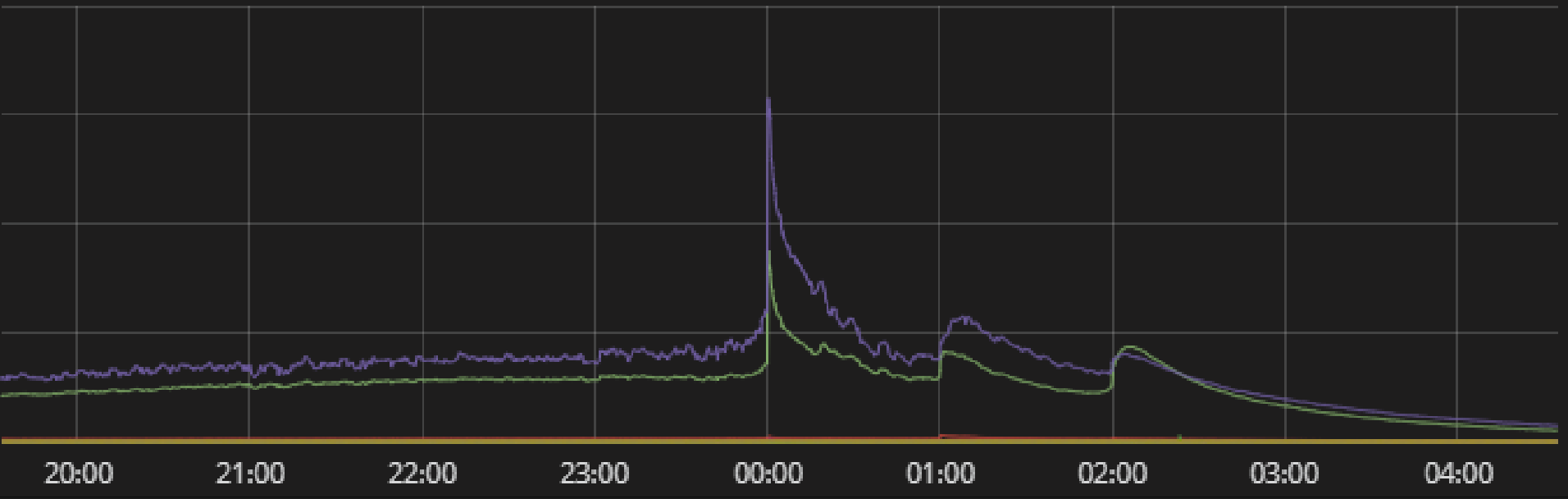
LINEのネットワークを ゼロから再設計した話

JANOG43 Meeting 2019/01/24

Masayuki Kobayashi
LINE Corporation

インフラの規模感

2019年1月1日 イベントトラフィック



165M+

Active Users in JP, TW, TH and ID

30,000+

Physical Servers

1Tbps+

User Traffic

直面していた課題

アーキテクチャの非効率性

1. **キャパシティの不足**

   East-Westトラフィックの増加によってボトルネックが発生

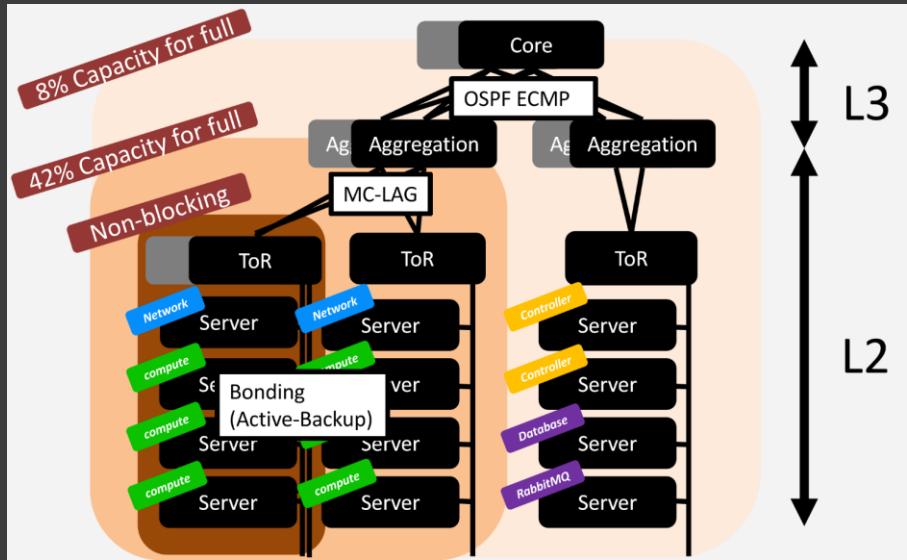
   水平スケールが困難な2N構成

2. **運用負荷の増大**

   複数のプロトコルと冗長化技術

   要件に応じたサーバ配置の限界

   手作業の運用による負荷



直面していた課題

課題を根本から解決することを決意

1. **アーキテクチャレベルでの見直し**

   East-Westトラフィックをノンブロッキングでさばくことができる

   N+1で水平スケールが可能な構成

2. **運用負荷を下げる**

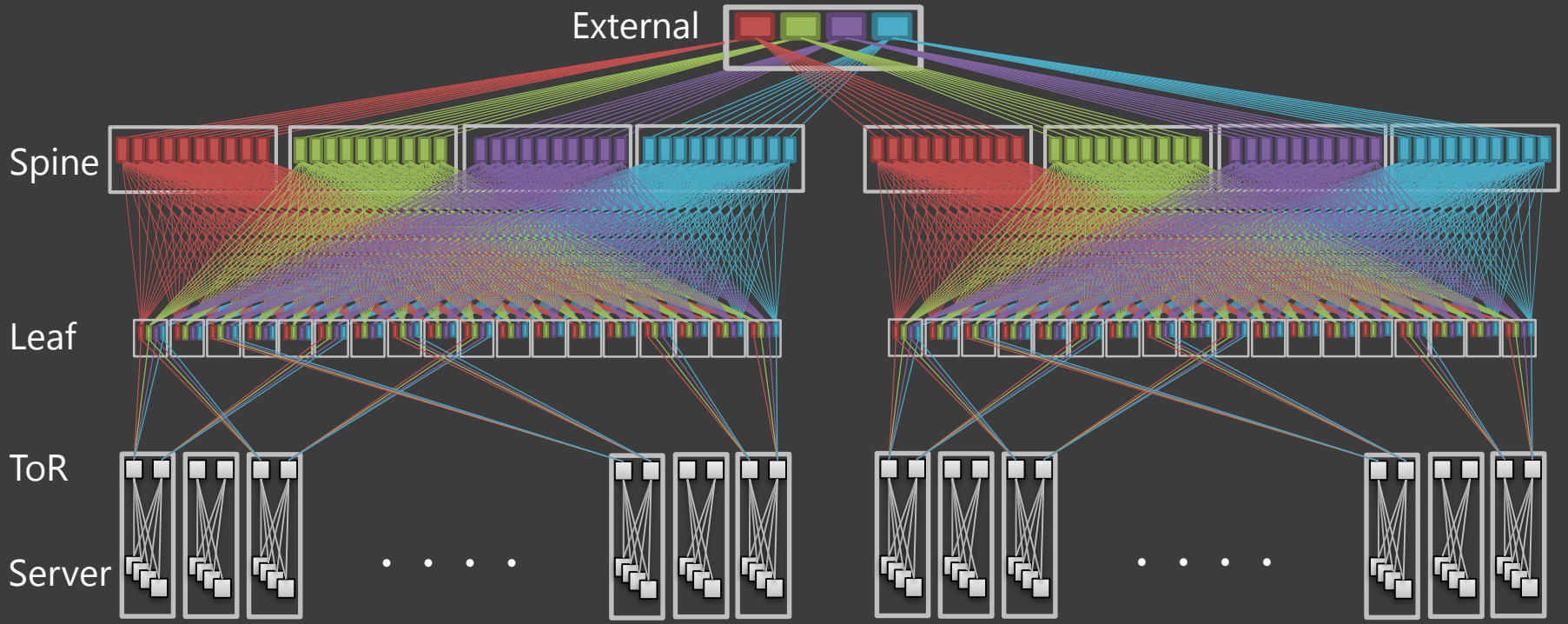
   ネットワークのシンプル化

   オープンかつ最少のプロトコルで構成され、ステートレスなネットワーク

   ネットワーク運用の自動化と効率化

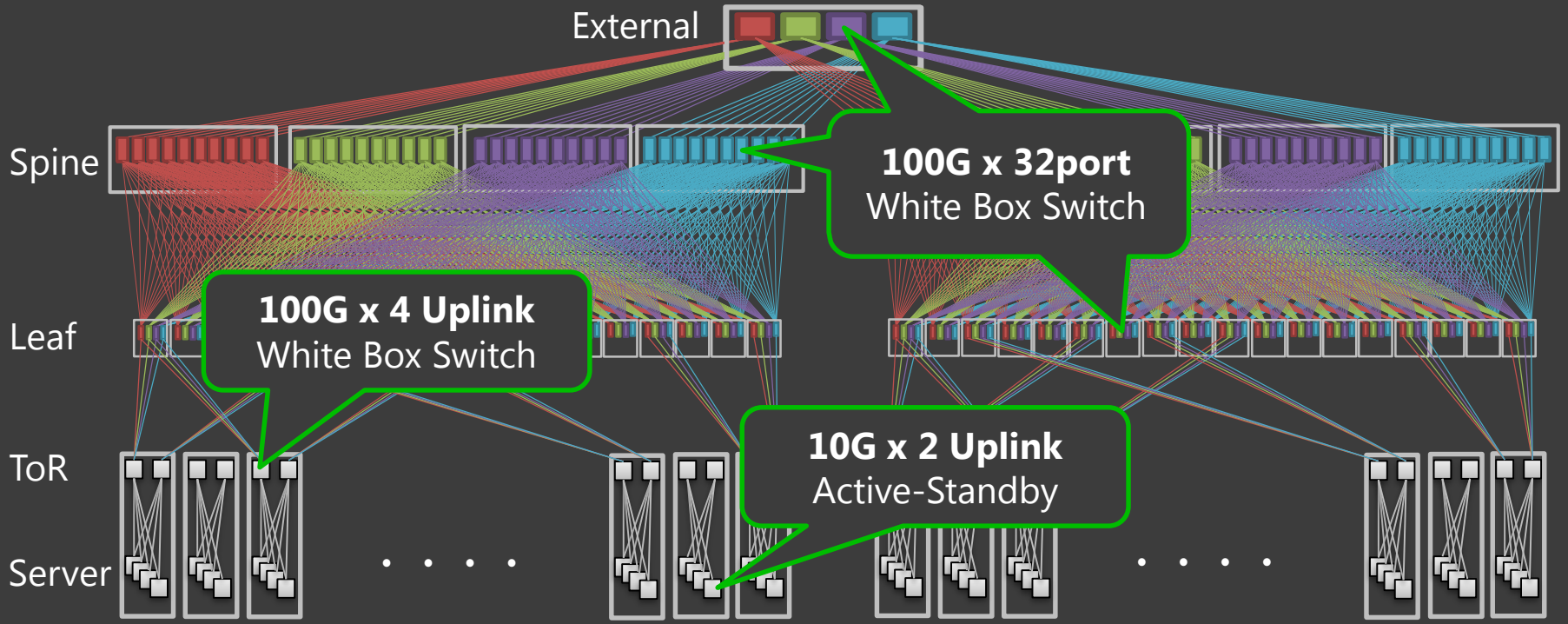
新しいアーキテクチャの概要

基本設計に3階層のCLOS Networkを採用



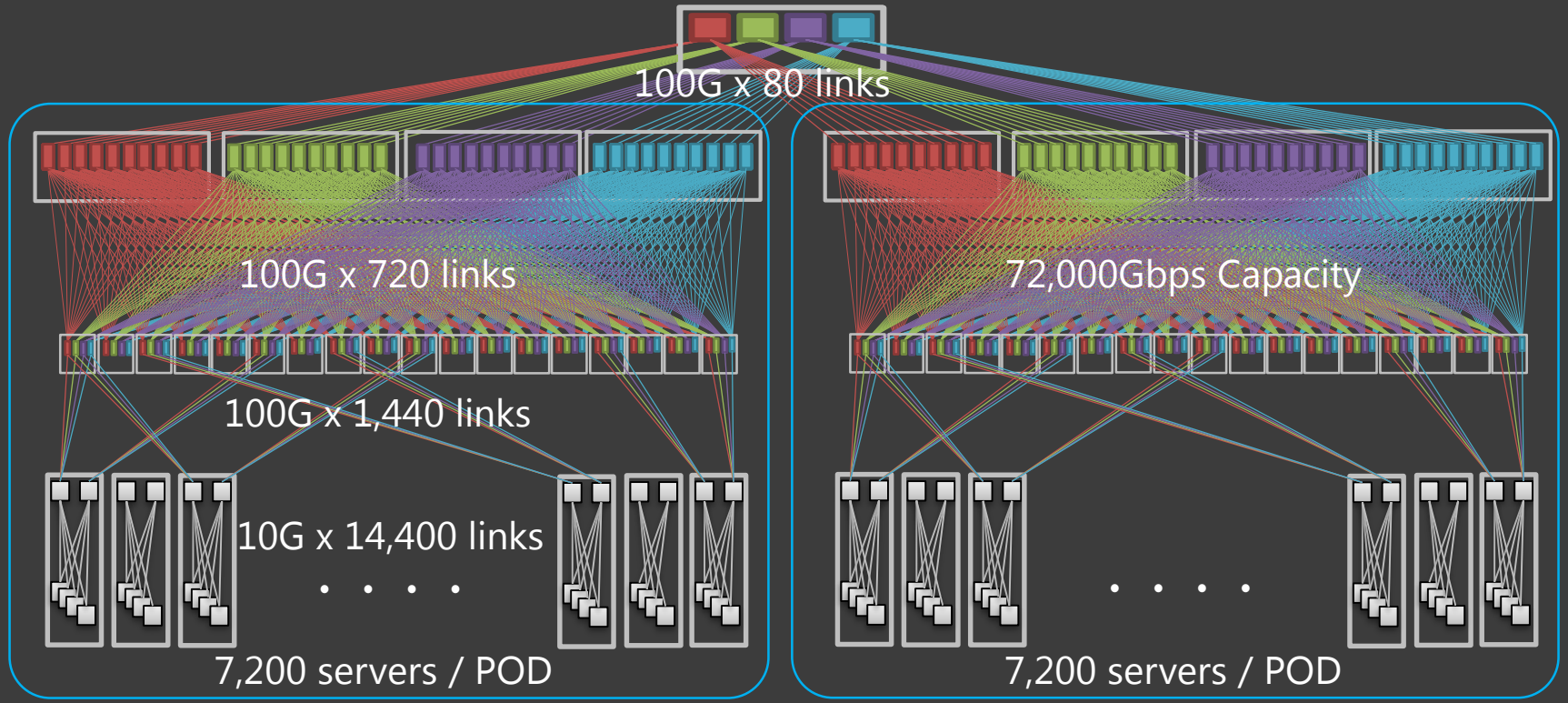
新しいアーキテクチャの概要

すべてホワイトボックススイッチで構築



新しいアーキテクチャの概要

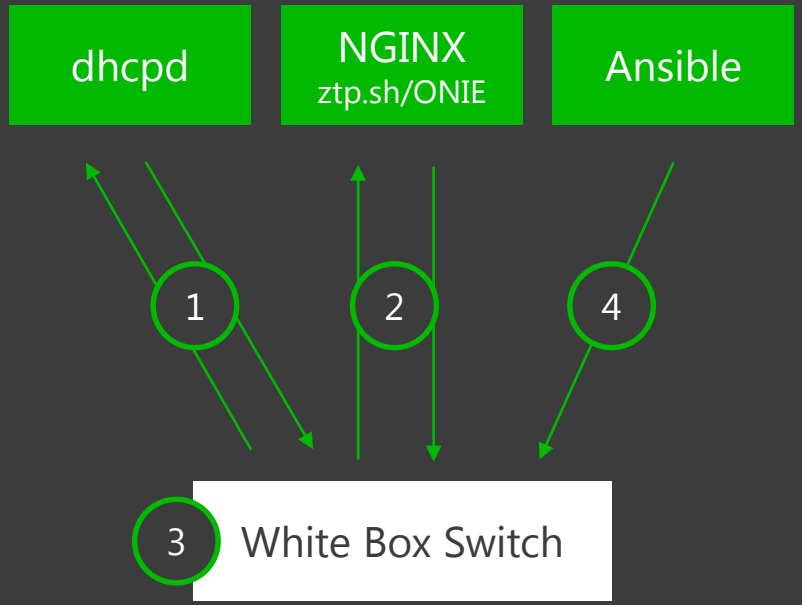
すべてのサーバが10Gbps Non-blockingで通信できる設計



CLOS Networkの構築方法

ZTPとAnsibleで構築時間を大幅に短縮

1. 電源投入後、DHCPで管理IPを割り当て
2. ZTPスクリプトをダウンロードして実行
    I. 工場出荷時のNOSのバージョンをチェック
    II. バージョンが異なる場合はONIEで再起動
    III. 利用するバージョンのNOSをインストール
    IV. 再起動して1に戻る
3. ZTPによるライセンスの投入と管理設定の実施
4. Ansibleの実行でサービス利用可能



1,000台以上の規模でも数時間で構築が完了

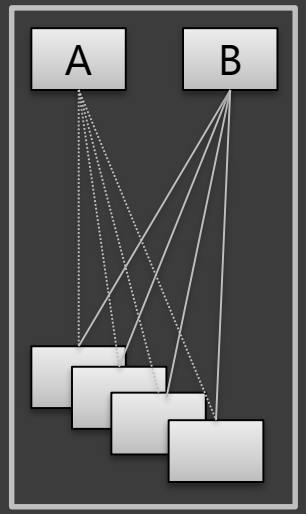
ToRとサーバの接続がL2であることの問題

トラフィックの切り替えでパケットロスが発生

L2接続

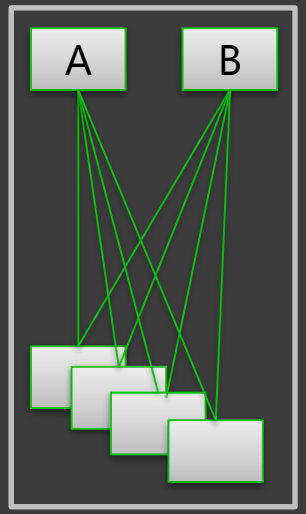
MC-LAG

BONDING



VS

L3接続



BGP

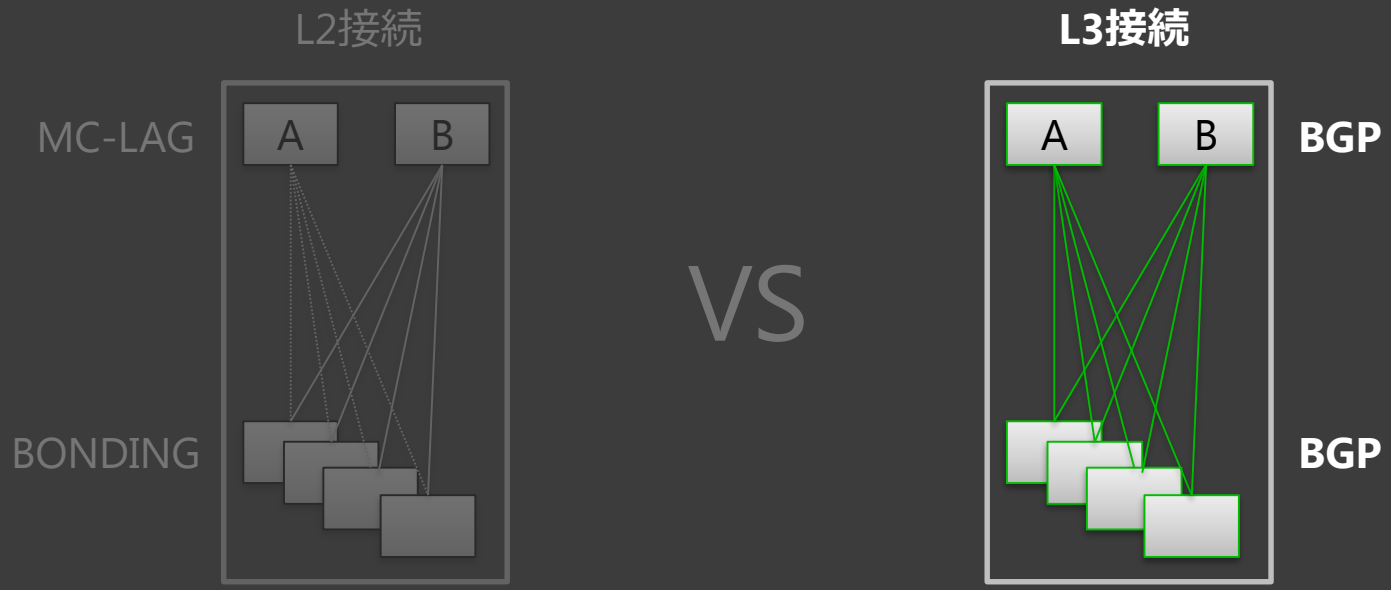
BGP

サーバ管理者への依頼が必要

ToR側で切り替えが可能

ToRとサーバの接続がL2であることの問題

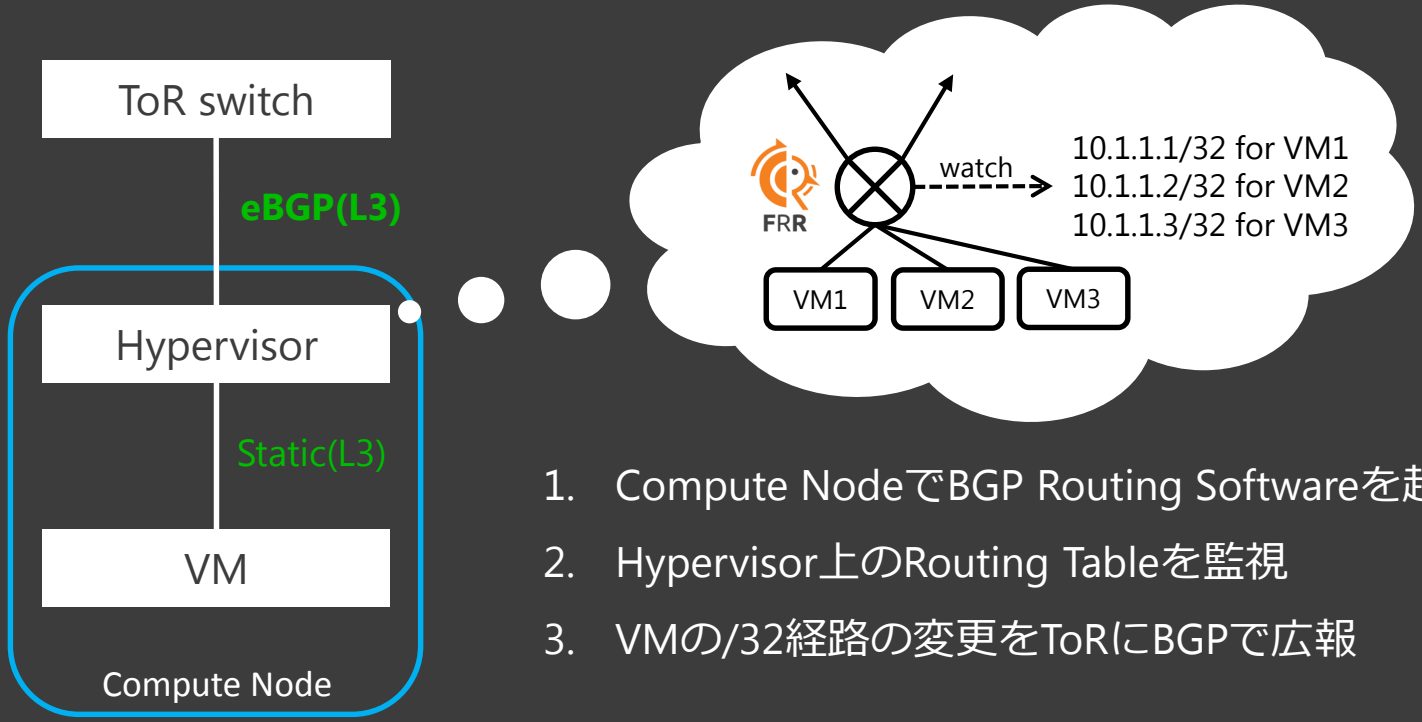
解決策: ToRをL2とL3の境界にしない



ToRのメンテナンス性が大幅に向上

BGP Routing on the Host

/32の直接ルーティングのみでエンドツーエンドの接続性を担保



1. Compute NodeでBGP Routing Softwareを起動
2. Hypervisor上のRouting Tableを監視
3. VMの/32経路の変更をToRにBGPで広報

VXLANなどのL2延伸が不要

Excitingly simple multi-path OpenStack networking: LAG-less, L2-less, yet fully redundant
[Blog] https://engineering.linecorp.com/ja/blog/openstack-summit-vancouver-2018-recap-2-2/

CLOS Network + Routing on the Host構成

達成したこと

新規構築と機器交換にかかる時間が大幅に短縮された

ネットワークのボトルネックを意識しなくてよくなった

ラック毎にVLANを管理する必要がなくなった

ベンダ依存のプロトコルや冗長化技術がなくなった

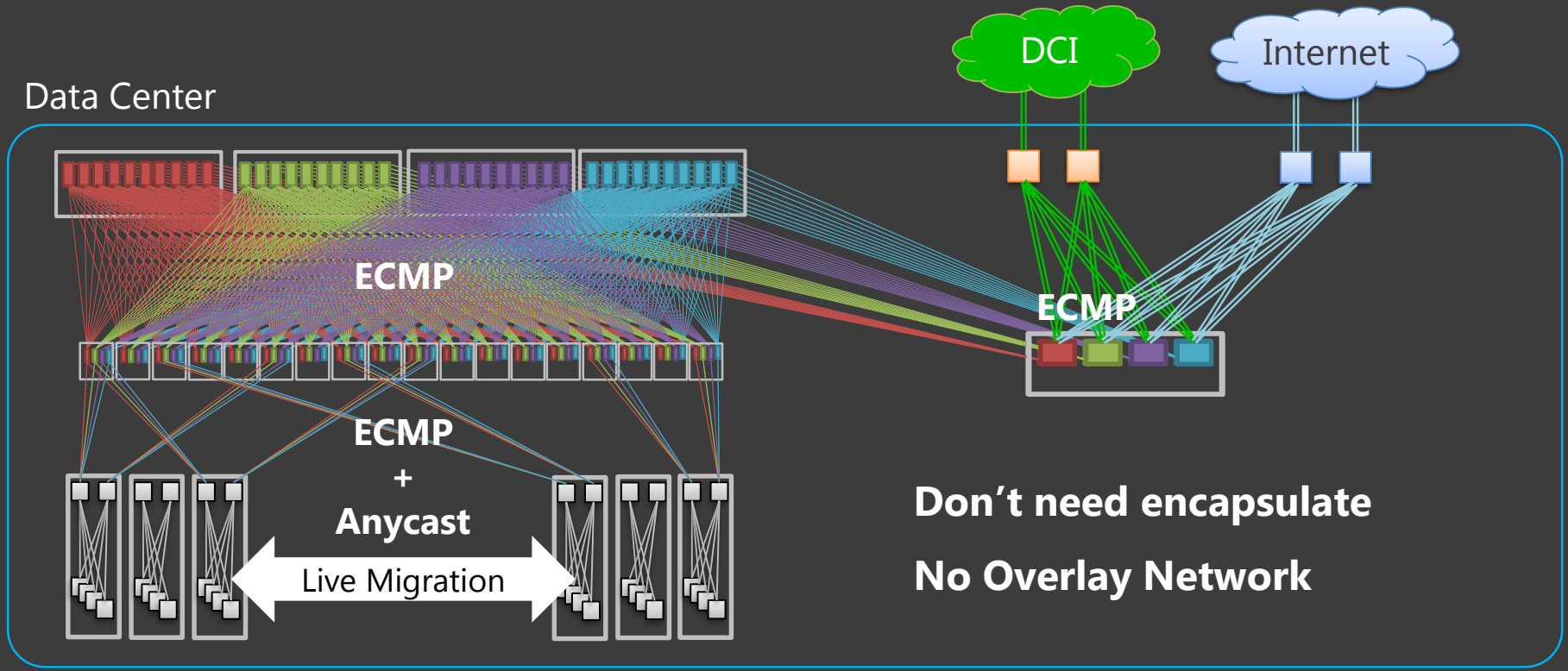
L2のオーバーレイネットワークを作らないことで設計の複雑化を回避した

トラフィックの迂回が簡単になり、メンテナンスがしやすくなった

機器の台数分の設定を手動で作る必要がなくなった

フルL3のシングルテナントネットワーク

ネットワークに必要な機能を一つの面で提供



BGP in DCをスケールさせるための仕組み

個別のパラメータ管理を最小限に

Separate 4-byte ASN per Node

Lo: 10.128.100.45/32

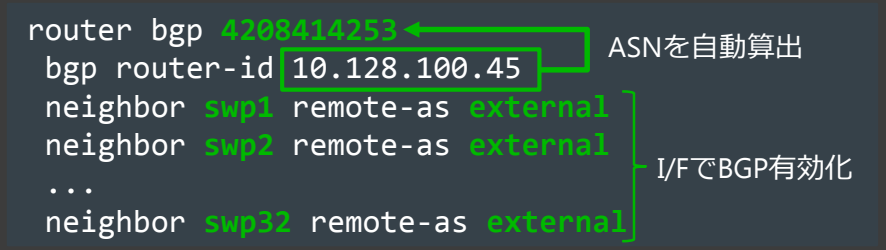
ASN: 4208414253

10.x.y.z → x*(2^16) + y*(2^8) + z + 4200000000

DCで必要なAS番号の数は1万以上

Loopback IPから一意なPrivate ASNを自動算出

BGP Unnumbered



RFC5549 Extended Next Hop Encoding capability

IPv6 LLAのeBGP session上でIPv4経路を交換

P2P Linkの明示的なアドレス設定が不要

サーバまでeBGP接続するLINEのネットワークでは、これらが必要不可欠

サーバから見た経路情報

out方向のトラフィックを片方のToRに寄せる

```
# vtysh -c "show ip bgp"
(snip)
   Network     Next Hop    Metric     Path
*> 0.0.0.0     eth0.100                4208258575 4208258904 4208258884 65001 38631 i

*              eth1.100     100        4208258576 4208258908 4208258884 65001 38631 i
```

bgp always-compare-med

ToR　Leaf　Spine　External　Router

Best Path Selection

```
# vtysh -c "show ip route bgp"
(snip)
B>* 0.0.0.0/0  [20/0] via fe80::ee0d:9aff:fe57:63b4, eth0.100, 09w2d21h
```

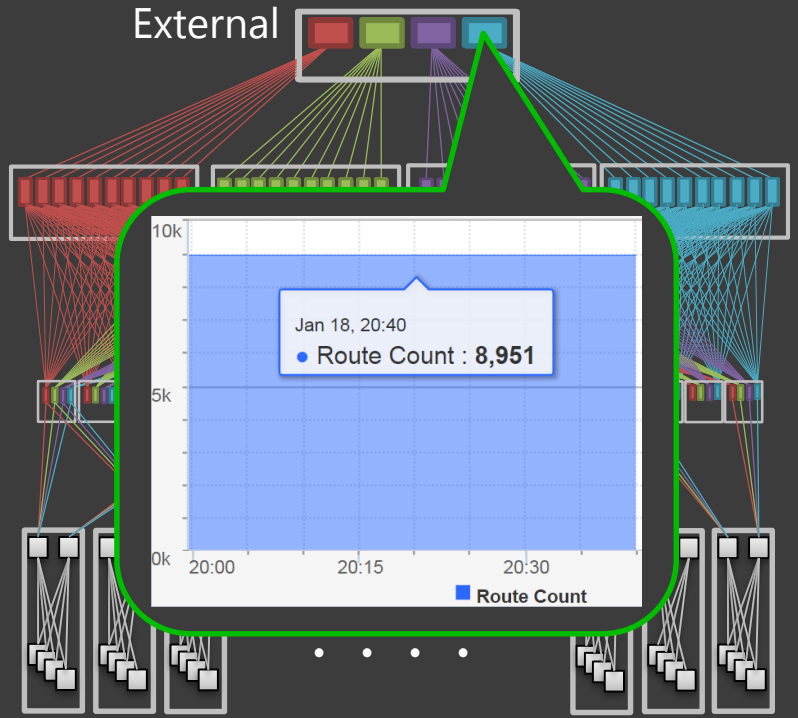
RFC5549 Next-Hop

Linux Kernel Routing Table

```
# ip r
default via 169.254.0.1 dev eth0.100  proto zebra metric 20 onlink
```

経路数

実際どれくらい？



Externalの機器のキャパシティ

- 設計上、最も多くの経路を学習する
- 他の機器とは異なる設定が多くなりがち
- 各種リソースの消費量に注意が必要

Link-Local Next-Hopを使うメリット

- P2P Linkのアドレスを広報する必要がない
- FIBのリソースを無駄に消費しない
- 外部から見えない

実際に起きた障害

経路集約の負荷によるBGP Peer Down

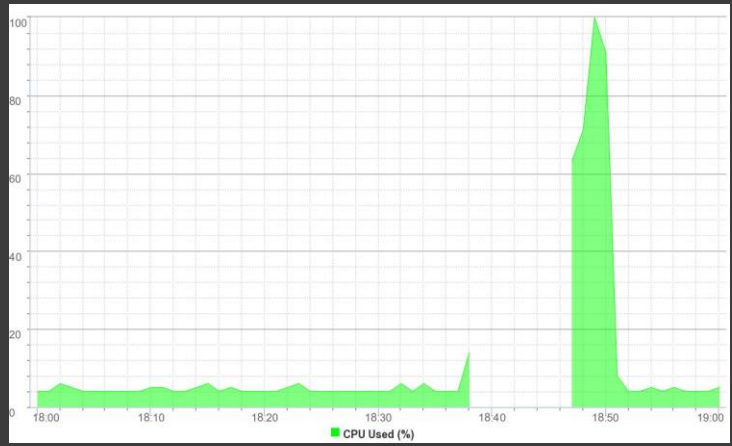
経路集約の設定を止めて解決
経路を生成する方式に変更

/16

/32 Update

1. 数百台のサーバを同時に再起動
大量のBGP Updateが走る

2. CPU負荷が急上昇して通信断



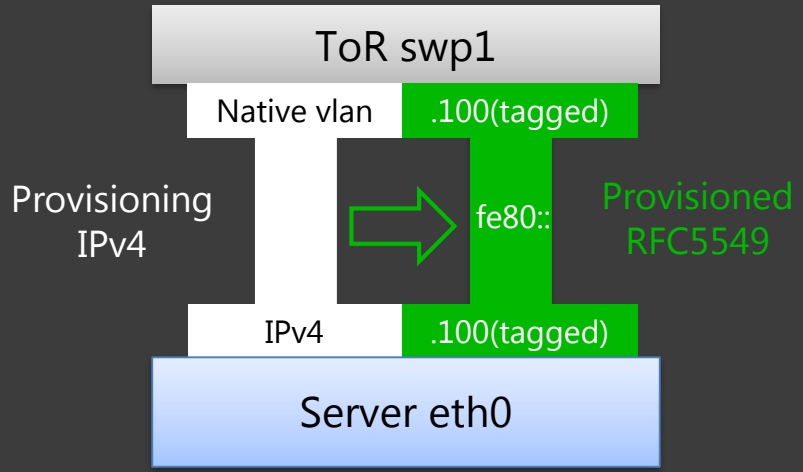
aggregate-addressの設定が原因

※ FRR 4.0以降では発生しない

運用面での工夫

フルL3化したことによって考慮が必要なこと

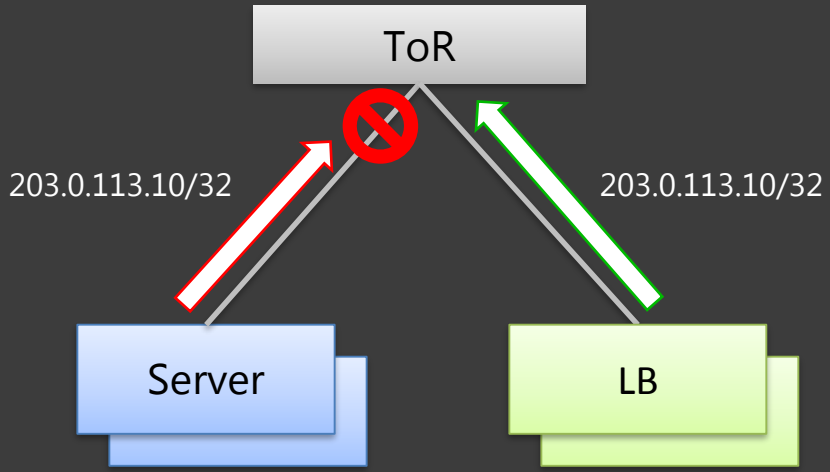
サーバのセットアップ



PXE Boot時はRFC5549が利用不可

完了後にサーバ側のProvisioning用IPを削除

経路フィルタ

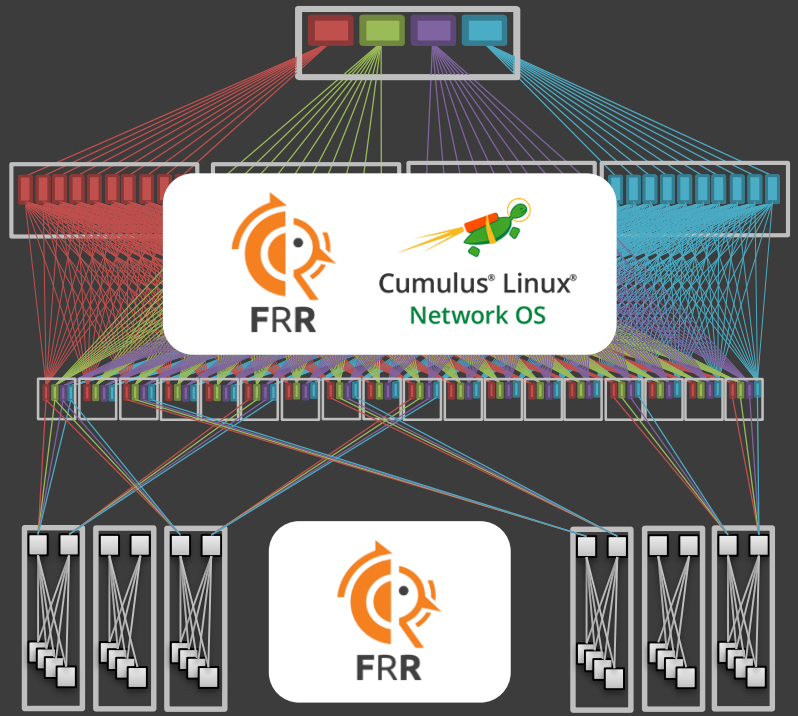


サーバの経路ハイジャックを防止

ToRでサーバの広報経路を適切にフィルタ

ホワイトボックススイッチの採用理由

必要なBGP Capabilityを満たすNOS + 運用の効率化



BGP Unnumbered

ピアの自動発見とRFC5549上での通信

I/Fを繋ぐだけで同時に実現できる実装

Hostname Capability for BGP

Open MessageにFQDNをエンコード

サーバの接続情報の取得に利用

サーバとスイッチで共通の実装を利用

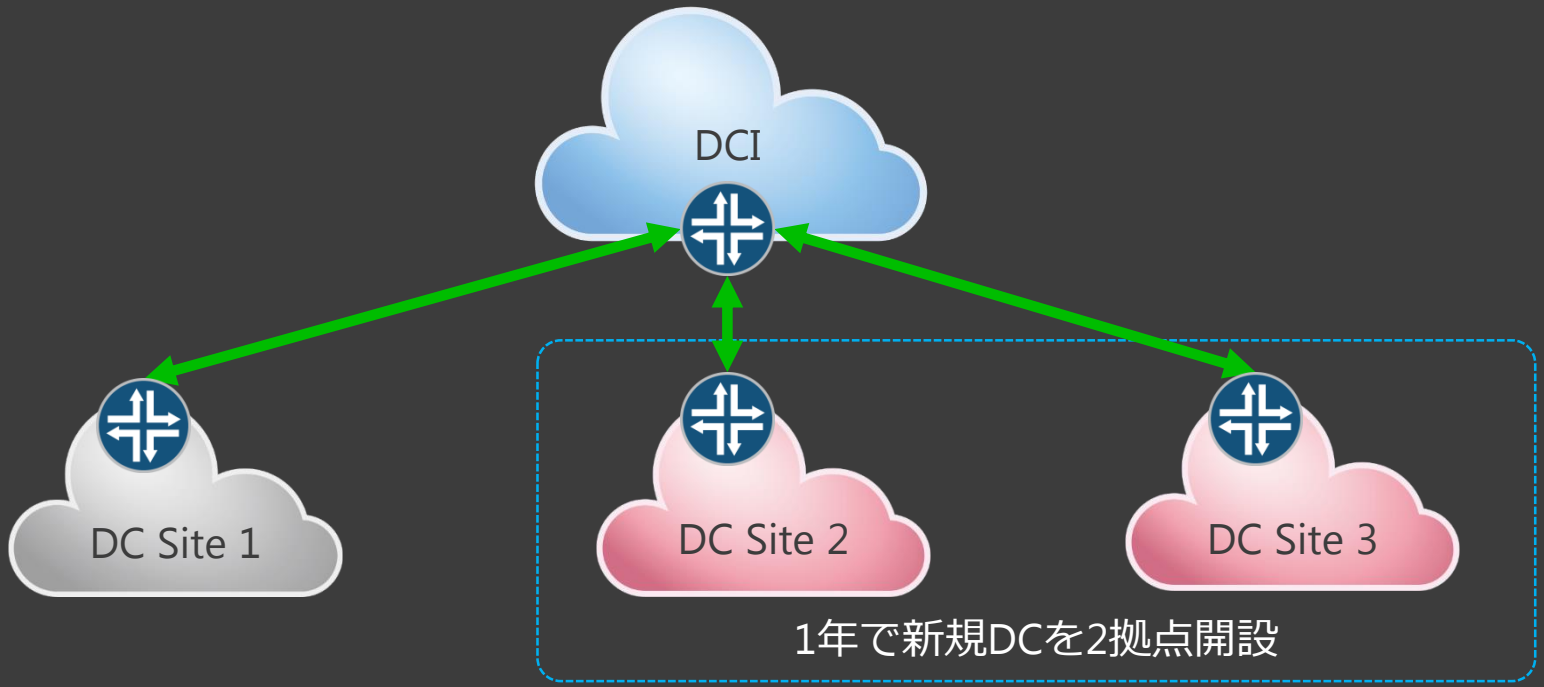
LINEの新DCアーキテクチャまとめ

Use BGP Everywhere

High Capacity
Fully L3 Redundant
Protocol Reduction
Horizontally Scalable

データセンタ間ネットワーク

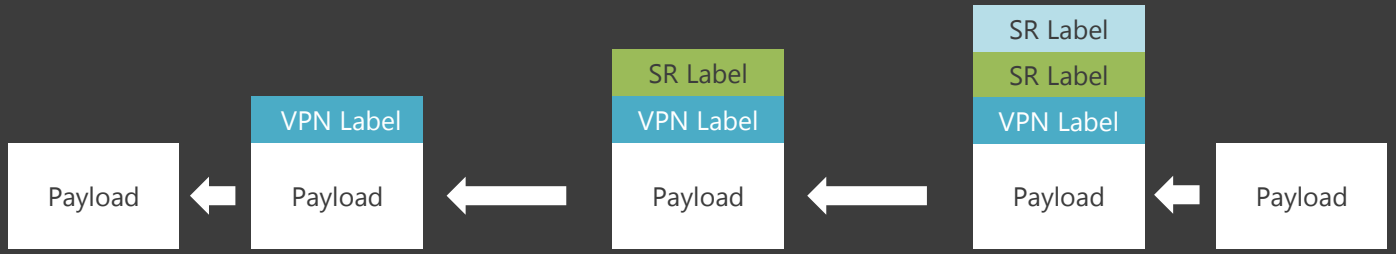
拠点間をL3で相互接続する構成



⇒サーバ間通信のためのL2延伸はしない

データセンタ間ネットワーク

MPLS-L3VPN + Segment Routing



データセンタ間ネットワーク

SR-MPLSの採用理由

C-PlaneのProtocol

IGPで動作するのでシンプル

- RSVP/LDPフリーのネットワーク
- パケットにステート情報を持たせることが可能

複雑な運用はしない

- 帯域確保はしないので単純な機能で足りる
- TI-LFAの利用を想定して導入

最適な導入タイミング

- 検証・構築と、実用段階の時期が一致(2017年)

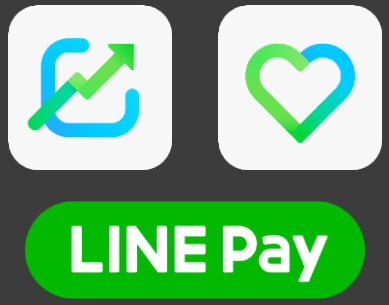
⇒ラベルスタック数やSID管理に注意すれば実運用は問題ないと判断

今後の展望と技術的挑戦の一例

※以降の内容は研究・開発中のものであり、商用環境へ導入が確定したものではありません。

次世代データセンタでの実現目標

ビジネスニーズに対して迅速に対応できるインフラ



Fintech Business

これまでとは異なる要件のネットワークが増加

現在は都度インフラを構築してサービスを提供

アンダーレイネットワークの断片化

構築にかかる時間の増加

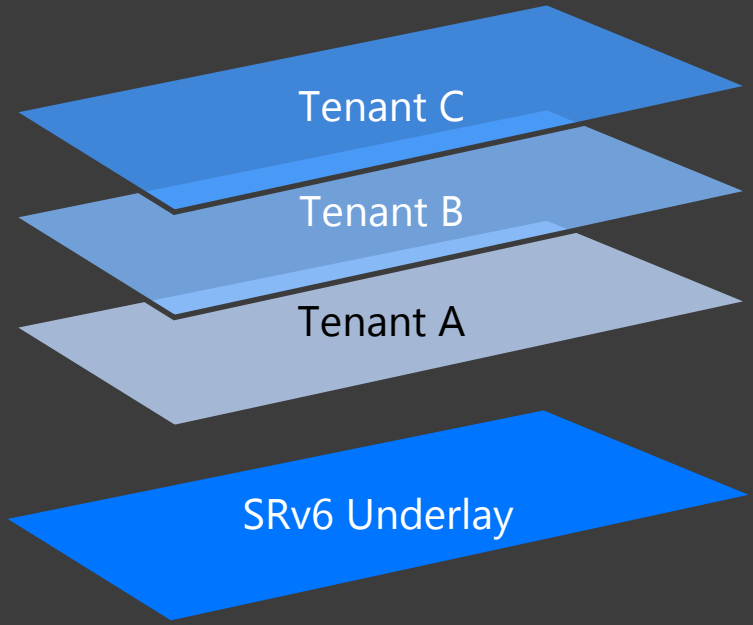
インフラエンジニアの負担増加

⇒一つのインフラの上で様々な要件を達成できるようにしたい

次世代データセンタでの実現手法

WIP

Multi-Tenancy & Service Programmability



柔軟にスケールするオーバーレイ

テナントで分離・独立したセキュリティ

将来的なサービスチェイニング

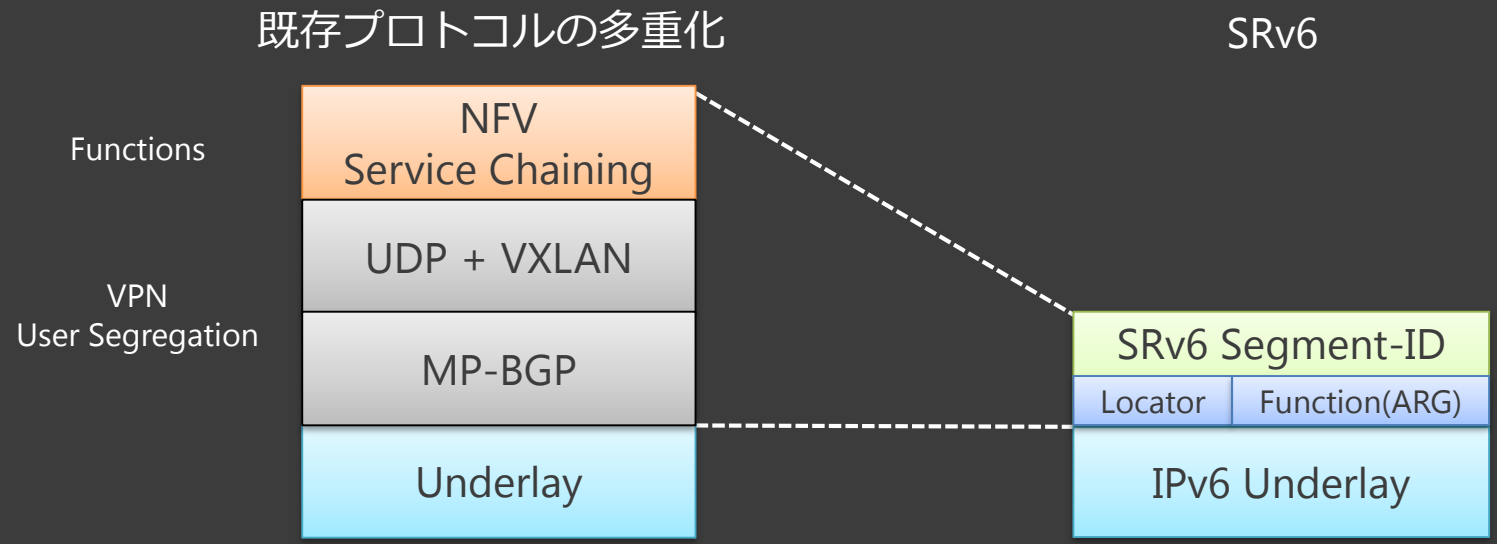
一つに統合されたアンダーレイ

有効な実現手法の一つとしてSRv6に注目

次世代データセンタでの実現手法

WIP

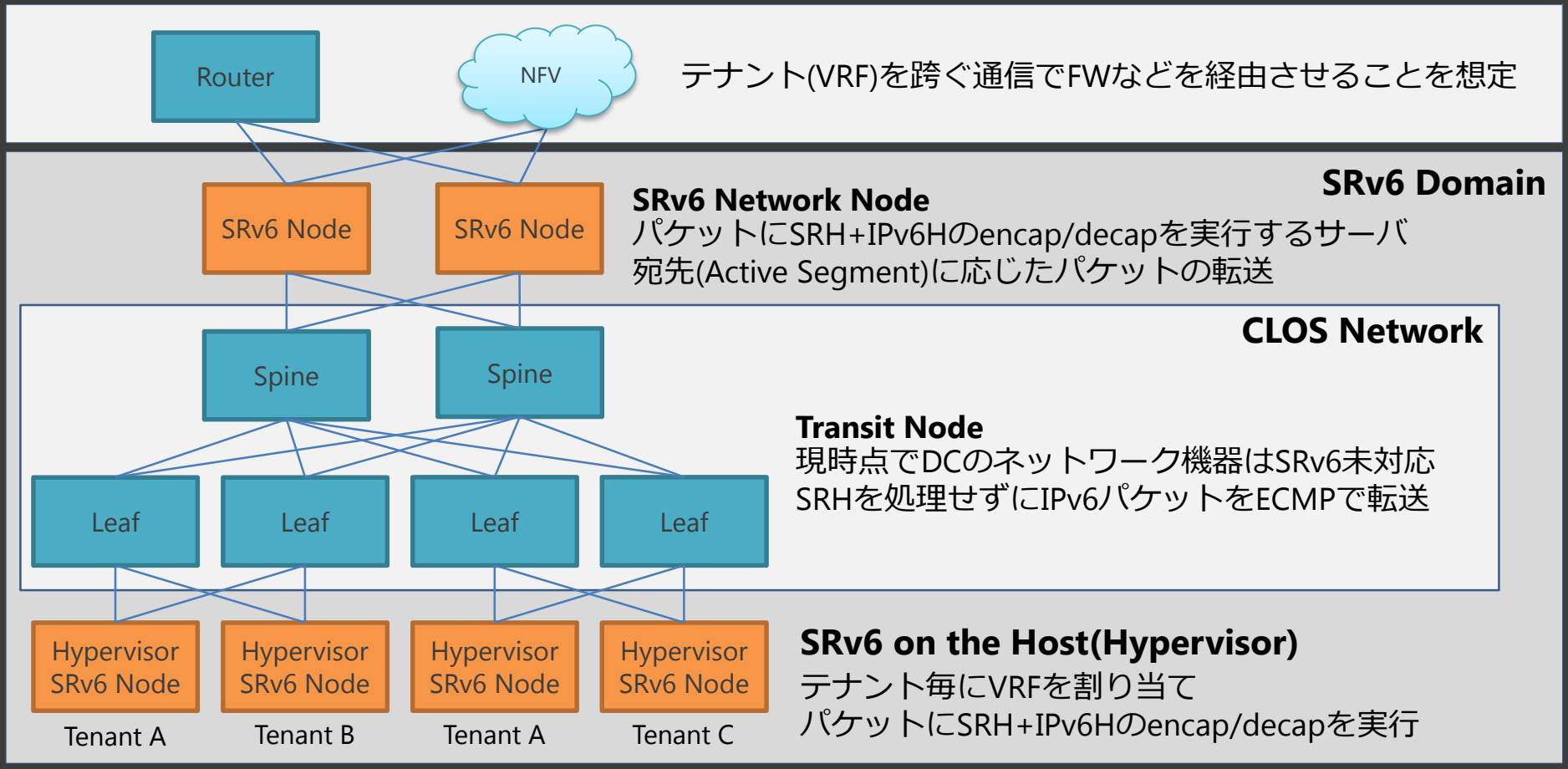
SRv6 in DCへの期待



⇒DC NWへの要求を128bit SIDのFunctionで表現する

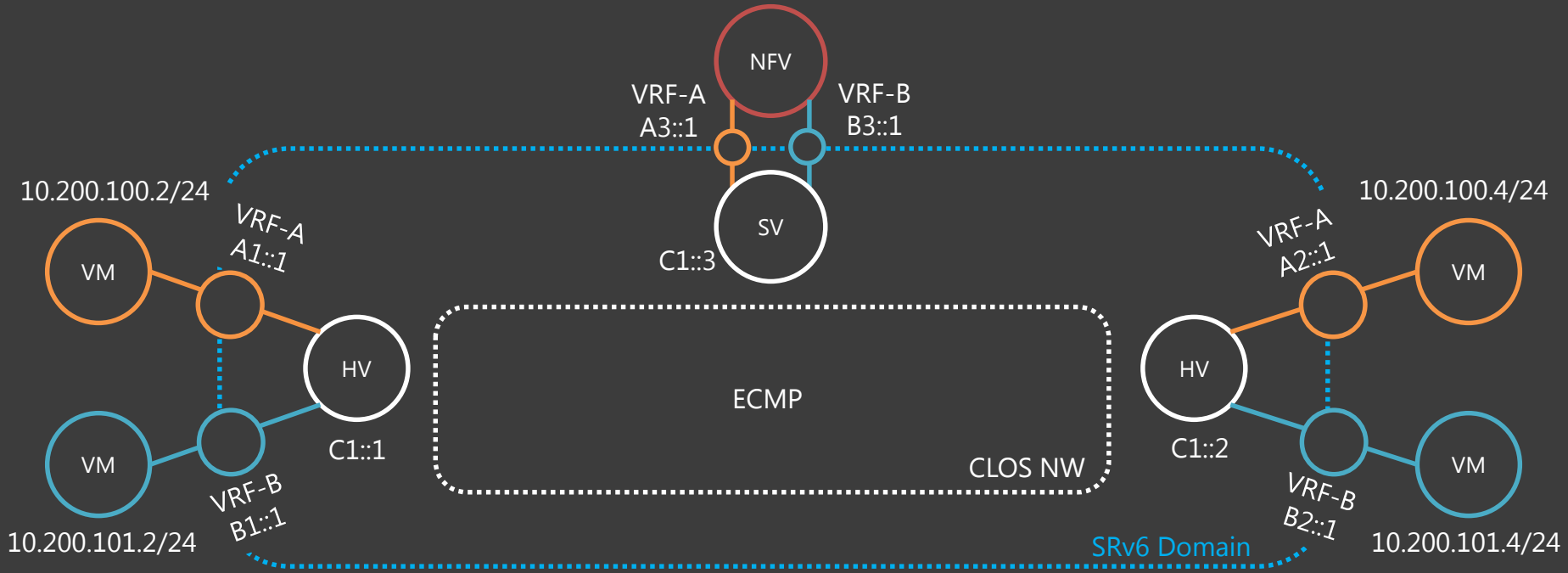
SRv6-L3VPN in DC (Concept Design)

WIP



SRv6-L3VPN in DC (Concept Design)

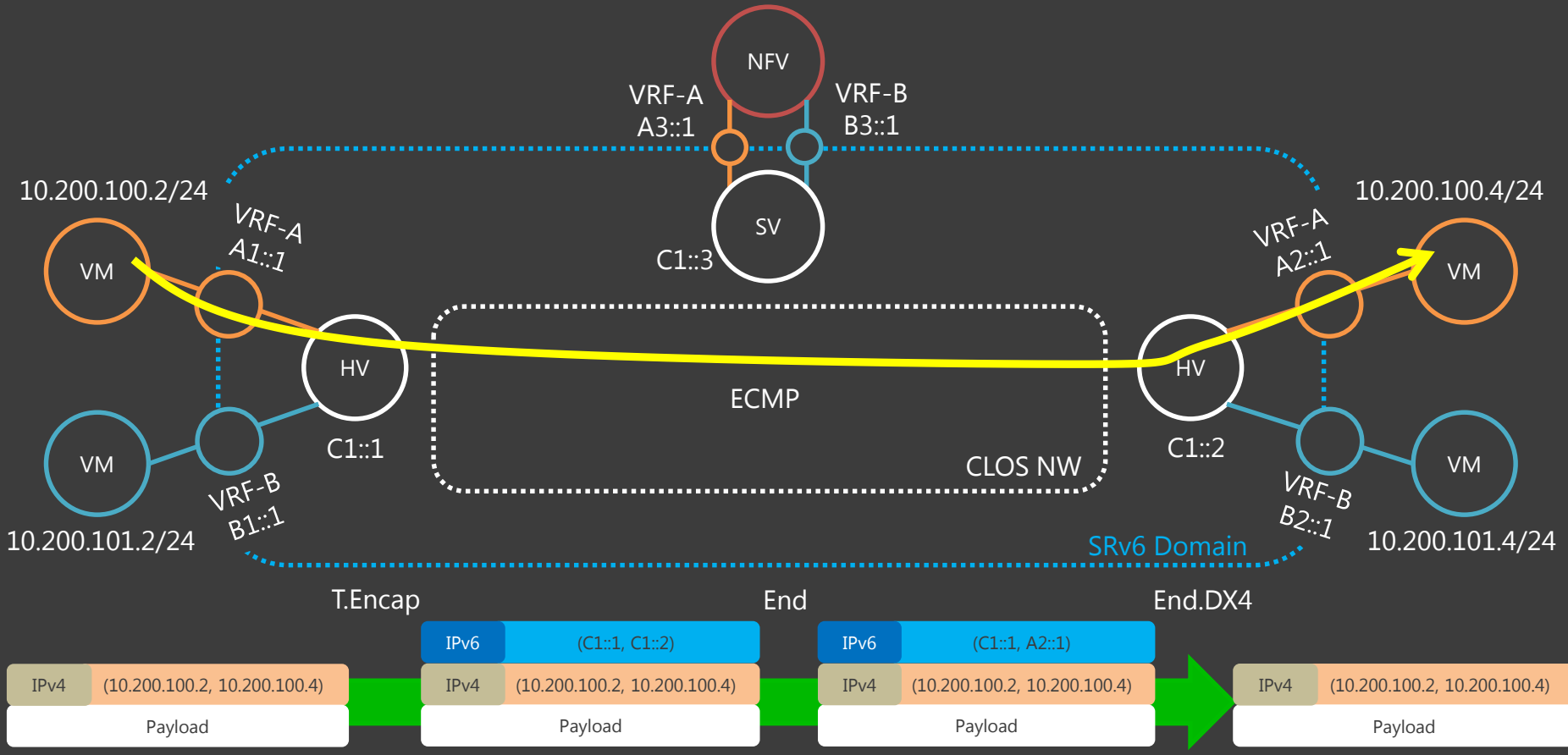
WIP



HV = Hypervisor(SRv6 enable)
SV = Server(SRv6 enable)

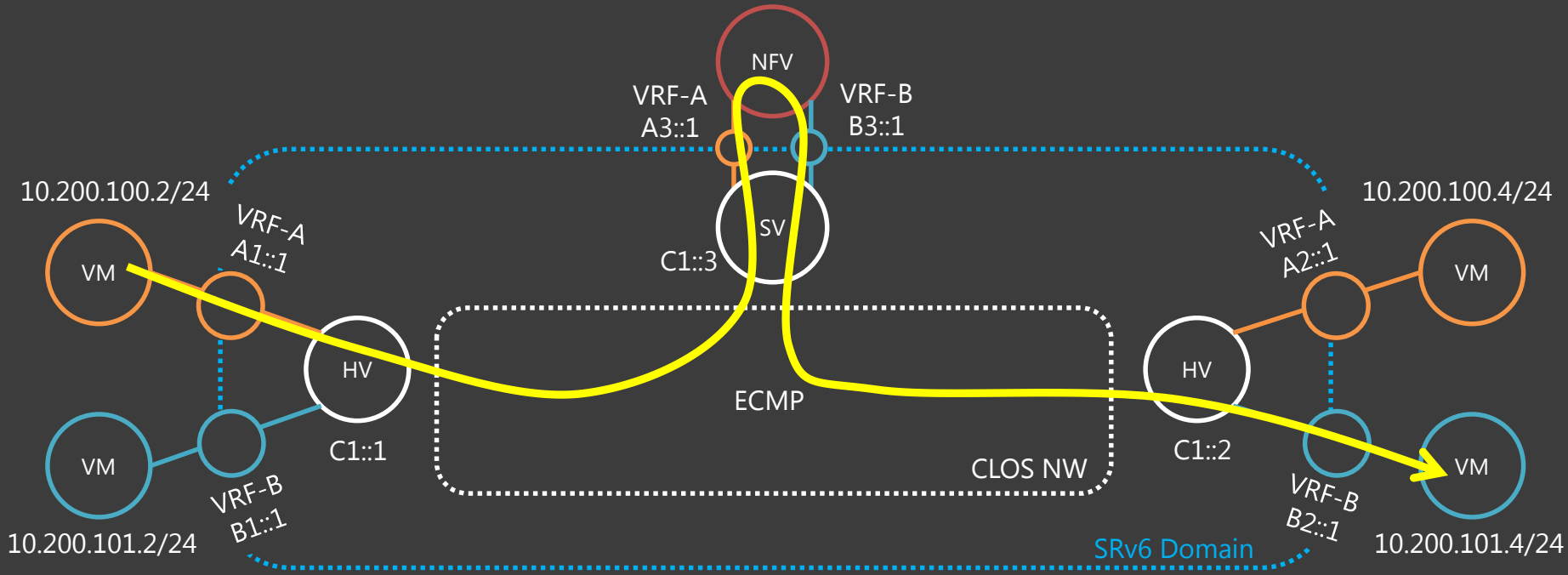
SRv6-L3VPN in DC (Concept Design)

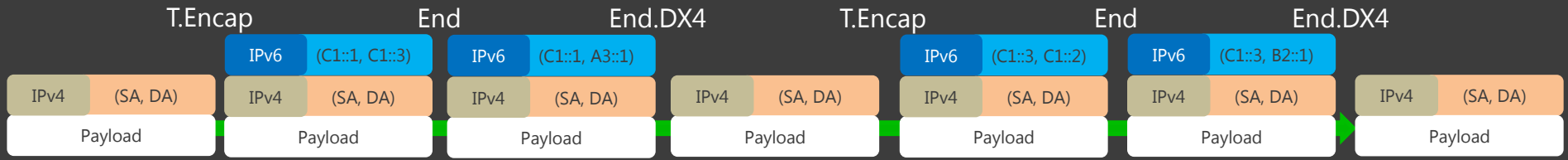
WIP



SRv6-L3VPN in DC (Concept Design)

WIP





SRv6の取り組みについてもっと詳しく

勉強会

- “SRv6 入門”
  - – 浅間 正和 (有限会社銀座堂)
- “Control Plane的何か（仮）”
  - – 河野 美也 (Cisco Systems)
- “（仮）SRv6の標準化やPoC”
  - – 松嶋 聡 (ソフトバンク)
- “（仮）SRv6データプレーン実装とモバイルへの適用”
  - – 海老澤 健太郎 (TOYOTA ITC)
- “LINEにおけるデータセンターでのSRv6ユースケース”
  - – 土屋 俊貴 (LINE株式会社)
- “IP anycast + SRv6によるNetwork APIの提供”
  - – 宮坂 拓也 (KDDI総合研究所)
- “Trellis (SR based DC Fabric Solution)(仮)”
  - – 調整中 (SK Telecom)
- “SKT’s R&D on Telco Networks(仮)”
  - – 調整中 (SK Telecom)
- “Kamuee SRv6対応 調査と報告と実装ステータス”
  - – 城倉 弘樹 (NTTコミュニケーションズ)
- “Huawei SRv6 update”
  - – 高嶋 隆一 (Huawei Technologies Japan K.K.)

※今回は SRv6 に関する発表のみのプログラムとなります。

お知らせ

ENOG55 Meeting 開催のお知らせ

by masakazu • 2018年12月13日 • 0 Comments

開催概要

日時

2019/2/22(金) 14:00～19:00(予定)

検証内容や現状の課題について話します

http://enog.jp/archives/2014

最後に

まだまだ改善が必要な事項、解決が必要な課題はたくさんある

- Legacy Networkとのインテグレーション
- 新たなアーキテクチャのネットワークへ組み込めないサーバ群
- 新たな監視方法の導入、機器の正常性チェック
- テストの自動化、設定の妥当性の検証

リファレンスデザインがまだ存在しない分野への取り組みも実施中

より運用負荷の低いスケールするネットワークへ


LINE