg` file to contain the following:

```
server.1=controller:2888:3888
server.2=network:2888:3888
server.3=compute1:2888:3888
```

VXLAN設定 (Mellanox)

- デフォルトでオフロードはEnable

```
# ethtool -k enp130s0 | grep udp
tx-udp_tnl-segmentation: on
```

- オフロードEnable / Disableの変更方法

```
# ethtool -K enp130s0 tx-udp_tnl-segmentation off
```

VXLAN設定 (midonet)

- midonetのデータパス "tnvxlan-overlay"を使います

```
# mm-dpctl  --show-dp midonet
Datapath name    : midonet
Datapath index : 6
Datapath Stats:
  Flows :0
  Hits  :1147
  Lost  :0
  Misses:211
Port #0 "midonet"   Internal
Port #1 "tngre-overlay"   Gre
Port #2 "tnvxlan-overlay"   VXLan
Port #3 "tnvxlan-vtep"   VXLan
Port #4 "tap4f973dca-6d"   NetDev
Port #5 "tap5052daea-da"   NetDev
```

VXLAN設定 (midonet)

ALTIMA

- VM間通信でVXLANのトンネルを張る設定が必要

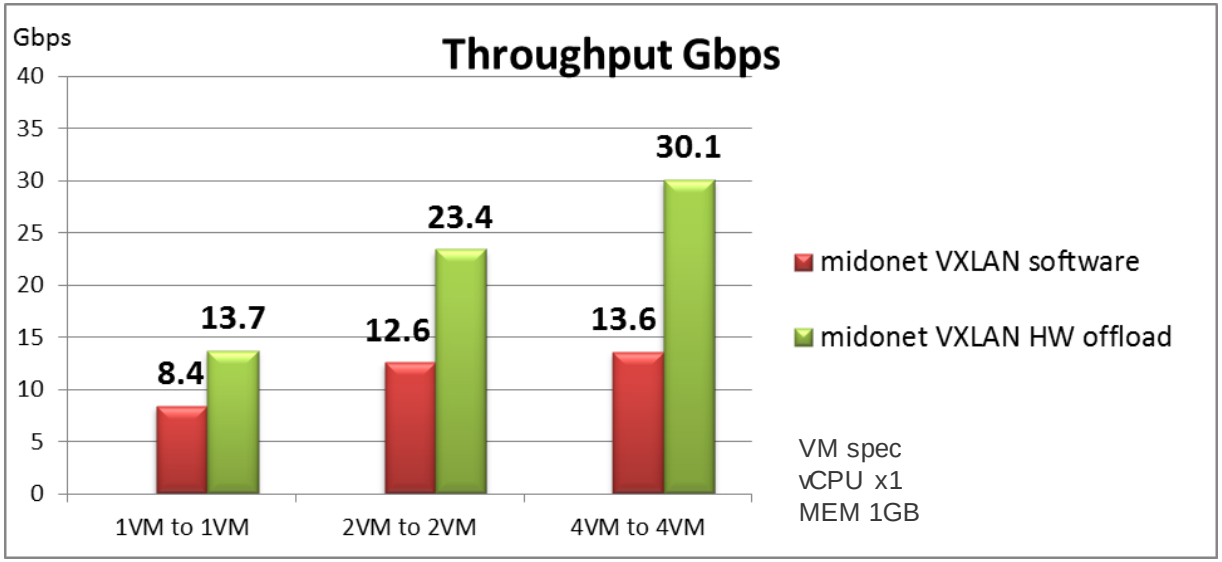
```
midonet> create tunnel-zone  name vxlan-tz type vxlan
tzone0

midonet> list tunnel-zone
tzone tzone0 name vxlan-tz type vxlan

midonet> tunnel-zone  tzone0 add member  host host0 address 192.168.100.201
zone tzone1 host host0 address 192.168.100.201
midonet> tunnel-zone  tzone0 add member  host host1 address 192.168.100.202
zone tzone1 host host1 address 192.168.100.202

midonet> tunnel-zone  tzone0 list member
zone tzone1 host host0 address 192.168.100.201
zone tzone1 host host1 address 192.168.100.202
```

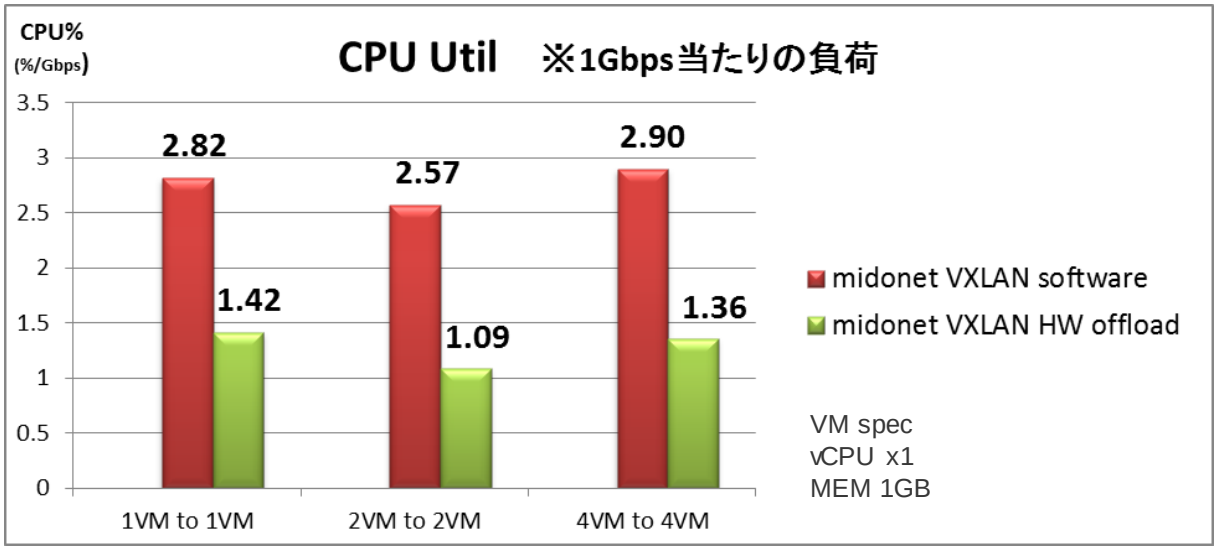
VXLANネットワーク疎通

- OpenStackダッシュボードの操作で動きます
  - 「ネットワーク作成」→「インスタンス作成」→「ネットワークアサイン」

VXLAN HWオフロードの性能データ

本資料に含まれる測定データは一例であり、条件によって変わることがあります。
また、本資料はMellanox Technologies社の公式見解を表すものではありません。





HWオフロード
スループット性能

✓ **2.2倍**

HWオフロード
CPU負荷率

✓ **52%削減**

テスト環境

OS	: **CentOS7 (3.10.0-123)**
OpenStack	: **Juno**
midonet	: **midolman-2015.01-0.1.rc0**
Driver	: **Mellanox OFED ver2.4.1**
Bench Tool	: **iPerf v2.0.5**

続いて

ミドクラ鈴木さんにOpenStack Ubuntu環境の
テストサマリをご紹介いただきます

MidoNet
+
Mellanox VXLAN offload

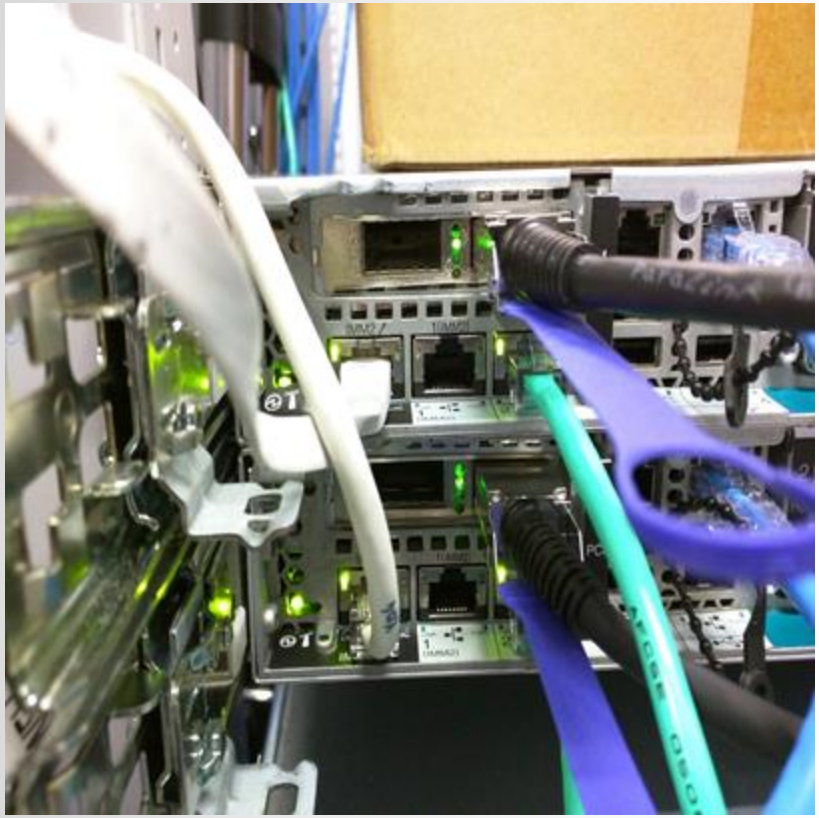
MidoNet VXLAN

2014/06~ 

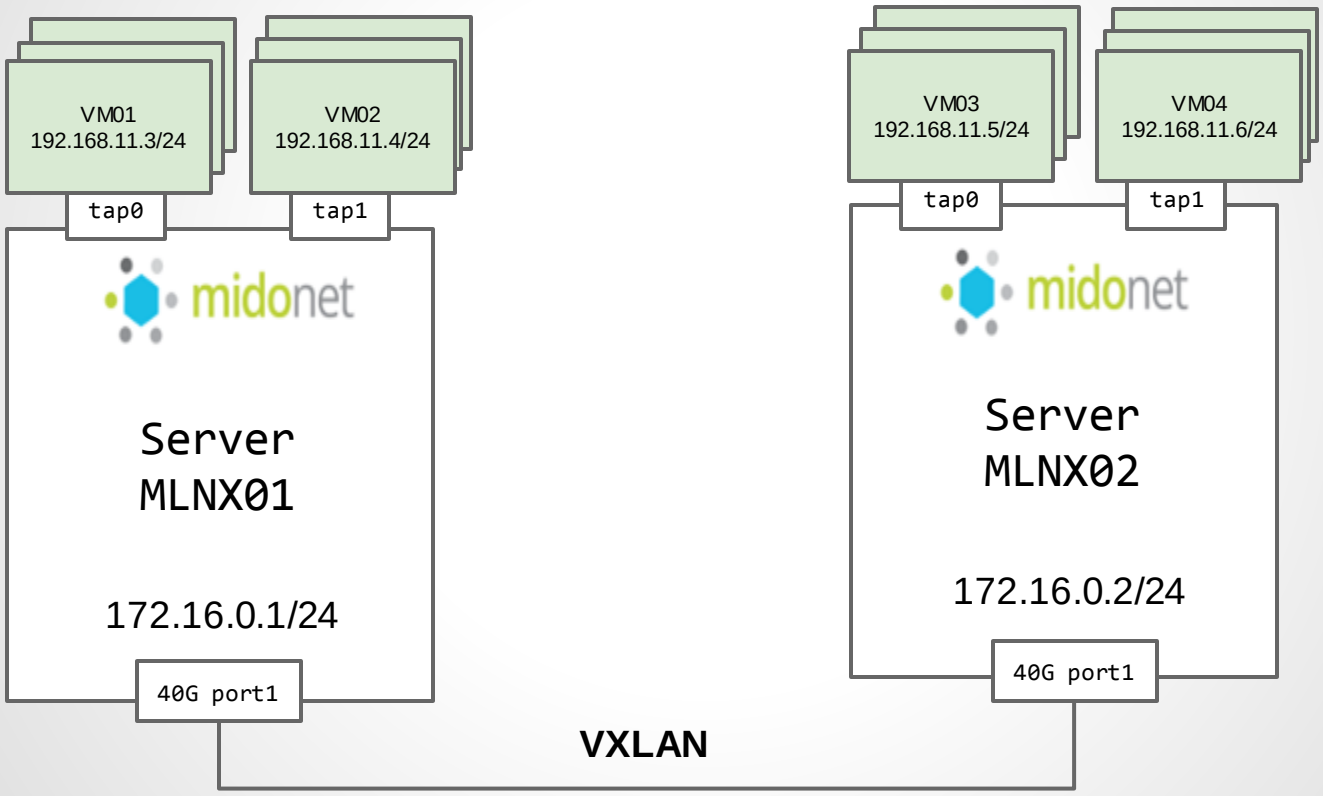
- 開発者によるVXLANコード追加
- VXLAN通信テスト
- VXLANオフロード機能テスト

*テストの為に物理環境を用意しました。

テスト環境



テスト環境


VM01
192.168.11.3/24
VM02
192.168.11.4/24
tap0
tap1
midonet
Server
MLNX01
172.16.0.1/24
40G port1
VM03
192.168.11.5/24
VM04
192.168.11.6/24
tap0
tap1
midonet
Server
MLNX02
172.16.0.2/24
40G port1
VXLAN

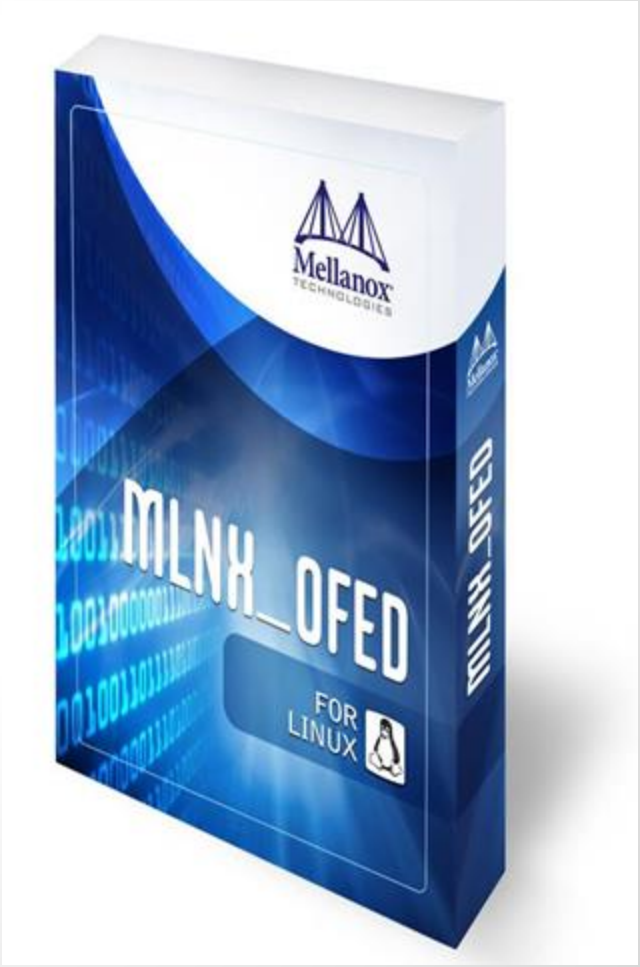
OSの選択

- RHEL/CentOS6.5 plus 3.14 kernel
- Ubuntu 12.04 plus 3.10 kernel

- この時はCentOSを選択

Mellanox 40G NIC driverの選択

- Kernel 3.14 inbox driver
- Mellanox OFED Driver
  *WEBから入手可能.



40Gポートが表示されない.

```
root@mlnx01:~# ifconfig -a
em1       Link encap:Ethernet  HWaddr 40:f2:e9:0b:41:3c
          BROADCAST MULTICAST  MTU:1500  Metric:1
          RX packets:0 errors:0 dropped:0 overruns:0 frame:0
          TX packets:0 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:0 (0.0 B)  TX bytes:0 (0.0 B)
          Memory:a8580000-a85a0000

em2       Link encap:Ethernet  HWaddr 40:f2:e9:0b:41:3d
          inet addr:192.168.100.193  Bcast:192.168.100.255  Mask:255.255.255.0
          inet6 addr: fe80::42f2:e9ff:fe0b:413d/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:8888  Metric:1
          RX packets:1198107 errors:0 dropped:0 overruns:0 frame:0
          TX packets:1387662 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:175380048 (175.3 MB)  TX bytes:173903881 (173.9 MB)
          Memory:a85a0000-a85c0000
```

*しかしながら、ちゃんとPCIデバイスは認識されている.

```
root@mlnx01:~# lspci | grep Mellanox
11:00.0 Network controller: Mellanox Technologies MT27520 Family [ConnectX-3 Pro]
```

40Gポートが表示されない.

原因: Infinibandモードだった.

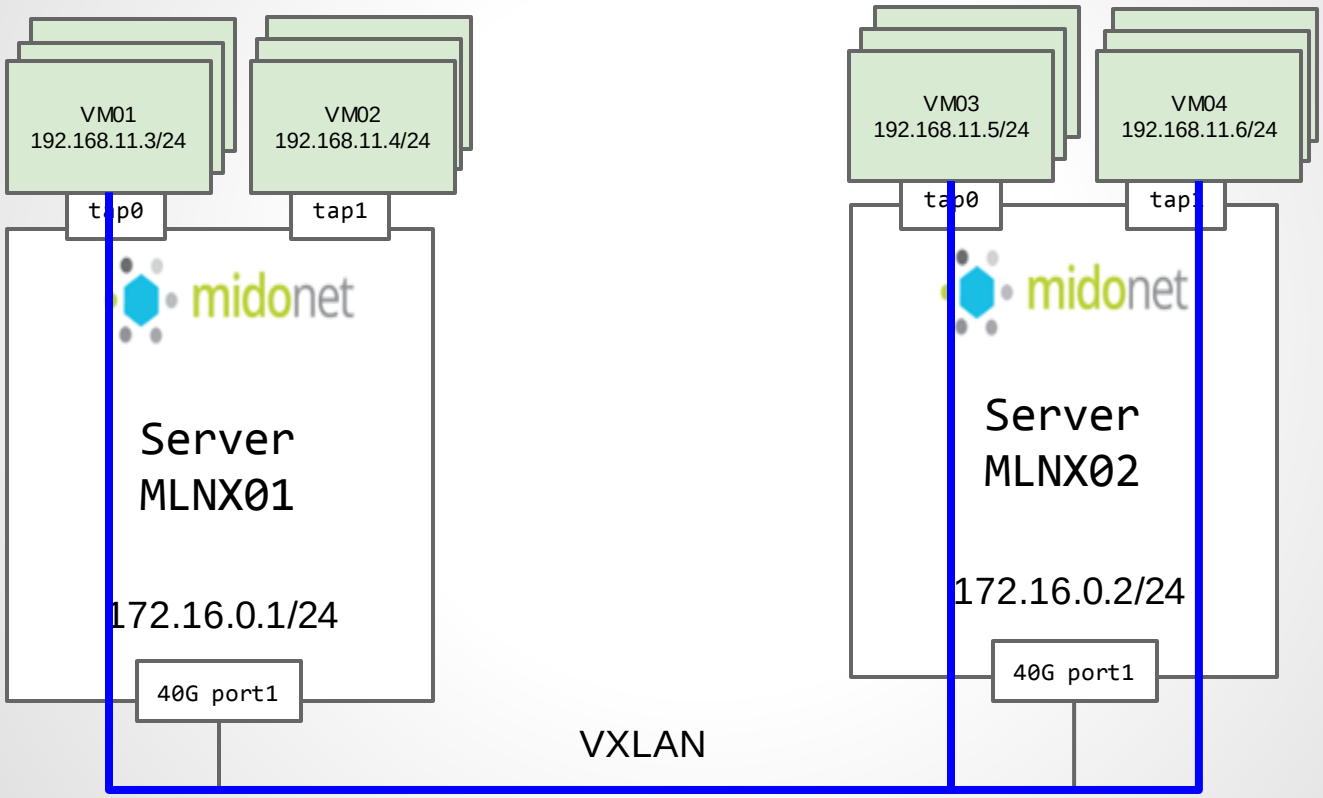
```
root@mlnx01:~# cat /sys/bus/pci/devices/0000:11:00.0/mlx4_port1
auto (eth)
root@mlnx01:~# cat /sys/bus/pci/devices/0000:11:00.0/mlx4_port2
auto (ib)
```

40Gポートが利用可能に

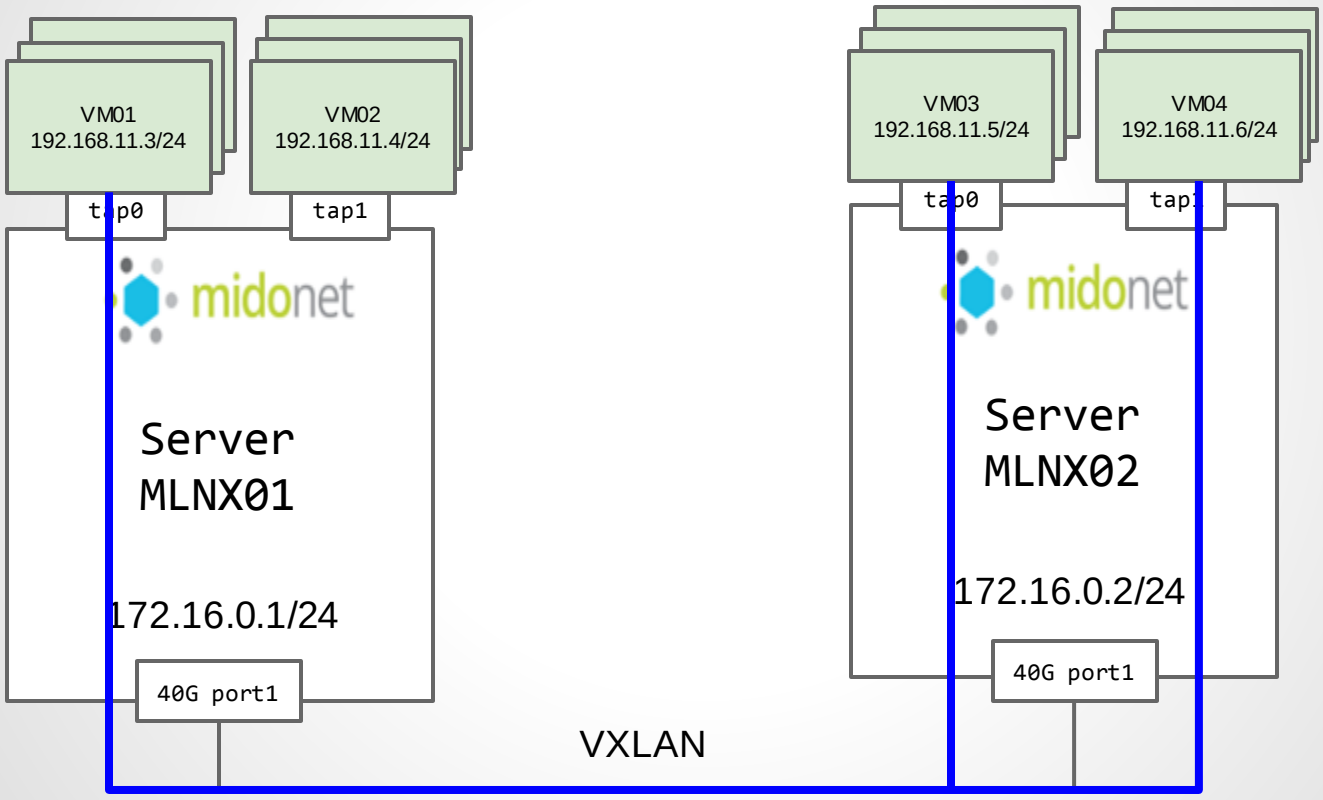
```
root@mlnx01:~# echo eth > /sys/bus/pci/devices/0000:11:00.0/mlx4_port1
root@mlnx01:~# echo eth > /sys/bus/pci/devices/0000:11:00.0/mlx4_port2
```

```
root@mlnx01:~# ethtool p1p1
Settings for p1p1:
        Supported ports: [ FIBRE ]
        Supported link modes:   1000baseKX/Full
                                10000baseKX4/Full
                                10000baseKR/Full
                                40000baseCR4/Full
                                40000baseSR4/Full
        Supported pause frame use: Symmetric Receive-only
        Supports auto-negotiation: Yes
        Advertised link modes:  1000baseKX/Full
                                10000baseKX4/Full
                                10000baseKR/Full
                                40000baseCR4/Full
                                40000baseSR4/Full
        Advertised pause frame use: Symmetric
        Advertised auto-negotiation: Yes
        Speed: 40000Mb/s
        Duplex: Full
        Port: Direct Attach Copper
        PHYAD: 0
        Transceiver: internal
        Auto-negotiation: off
        Supports Wake-on: d
        Wake-on: d
        Current message level: 0x00000014 (20)
                               link ifdown
        Link detected: yes
```

MidoNet VXLAN通信の確認



MidoNet VXLAN通信の確認



ping/ssh 192.168.11.5 OK
ping/ssh 192.168.11.6 OK

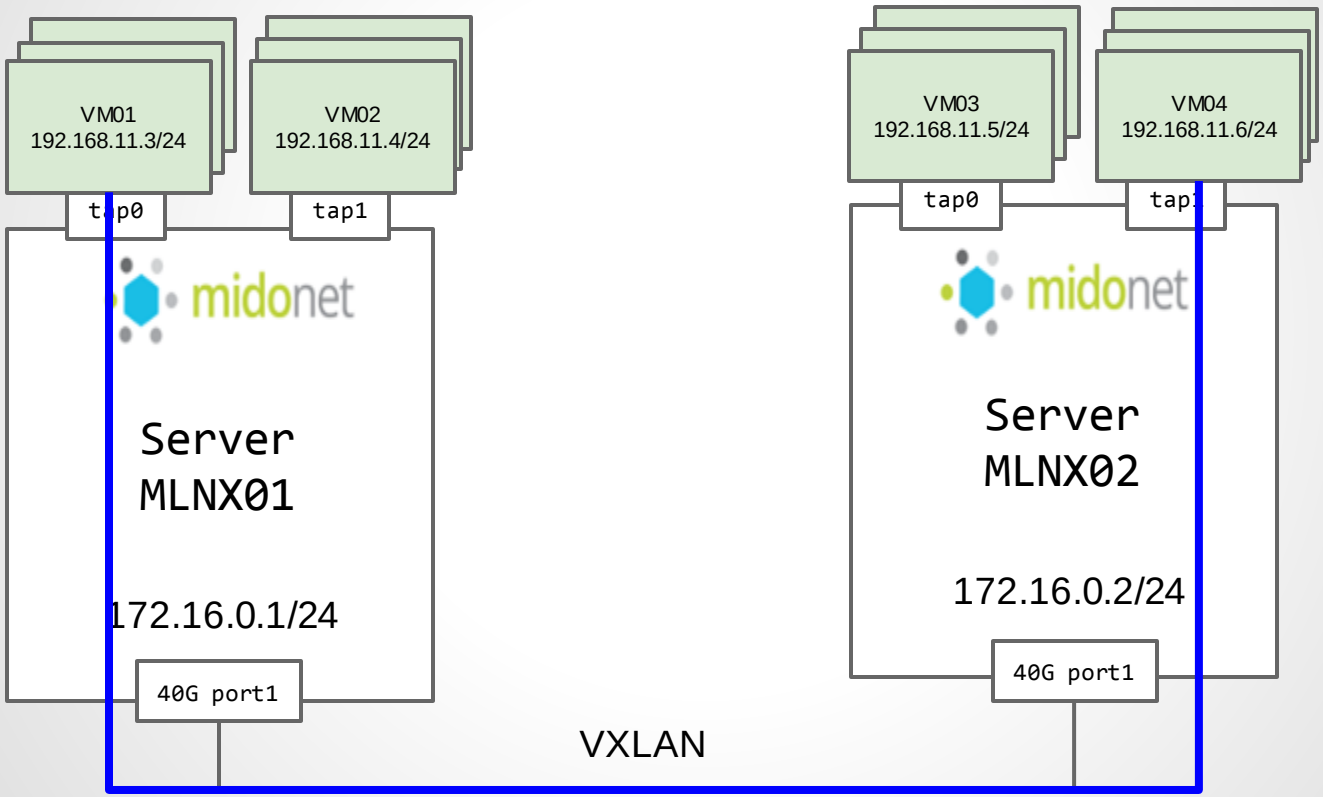
VXLANオフロードの有効化

```
root@mlnx01:~# cat /etc/modprobe.d/mlx4.conf
# mlx4_core gets automatically loaded, load mlx4_en also (LP: #1115710)
options mlx4_core log_num_mgm_entry_size=-1 debug_level=1
```

```
root@mlnx01:~# cat /etc/modprobe.d/mlnx.conf
# Module parameters for MLNX_OFED kernel modules
blacklist mlx4_core
blacklist mlx4_en
blacklist mlx5_core
blacklist mlx5_ib
```

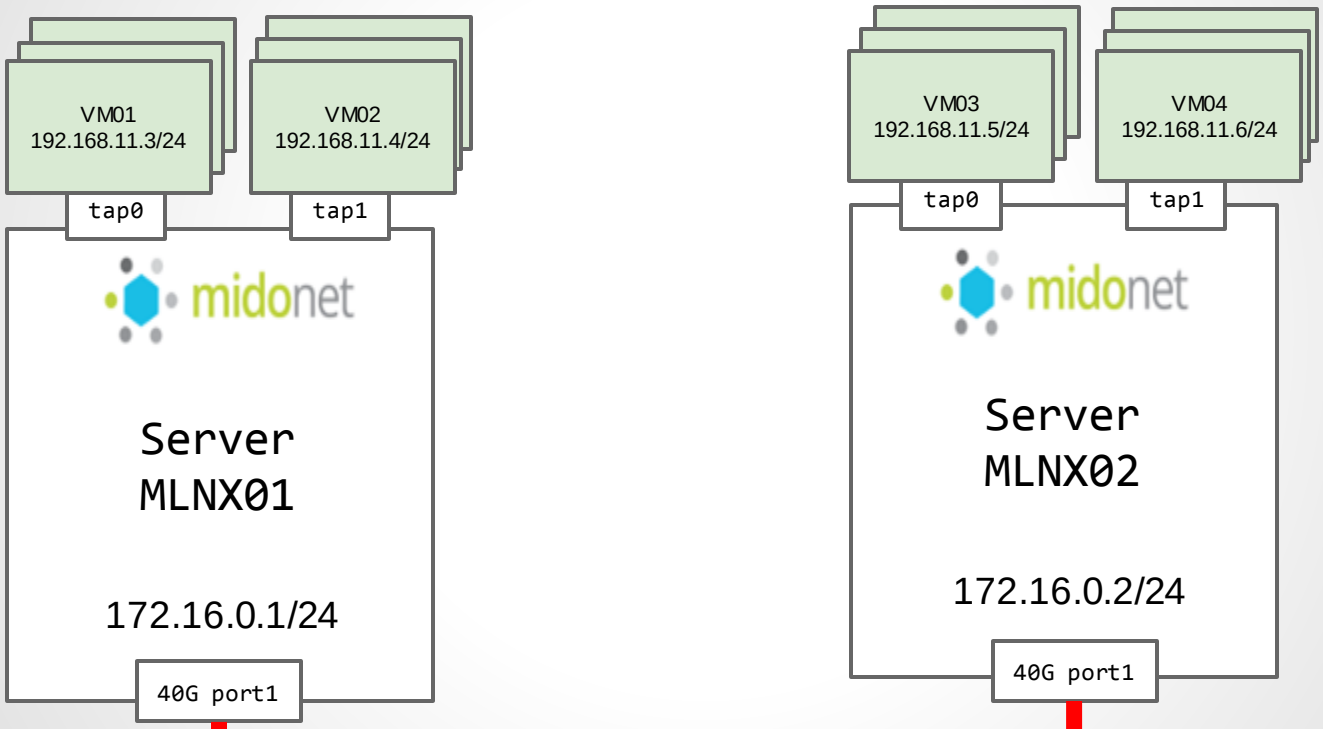
```
[ 3475.361329] mlx4_core 0000:11:00.0:     Device manage flow steering support
[ 3475.361345] mlx4_core 0000:11:00.0:     Device managed flow steering IPoIB support
[ 3475.361369] mlx4_core 0000:11:00.0:     TCP/IP offloads/flow-steering for VXLAN support
[ 3475.361378] mlx4_core 0000:11:00.0:     Device madaged flow steering VLAN tag mode support
[ 3476.641412] mlx4_core 0000:11:00.0: Steering mode is: Device managed flow steering, oper_log_mgm_entry_size = 1
```

VM間通信がとてつもなく遅い



iperf: 1~3”M”bps

そもそもサーバ間での速度はどうか



iperf: 21Gbps

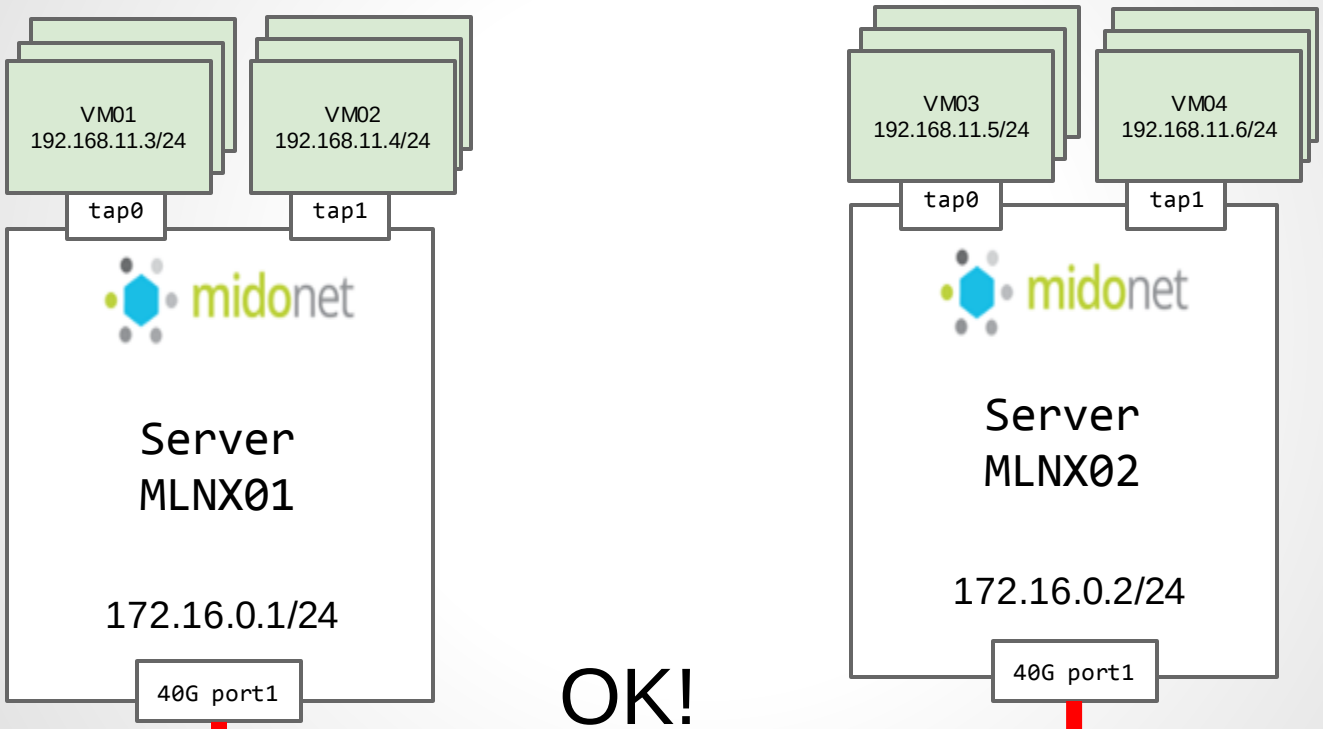
PCIe 3.0 x16速度の確認

```
11:00.0 Network controller: Mellanox Technologies MT27520 Family [ConnectX-3 Pro]
        Subsystem: Mellanox Technologies Device 0003
        Control: I/O- Mem+ BusMaster+ SpecCycle- MemWINV- VGASnoop- ParErr+ Stepping- SERR- FastB2B- DisINTx+
        Status: Cap+ 66MHz- UDF- FastB2B- ParErr- DEVSEL=fast >TAbort- <TAbort- <MAbort- >SERR- <PERR- INTx-
        Latency: 0, Cache Line Size: 64 bytes
        Interrupt: pin A routed to IRQ 32
        Region 0: Memory at a8300000 (64-bit, non-prefetchable) [size=1M]
        Region 2: Memory at aa800000 (64-bit, prefetchable) [size=8M]
        Expansion ROM at <ignored> [disabled]
```

```
LnkSta: Speed 8GT/s, Width x8, TrErr- Train- SlotClk+ DLActive- BWMgmt- ABWMgmt-
```

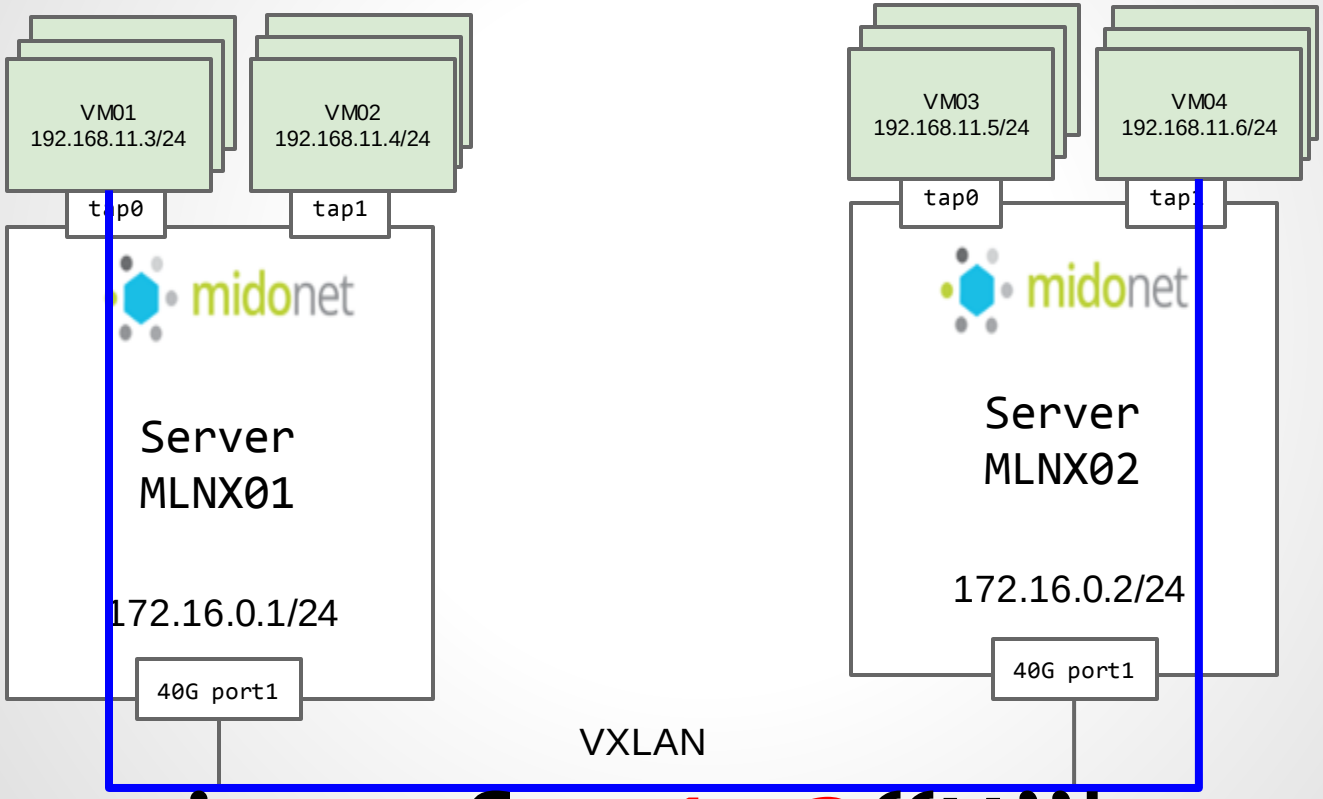
* CPUの制限で5GT/sだったのでCPUを交換.

そもそもサーバ間での速度はどうか



iperf: 37Gbps

やっぱりVM間通信がとてつもなく遅い

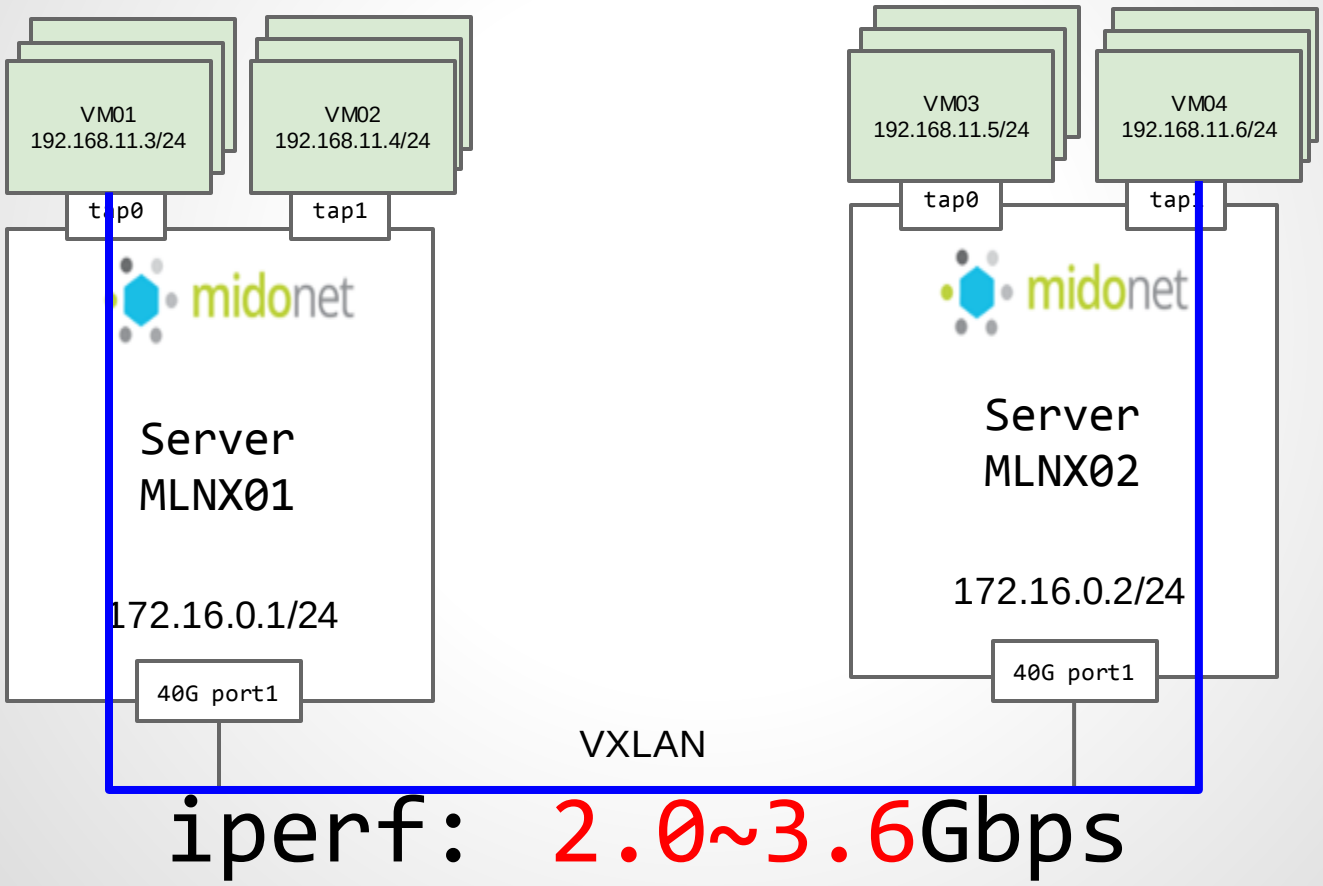


iperf: 1~3“M”bps

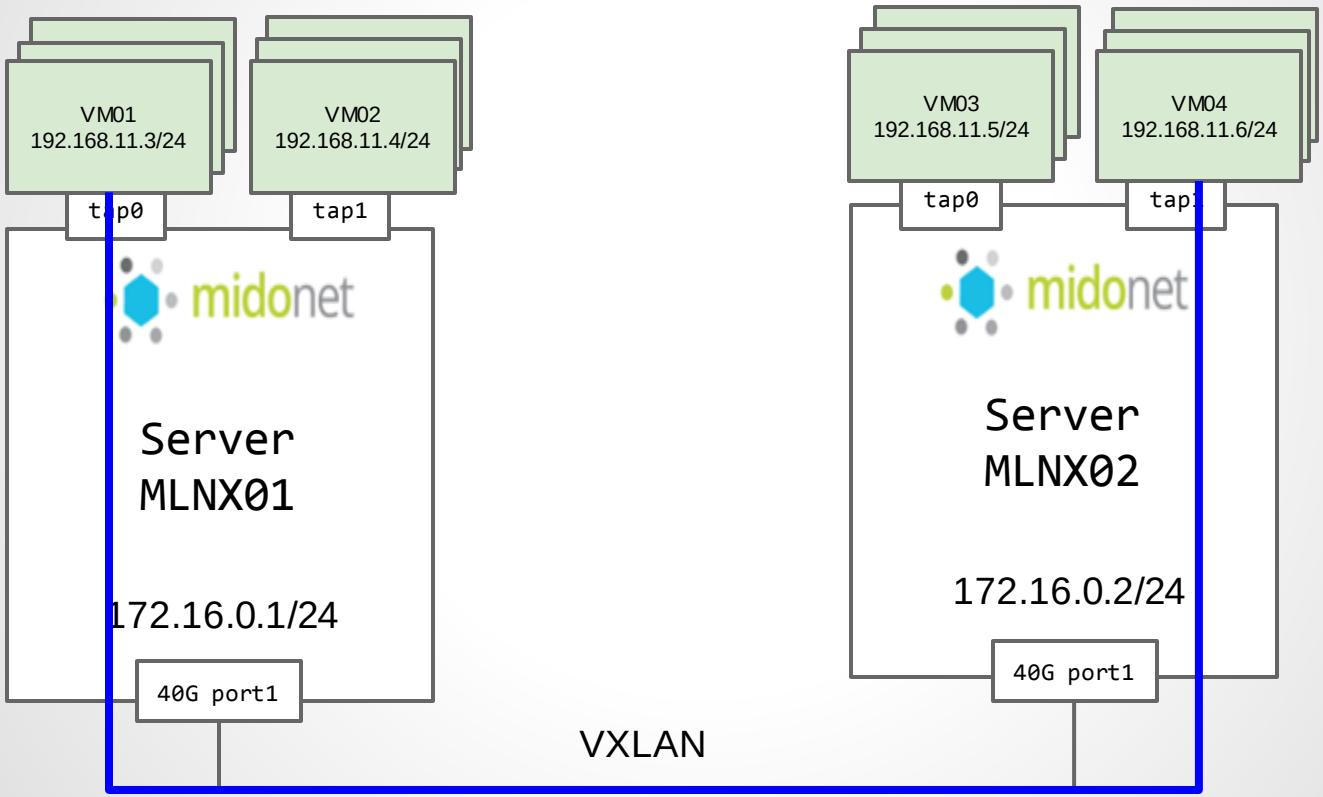
MidoNet開発者によるコード変更

```
commit c9d544701189d54510edf2f6662f8ab30db00ee7
Merge: b42487d 5e7580a
Author: Hugo Benichi <hugo@midokura.com>
Date:   Tue Jun 17 16:47:17 2014 +0200

    Merge topics 'vxlan_tunnelling', 'vxlan'

    * changes:
      Fixing HBT leak when failing to create a DpPort
      Per tunnel route tunnelling output action
```

通信速度改善された（しかし？）



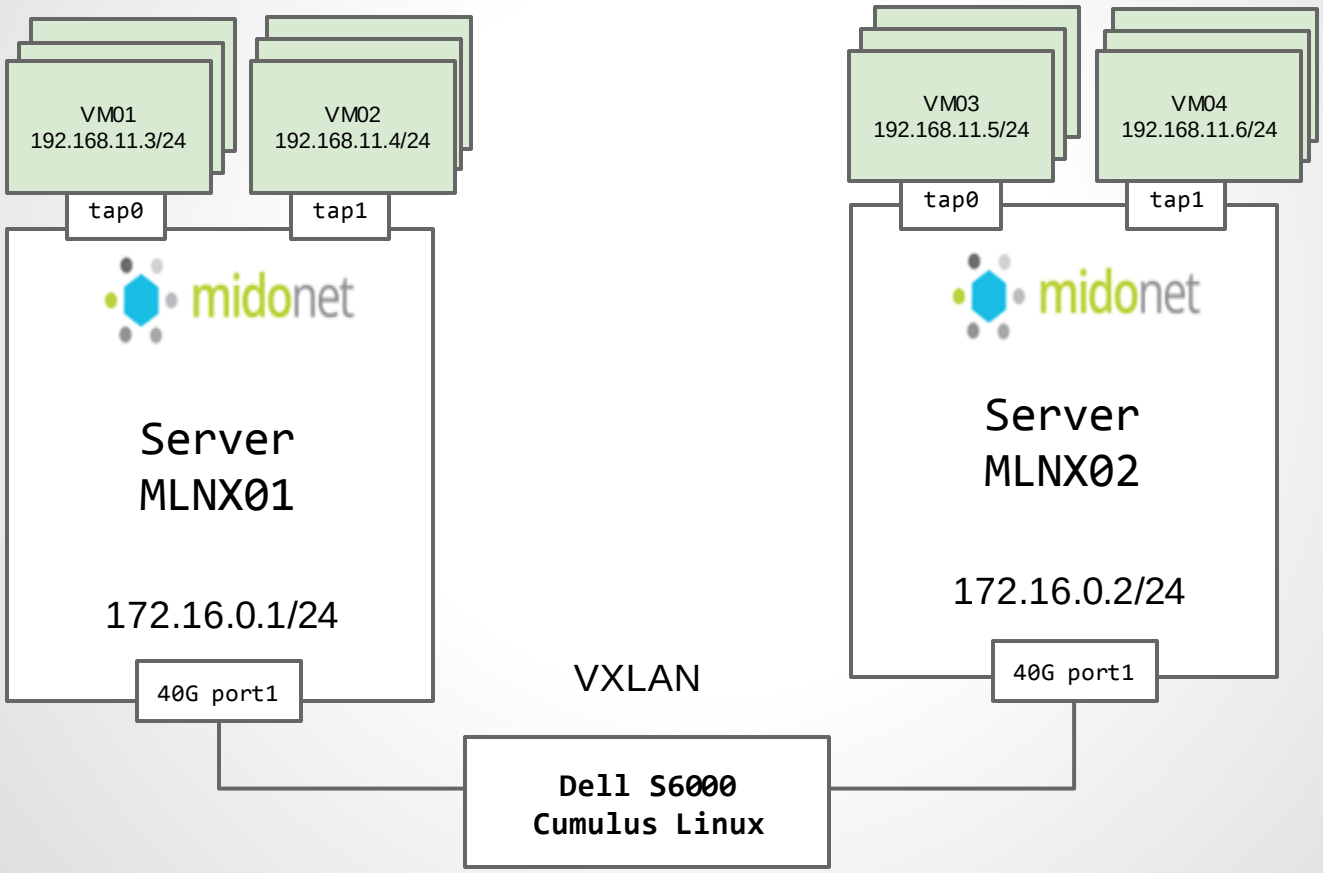
北島さんのテスト環境



iperf: 9~11Gbps

Ubuntu 14.04で試してみることに

テスト環境 (Ubuntu14.04)



VM間通信 (Ubuntu14.04環境)

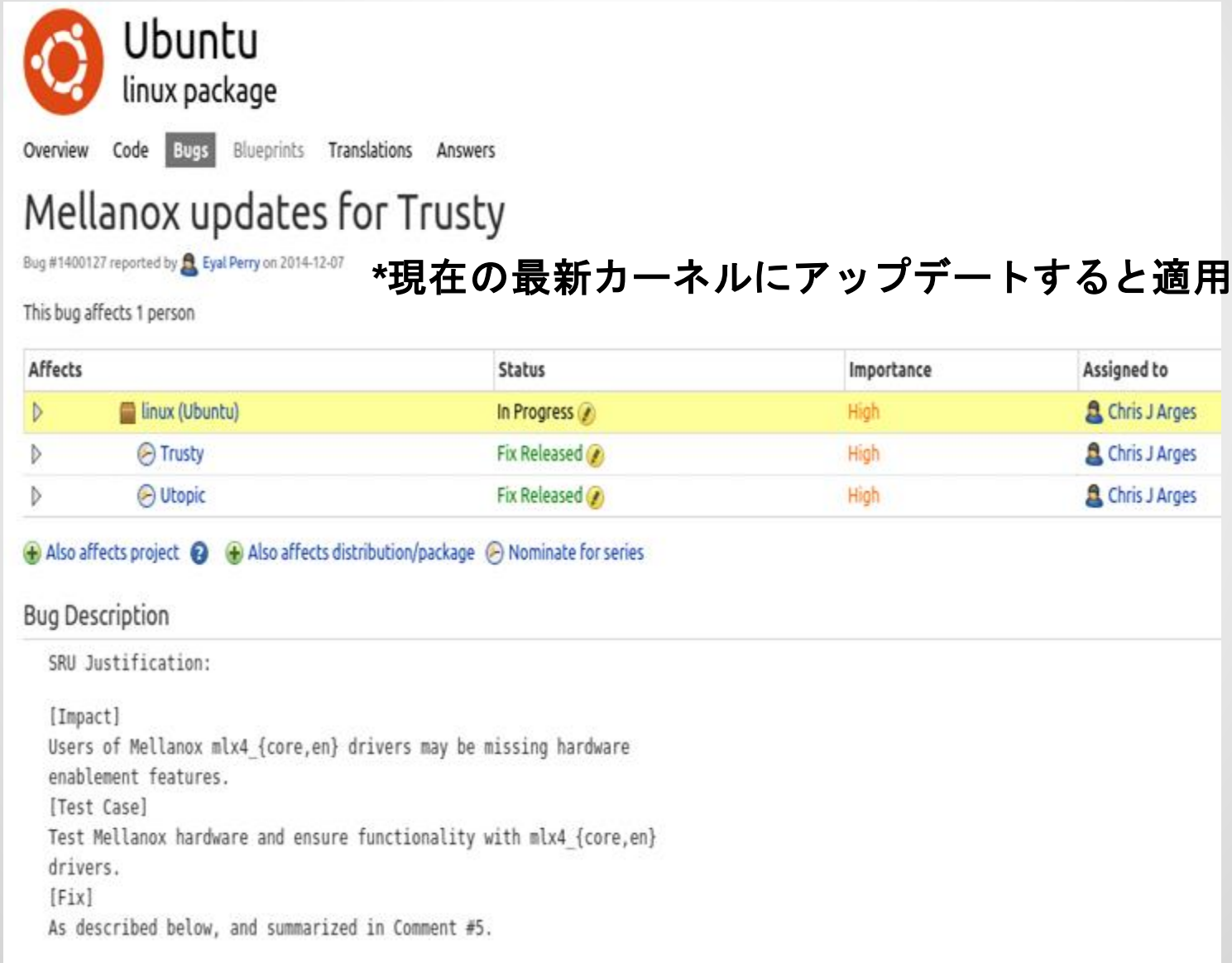
[協力ありがとうございます]
アルティマ北島さん、Canonical 松本さん

前回同様　**2.4 ~ 3.0**Gbps

NICドライバ設定の確認

CentOS	Ubuntu
Features for enp130s0:	Features for p1p1:
rx-checksumming: on	rx-checksumming: on
tx-checksumming: on	tx-checksumming: on
tx-checksum-ipv4: on	tx-checksum-ipv4: on
tx-checksum-ip-generic: off [fixed]	tx-checksum-ip-generic: off [fixed]
tx-checksum-ipv6: on	tx-checksum-ipv6: on
tx-checksum-fcoe-crc: off [fixed]	tx-checksum-fcoe-crc: off [fixed]
tx-checksum-sctp: off [fixed]	tx-checksum-sctp: off [fixed]
scatter-gather: on	scatter-gather: on
tx-scatter-gather: on	tx-scatter-gather: on
tx-scatter-gather-fraglist: off [fixed]	tx-scatter-gather-fraglist: off [fixed]
tcp-segmentation-offload: on	tcp-segmentation-offload: on
tx-tcp-segmentation: on	tx-tcp-segmentation: on
tx-tcp-ecn-segmentation: off [fixed]	tx-tcp-ecn-segmentation: off [fixed]
tx-tcp6-segmentation: on	tx-tcp6-segmentation: on
udp-fragmentation-offload: off [fixed]	udp-fragmentation-offload: off [fixed]
generic-segmentation-offload: on	generic-segmentation-offload: on
generic-receive-offload: on	generic-receive-offload: on
large-receive-offload: off	large-receive-offload: off
rx-vlan-offload: on	rx-vlan-offload: on
tx-vlan-offload: on [fixed]	tx-vlan-offload: on
ntuple-filters: off	ntuple-filters: off
receive-hashing: on	receive-hashing: on
highdma: on [fixed]	highdma: on [fixed]
rx-vlan-filter: on [fixed]	rx-vlan-filter: on [fixed]
vlan-challenged: off [fixed]	vlan-challenged: off [fixed]
tx-lockless: off [fixed]	tx-lockless: off [fixed]
netns-local: off [fixed]	netns-local: off [fixed]
tx-gso-robust: off [fixed]	tx-gso-robust: off [fixed]
tx-fcoe-segmentation: off [fixed]	tx-fcoe-segmentation: off [fixed]
tx-gre-segmentation: off [fixed]	tx-gre-segmentation: off [fixed]
tx-ipip-segmentation: off [fixed]	tx-ipip-segmentation: off [fixed]
tx-sit-segmentation: off [fixed]	tx-sit-segmentation: off [fixed]
tx-udp_tnl-segmentation: on	tx-udp_tnl-segmentation: on
tx-mpls-segmentation: off [fixed]	tx-mpls-segmentation: off [fixed]
fcoe-mtu: off [fixed]	fcoe-mtu: off [fixed]
tx-nocache-copy: on	tx-nocache-copy: on
loopback: off	loopback: off
rx-fcs: off [fixed]	rx-fcs: off [fixed]
rx-all: off [fixed]	rx-all: off [fixed]
tx-vlan-stag-hw-insert: off [fixed]	tx-vlan-stag-hw-insert: off [fixed]
rx-vlan-stag-hw-parse: off [fixed]	rx-vlan-stag-hw-parse: off [fixed]
rx-vlan-stag-filter: off [fixed]	rx-vlan-stag-filter: off [fixed]
	l2-fwd-offload: off [fixed]

Mellanoxエンジニアによるアップデート



Ubuntu
linux package

Overview Code **Bugs** Blueprints Translations Answers

Mellanox updates for Trusty

Bug #1400127 reported by Eyal Perry on 2014-12-07

This bug affects 1 person

Affects	Status	Importance	Assigned to
linux (Ubuntu)	In Progress	High	Chris J Arges
Trusty	Fix Released	High	Chris J Arges
Utopic	Fix Released	High	Chris J Arges

Also affects project | Also affects distribution/package | Nominate for series

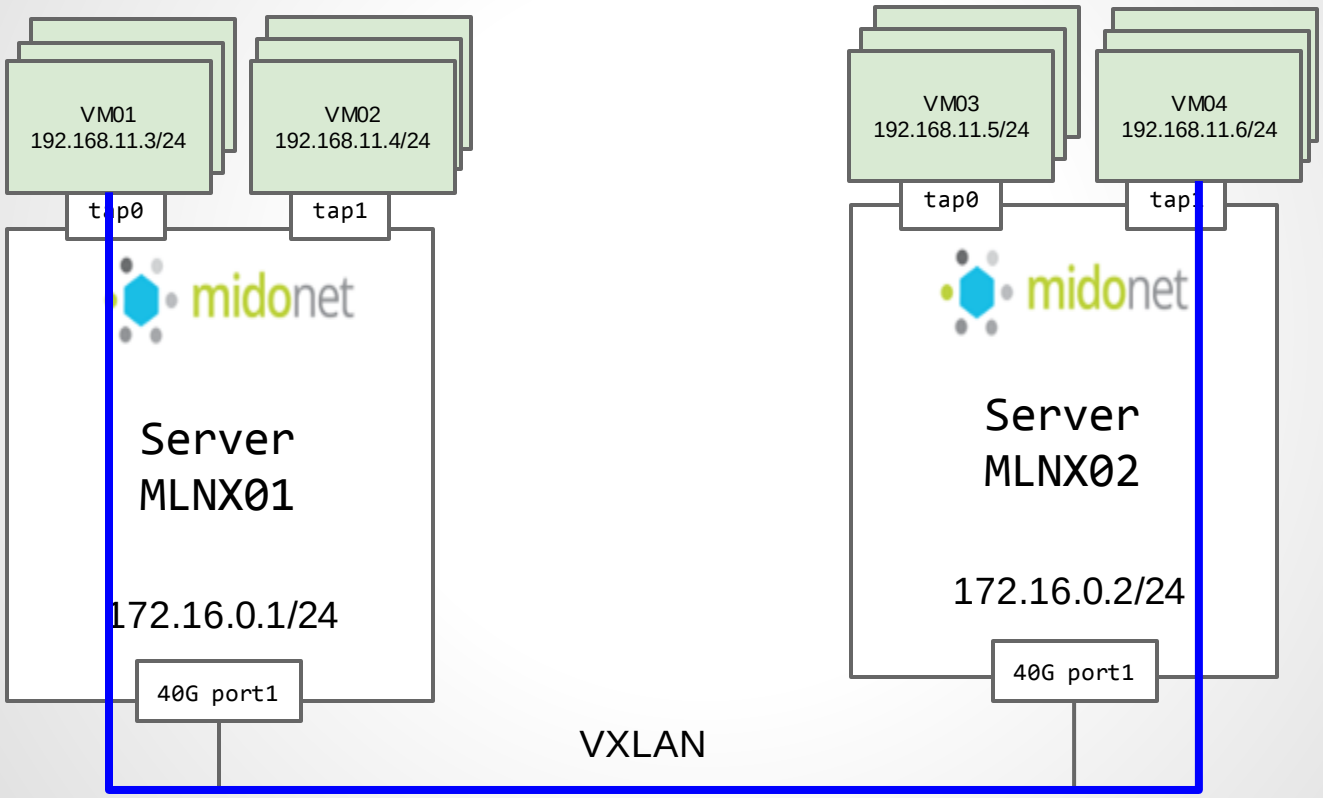
Bug Description

SRU Justification:

[Impact]
Users of Mellanox mlx4_{core,en} drivers may be missing hardware
enablement features.
[Test Case]
Test Mellanox hardware and ensure functionality with mlx4_{core,en}
drivers.
[Fix]
As described below, and summarized in Comment #5.

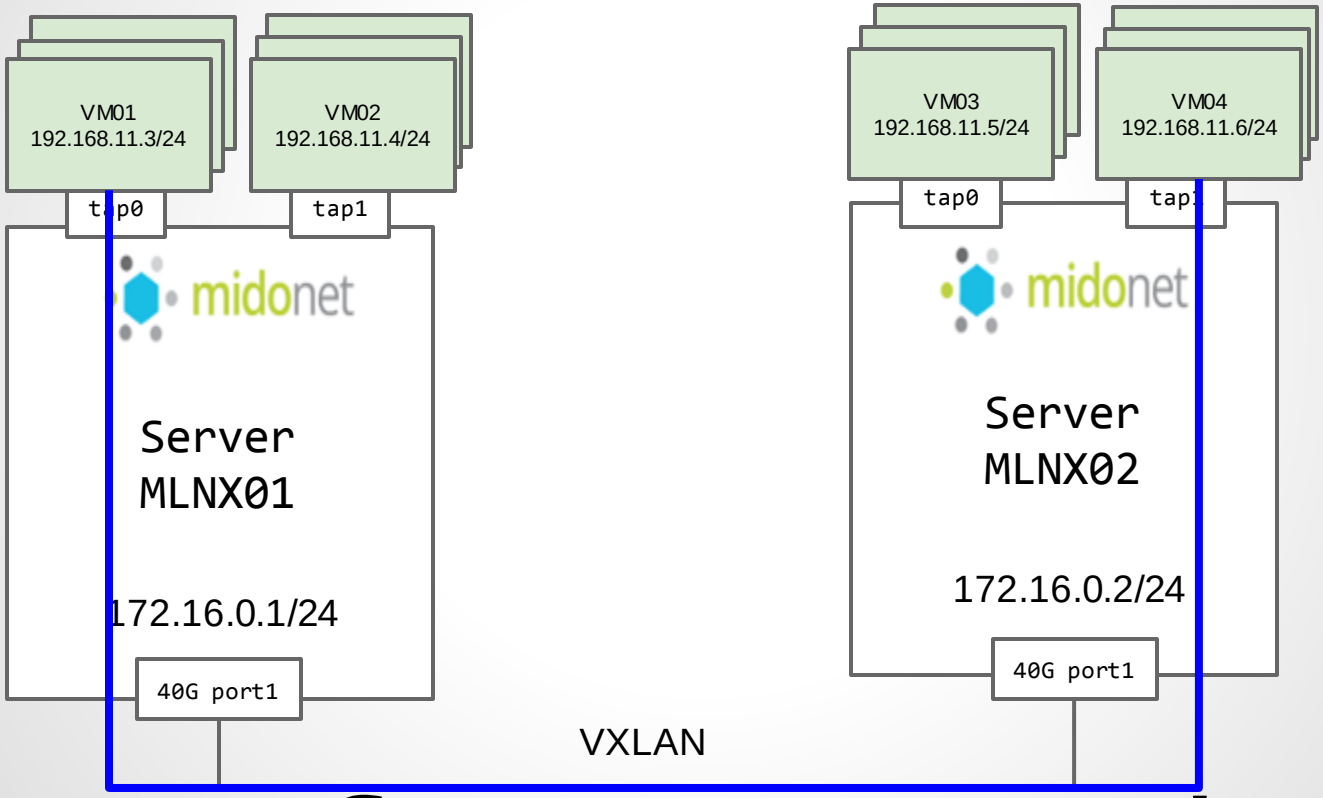
*現在の最新カーネルにアップデートすると適用される.

テスト環境(オフロード無し Ubuntu14.04)



iperf: 7.2Gbps

テスト環境(オフロード有り Ubuntu14.04)



iperf: 8~10.2Gbps

今後の課題

- 40Gbpsを使い切りたい
  (複数のVM間通信).
- KernelやNIC Driverの組み合わせや環境の違いによっては, 期待している性能速度を出せない場合があるので解決していきたい.