All-Photonics Networkをバックボーンにして EVPN Multi-Site Architectureで 分散DCを構築してみた

西日本電信電話株式会社

芳政信　吉川尚吾

アジェンダ

All-Photonics NetworkでDC間を接続した
サーバクラスタの簡易性能検証

All-Photonics Networkをバックボーンにした
EVPN Multi-Site Architectureの検証

アジェンダ

All-Photonics NetworkでDC間を接続した
サーバクラスタの簡易性能検証

All-Photonics Networkをバックボーンにした
EVPN Multi-Site Architectureの検証

自己紹介

NTT西日本

吉川　尚吾
ヨシカワ　ショウゴ

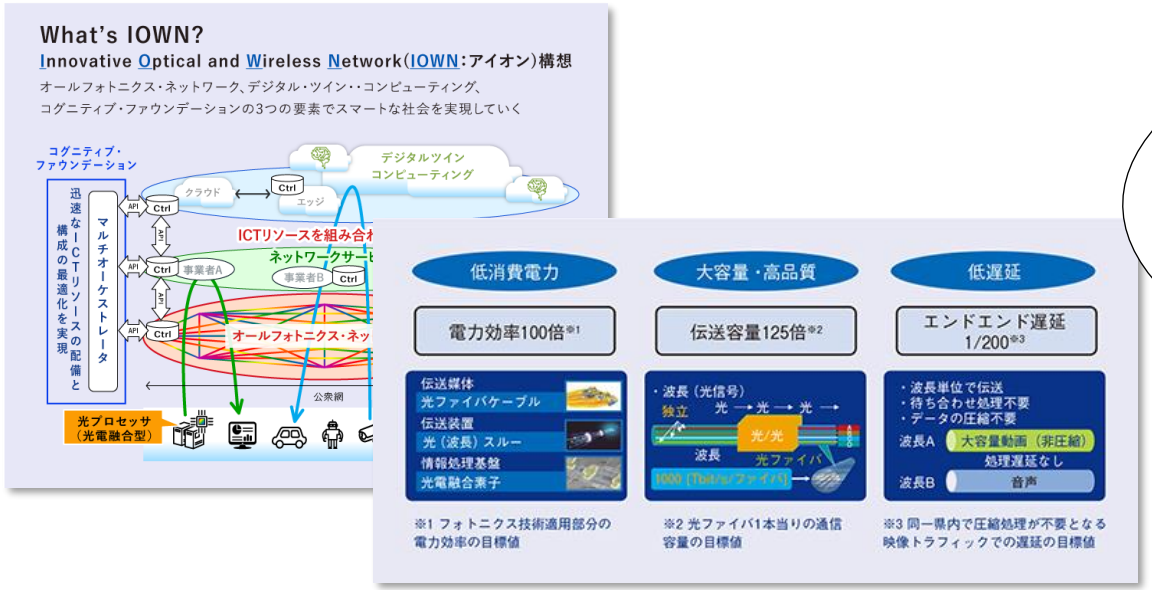
■ 入社から主にキャリアネットワークの保守・開発に従事
- NWマイグレーション・故障交換の地域収容局での現地対応
- 開発ではBGP・PPP・L2TPなどを解析

■ データ分析業務を経て、現在はR&D部門で技術評価・実証実験

All-Photonics Network (APN) とは

NTT西日本

NTTが推進する**IOWN構想**の要素のひとつ

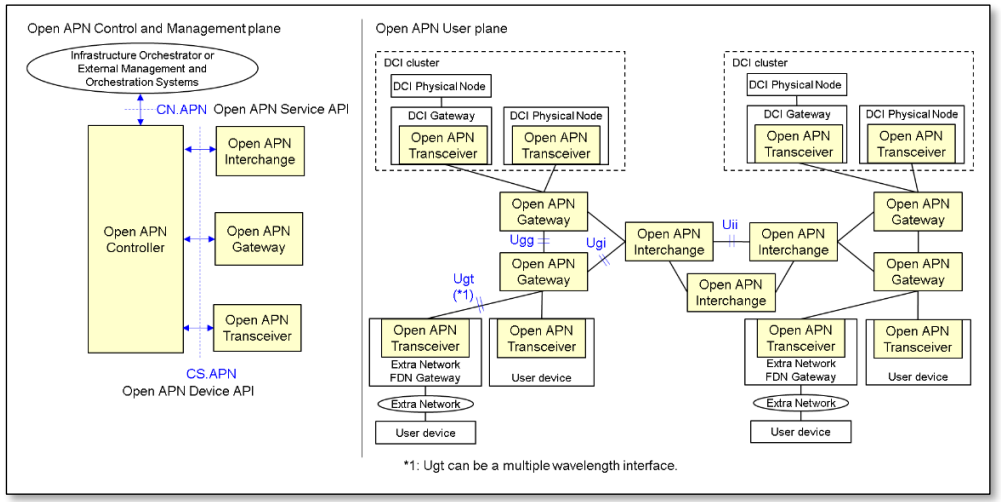


「すべてにフォトニクス（光）ベースの技術を導入し（中略）低消費電力、高品質・大容量、低遅延の伝送」

オールフォトニクス・ネットワークとはなにか｜NTT R&D Website より

All-Photonics Network (APN) の今

将来にむけた仕様策定



Open All-Photonic Network Functional Architecture より

構想実現をめざしフォーラムで議論

提供中のサービス

News Release

西日本電信電話株式会社
東日本電信電話株式会社

APN IOWN1.0の提供開始について

2023年3月2日

東日本電信電話株式会社（以下、NTT東日本）および西日本電信電話株式会社（以下、NTT西日本）は、IOWN構想※1の実現に向けた初めての商用サービスとして、通信ネットワークの全区間で光波長を専有するオールフォトニクス・ネットワーク（All-Photonics Network、以下、APN）IOWN1.0を2023年3月16日（木）に提供開始いたします。また、APN IOWN1.0上での遅延の可視化と遅延調整機能を備えた端末装置「OTN Anywhere」も販売開始いたします。

※1 IOWN（Innovative Optical and Wireless Network）構想とは、あらゆる情報を基に個と全体との最適化を図り、光を中心とした革新的技術を活用し、高速大容量通信ならびに膨大な計算リソースなどを提供可能な、端末を含むネットワーク・情報処理基盤の構想です。詳しくは以下ホームページをご覧ください。
IOWN構想とは？https://www.rd.ntt/iown/index.html

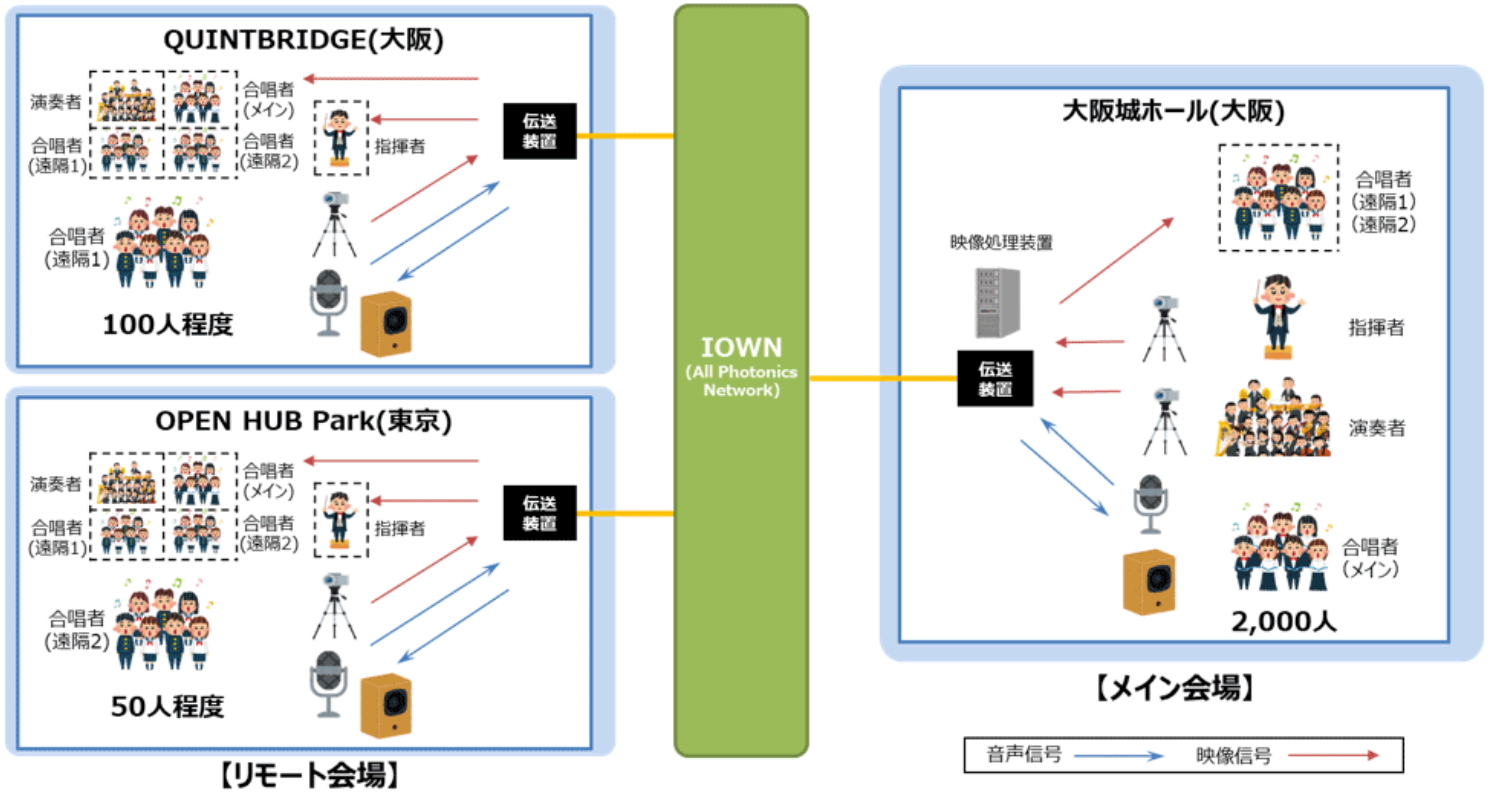
【NTT西日本】APN IOWN1.0の提供開始について より

OTU4をIFとした通信サービス

All-Photonics Network (APN) の今

NTT西日本

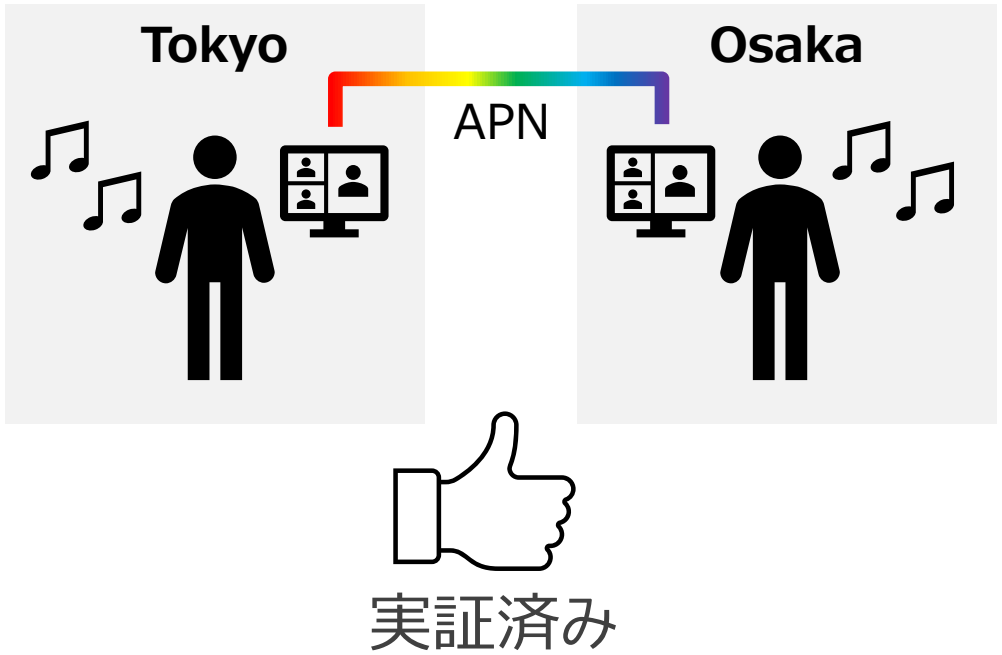
実証実験として東京・大阪間でリアルタイム遠隔合唱



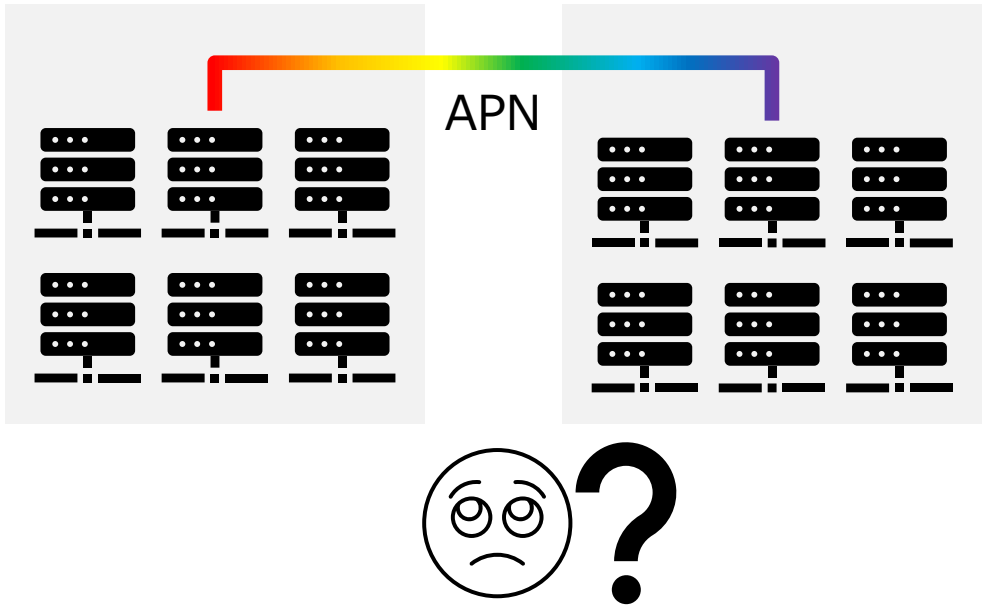
【NTT西日本】IOWN APN関連技術を用いたリアルタイム遠隔合唱実証実験を「サントリー1万人の第九」で実施 ～世界初、東京―大阪間をIOWN APN関連技術でつなぐ～｜ニュースリリース - 通信・ICTサービス・ソリューション より

新たに試したかったこと

計算機同士のコミュニケーションにもAPNを活用したい



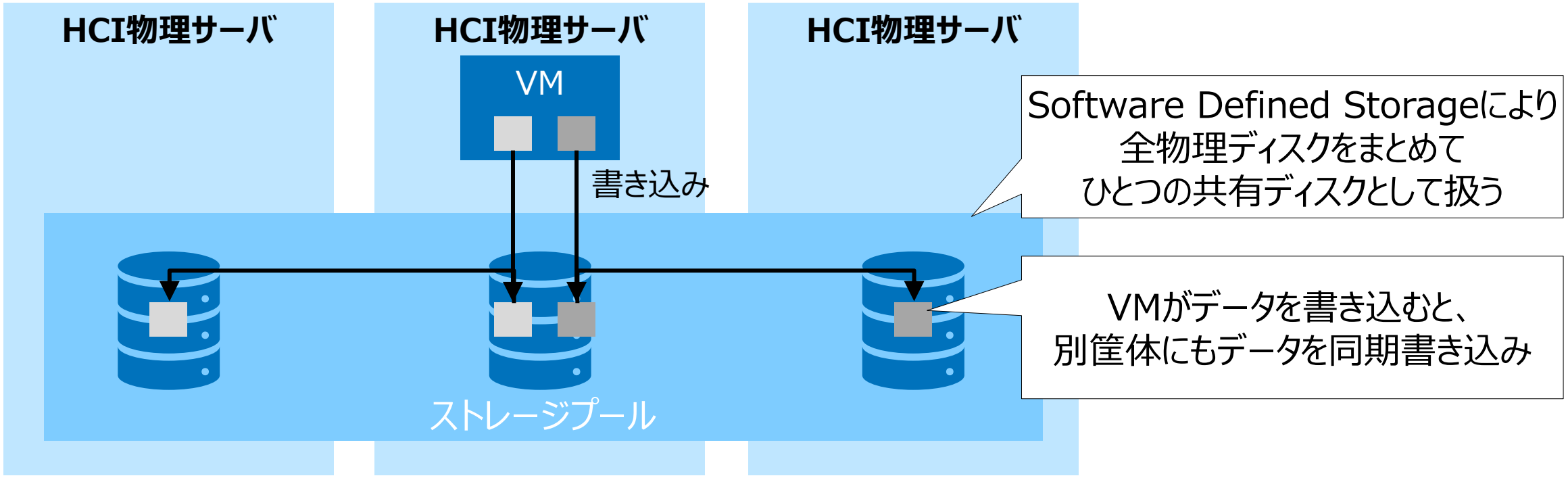
実証済み



今回検証

使用したサーバシステム

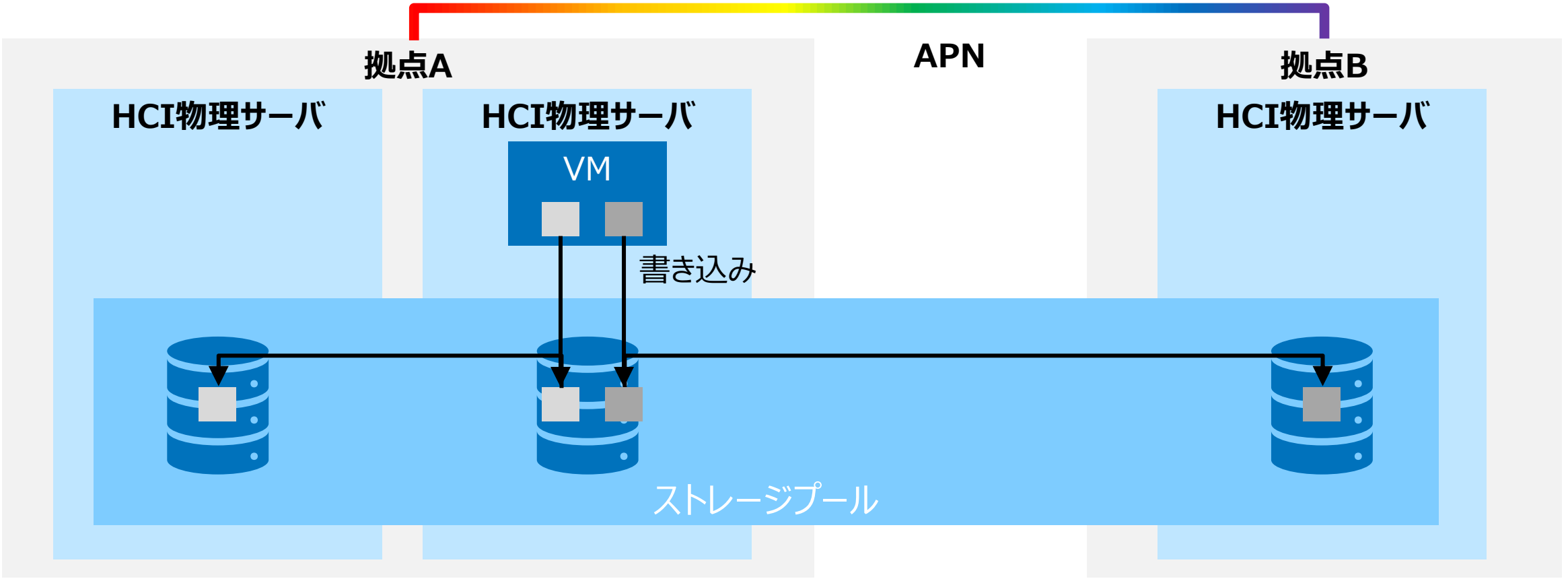
Hyper Converged Infrastructure（HCI）を使用



耐障害性・スケーラビリティが得られる反面、物理サーバ間のNW帯域・遅延が性能に影響

使用したサーバシステム

Hyper Converged Infrastructure（HCI）を使用



遠隔地の物理サーバをAPNで接続し、HCIクラスタとして使えるか検証

できると何がうれしいか

NTT西日本

距離を越えてサーバをクラスタリングし、VMなどを柔軟に配置することで・・・

可用性の向上

停止時に他のDCへ
フェイルオーバー

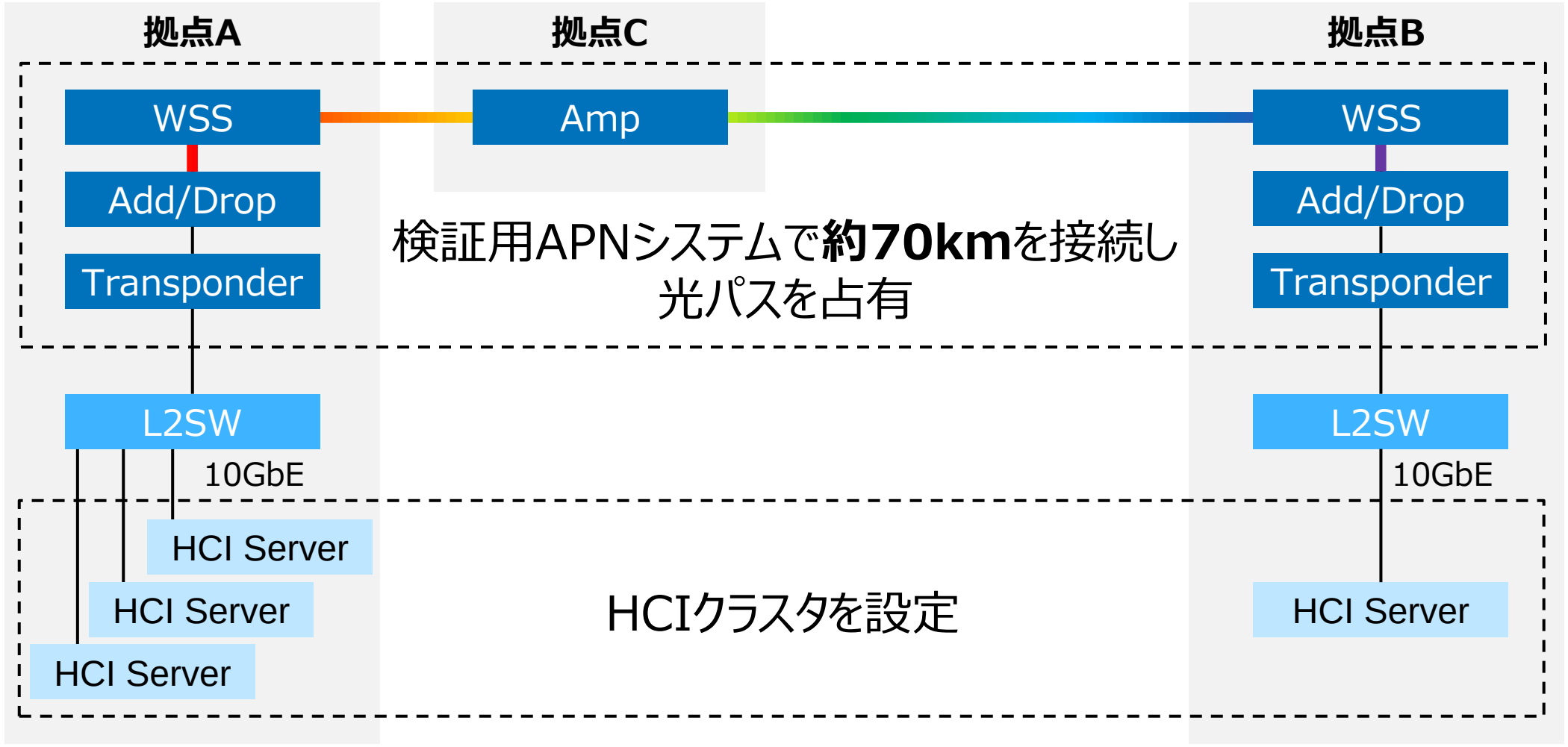
消費電力の低減

電力使用効率の高い
DCを動的に選択

施設のスケールアウト

小規模DCを組み合わせ
大きなワークロードを処理

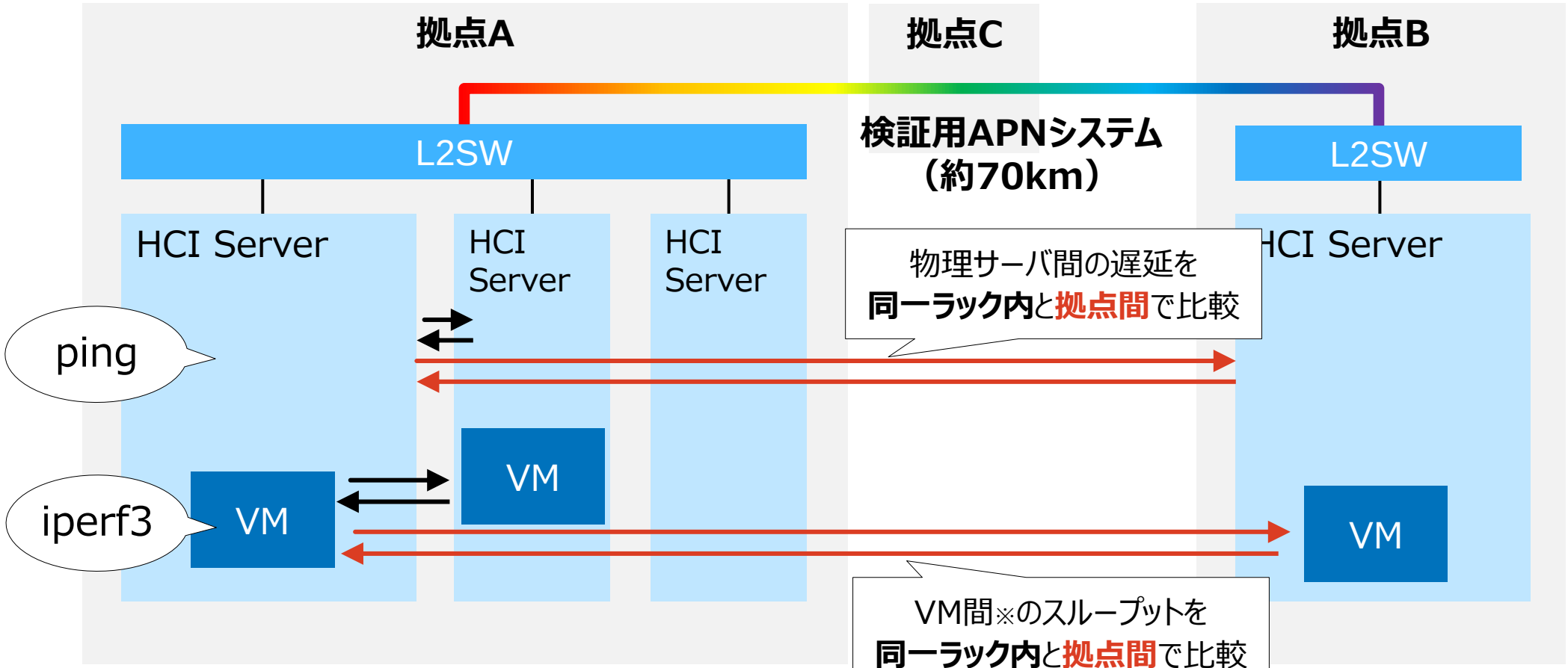
検証構成



WSS: Wavelength Selective Switch

検証内容：ネットワーク

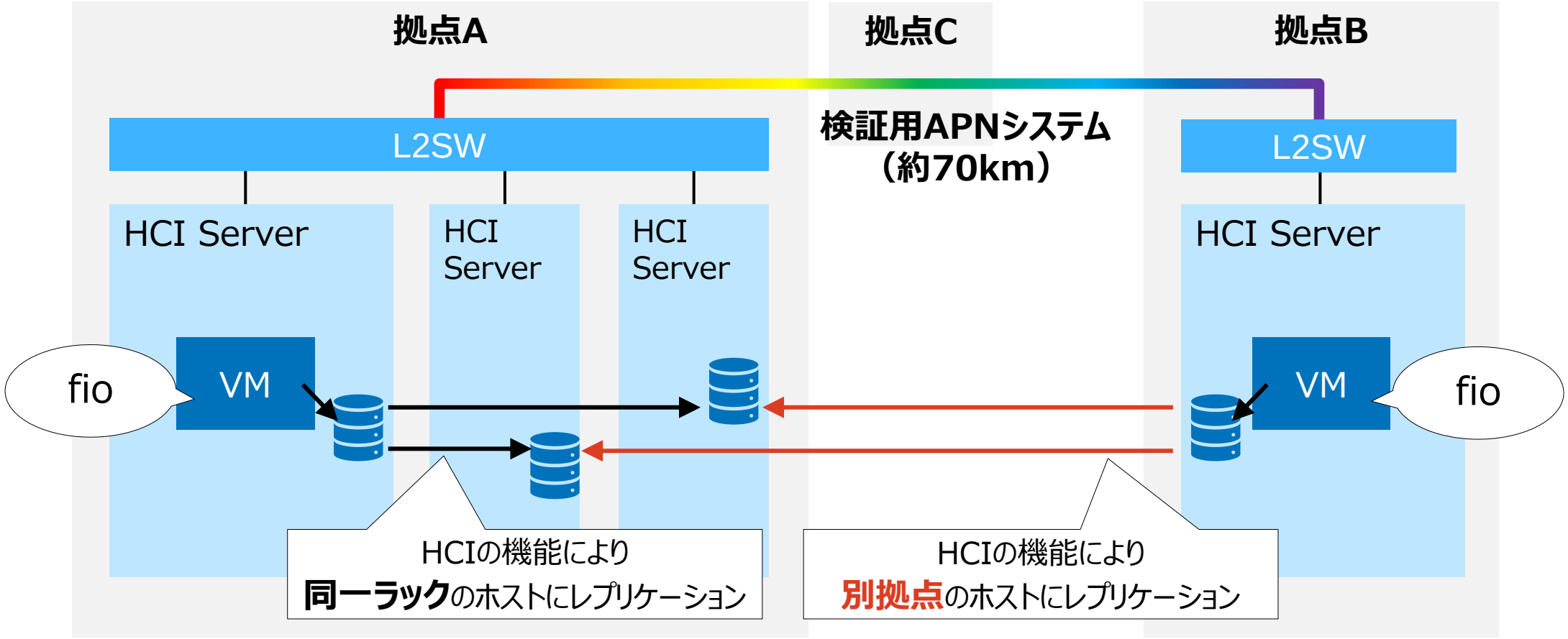
pingによるRTTとiperf3によるスループットを、同一ラック内／拠点間で比較



※HCIホストにソフトウェアインストールの制限があったためVMで実施

検証内容：ディスクI/O

fioによる書き込み性能を、HCIのレプリケーション同一ラック内／拠点間で比較



サーバ構成・実施コマンド

HCI物理サーバ	
CPU	24core (HT有効化で48)
メモリ	192GB
ディスク	1.92TB SATA-SSD ×2 4TB HDD × 4

VM	
CPU	4core
メモリ	8GB
ディスク	30GB

ネットワーク検証

```
ping -c 30  $server_addr
```

3回取得し平均

```
iperf3 -b 10000M -c $server_addr
```

5回取得し平均

ディスクI/O検証

```
fio --bs=${block_size}
   --direct=1
   --readwrite=randwrite
   --filename=${tmpfile}
   --gtod_reduce=1
   --iodepth=64
   --ioengine=libaio
   --name=fiotest
   --output=${logfilename}.log
   --size=16G
```

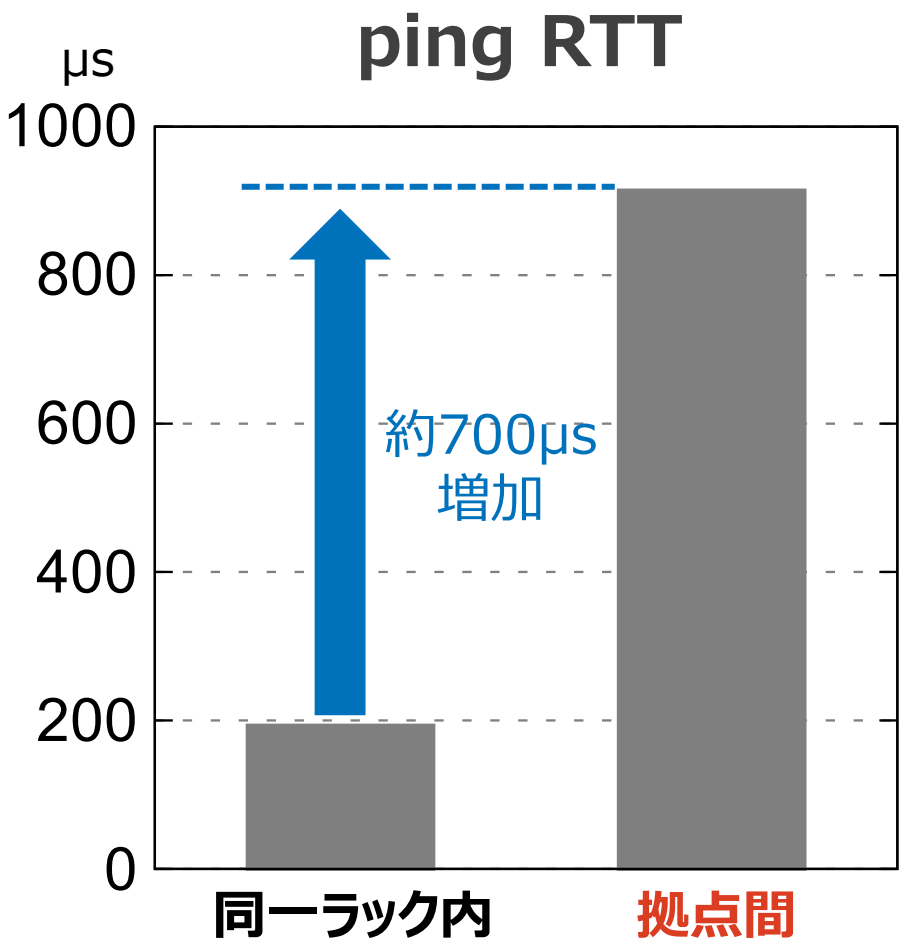
書き込み頻度による比較のため
ブロックサイズ4KBと1024KBの
両方を測定

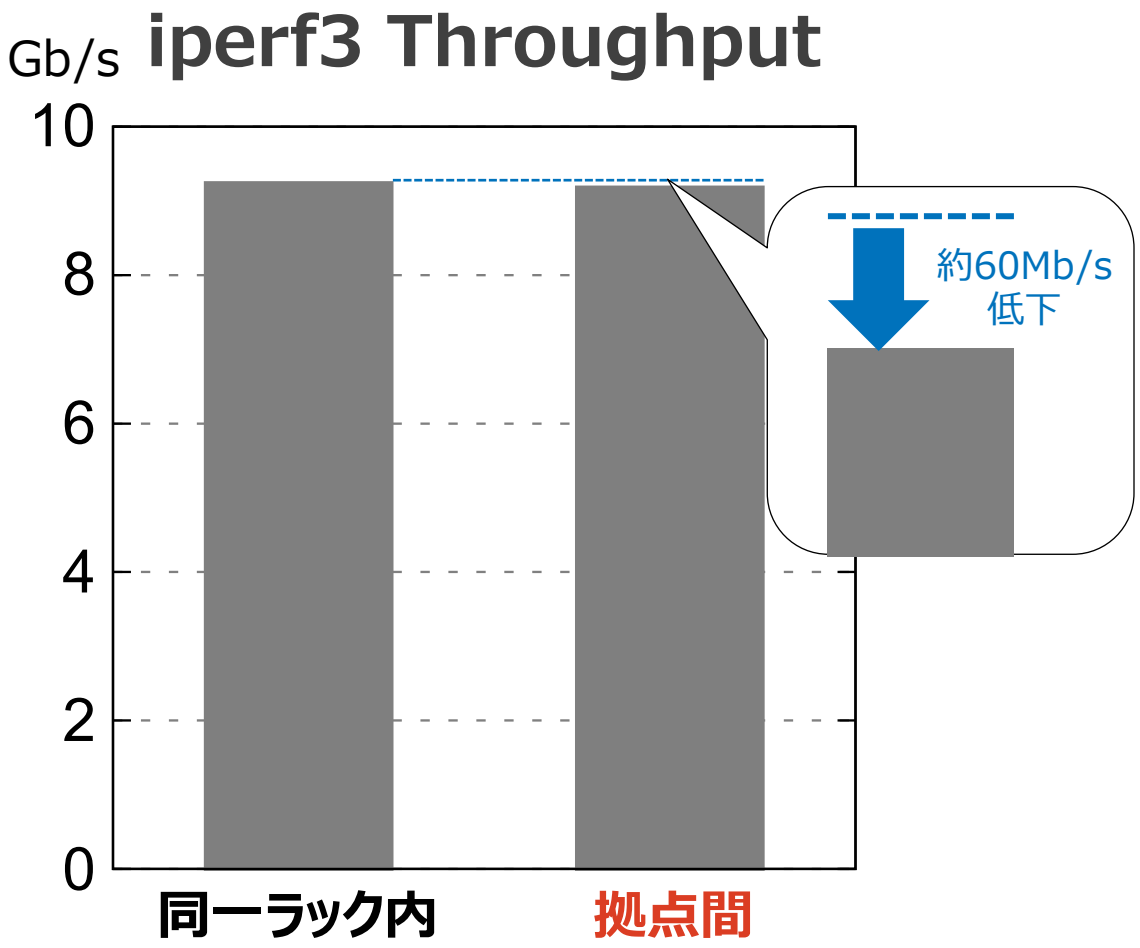
性能影響を顕在化させるために
ディスク同期書き込み・
ランダム書き込みで実行

5回取得し平均

検証結果：ネットワーク

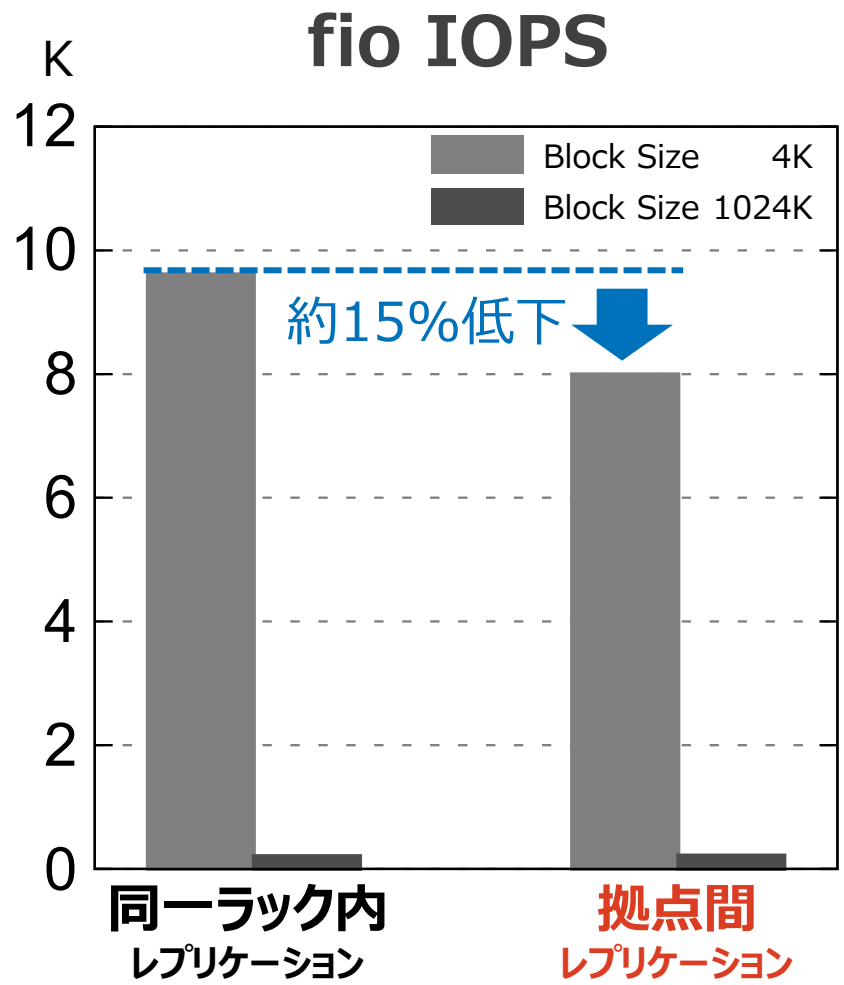
伝送区間が加わることで約700μsの遅延増加。スループットはほぼ同等を維持

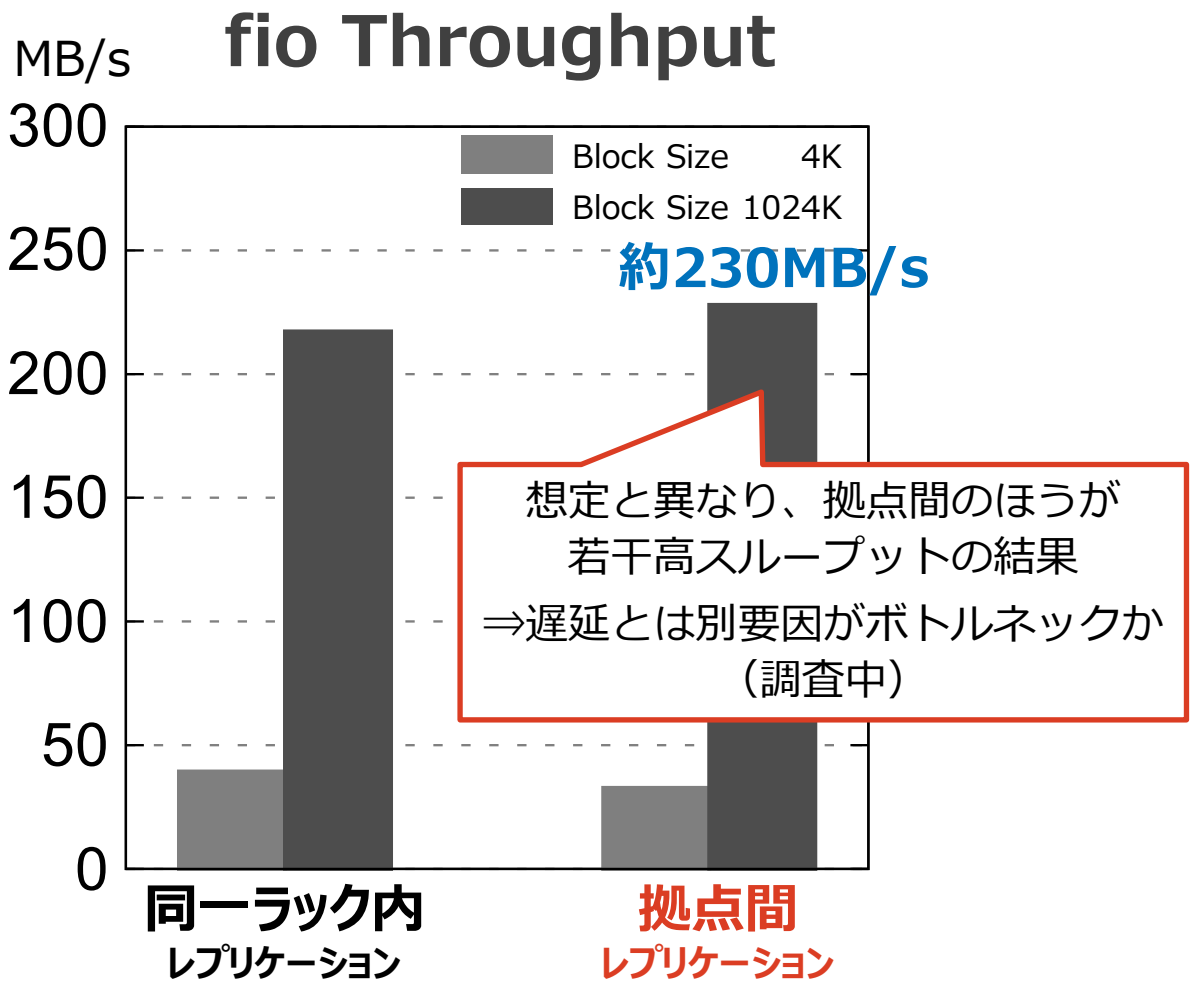




検証結果：ディスクI/O

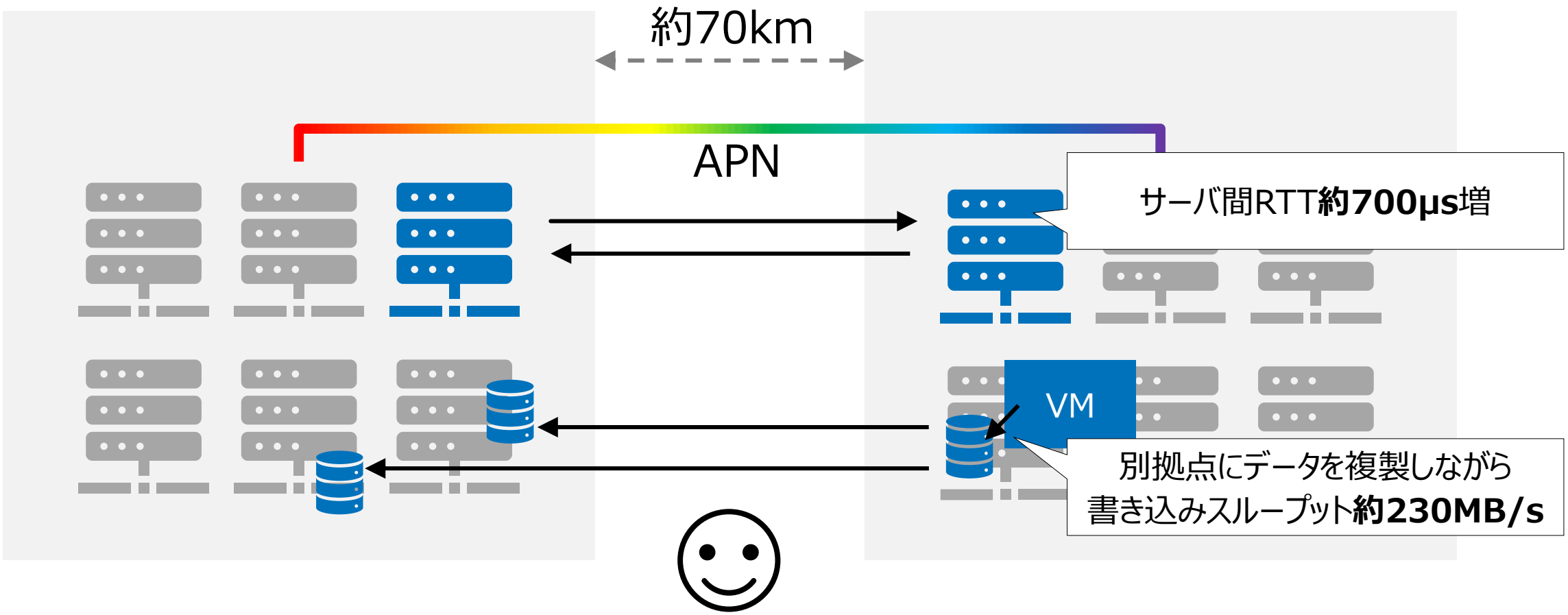
ブロックサイズ4KBでIOPSが約15%低下するが、1MBでスループット約230MB/s





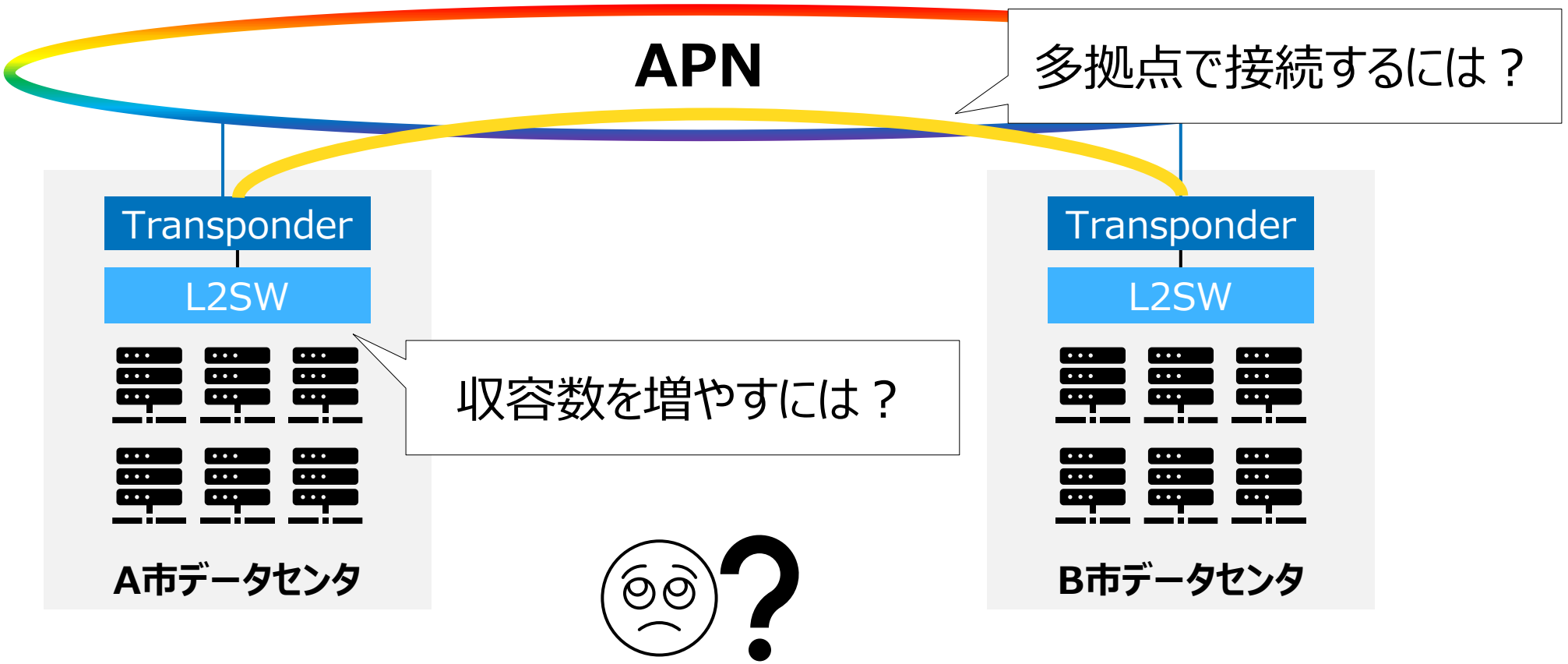
検証結果まとめ

ワークロードによっては、APNで接続した複数拠点のサーバ群をクラスタとして扱えそう



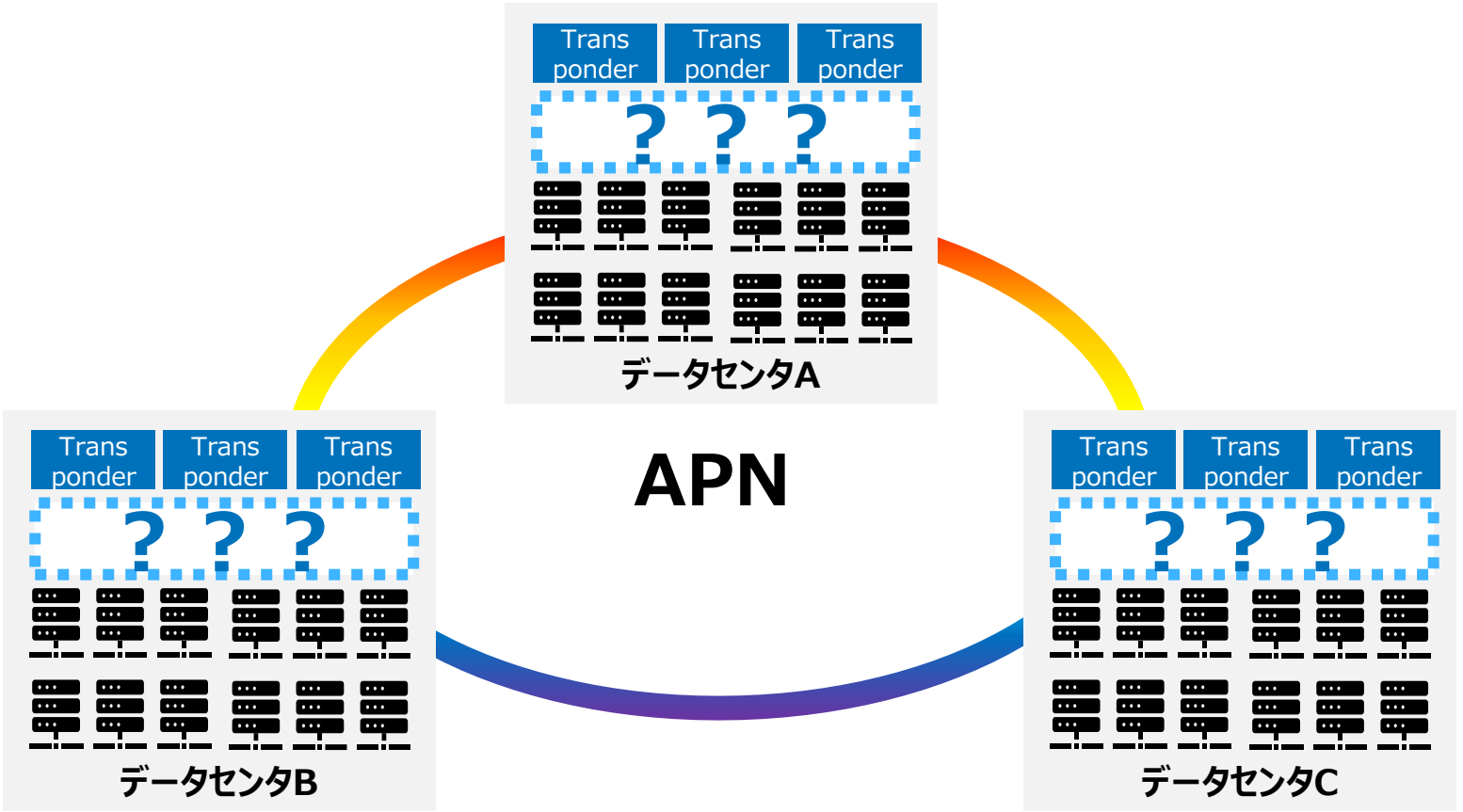
検証構成の課題

APNの光パスを１対１の伝送路として使用しているだけで、拡張性がない



検証構成の課題

柔軟性・拡張性を高められるよう、APNとIPが連動するネットワークを検討した



アジェンダ

NTT西日本

All-Photonics NetworkでDC間を接続した
サーバクラスタの簡易性能検証

**All-Photonics Networkをバックボーンにした
EVPN Multi-Site Architectureの検証**

自己紹介

芳　政信
ヨシ　マサノブ

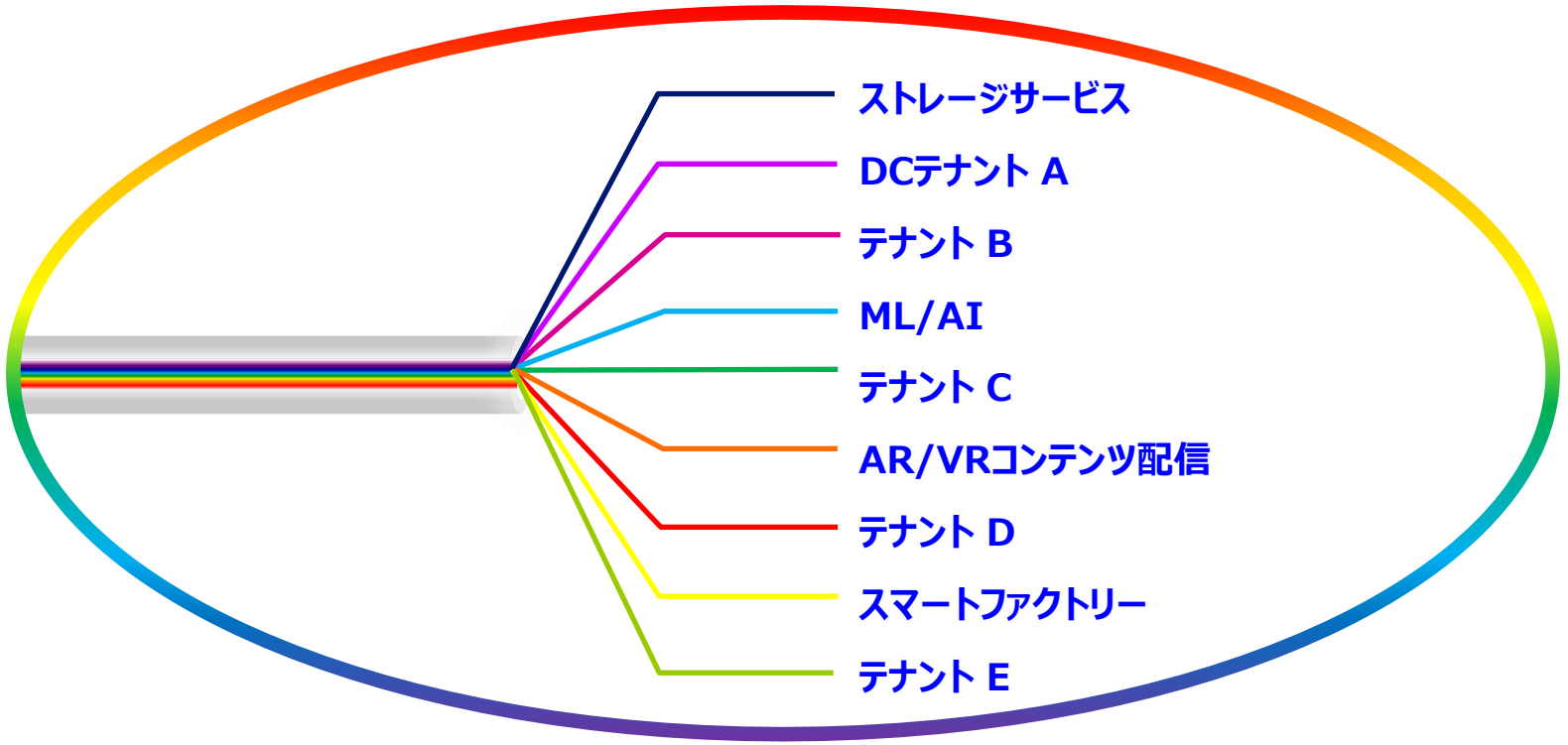
- 造園土木職人からネットワークエンジニアへ
- 企業VPNの詳細設計、社内NW運用管理、ネットワークインストラクタなど
- Model Driven Programmability,
  Network Programmability, DevNet

データセンター相互接続（DCI）

DCIでは距離や帯域容量が課題となっており、光波長も帯域の束として扱われる

波長にもっと個性を！！

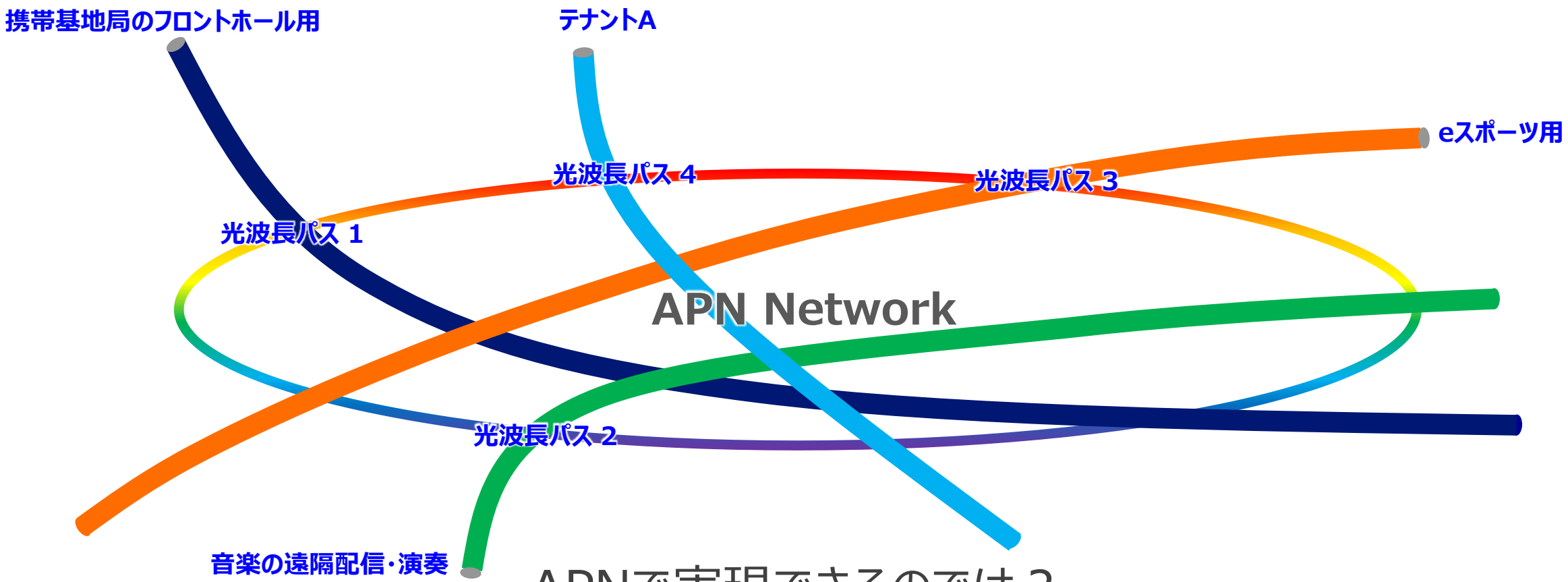
波長に個性を与えることで、波長多重にもっと価値を持たせることができるかも



うまく組み合わせればDCIを含めた仮想DC群をテナントごとに提供できたりとか

All-Photonics Network ?

APNは端末・ユーザ・サービスごとに、光パスを波長単位で提供するネットワーク



APNで実現できるのでは？

NWテクノロジーと波長の統合

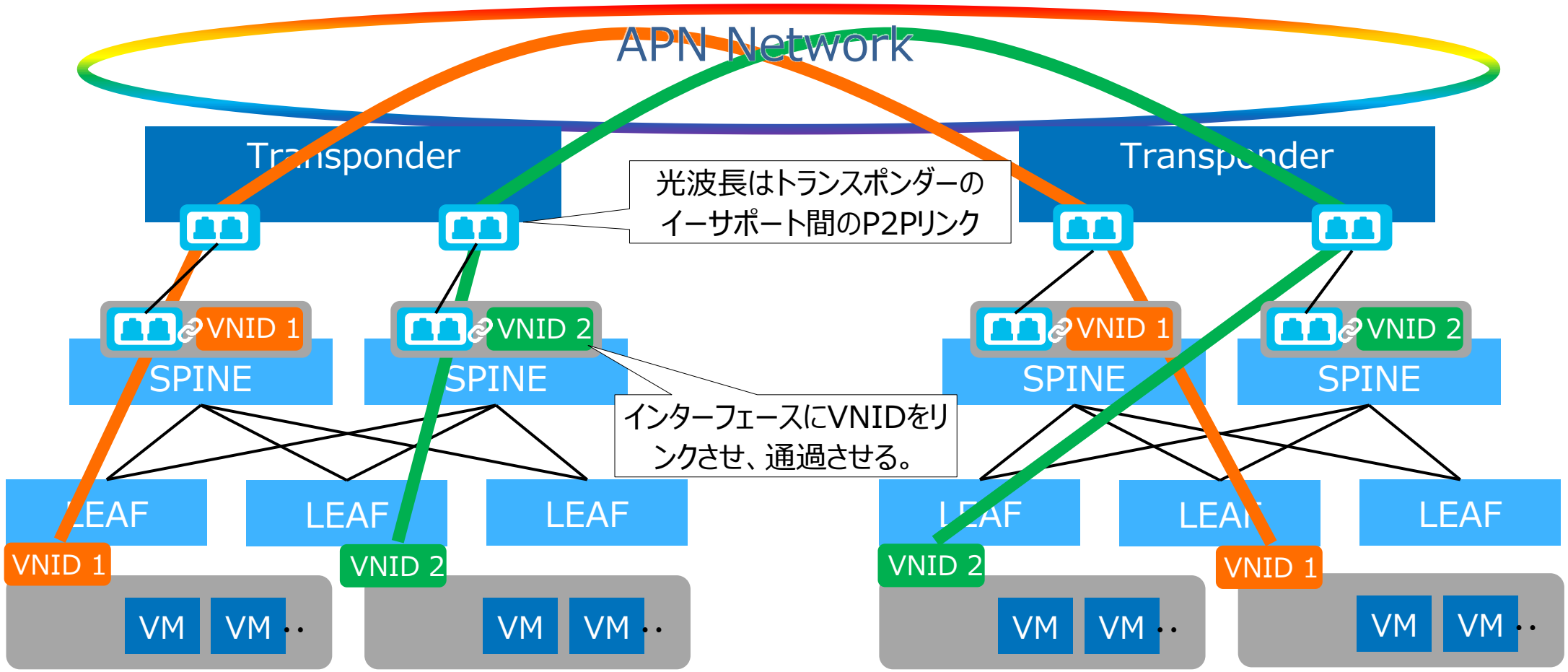
エンド-エンドの波長化はまだ困難なので、IP-NWとAPNを統合して実現する

APNの波長に、NWテクノロジーのIDを組み合わせて波長に個性（ID）を与える

波長と組み合わせるのはVXLANのVNIが最適

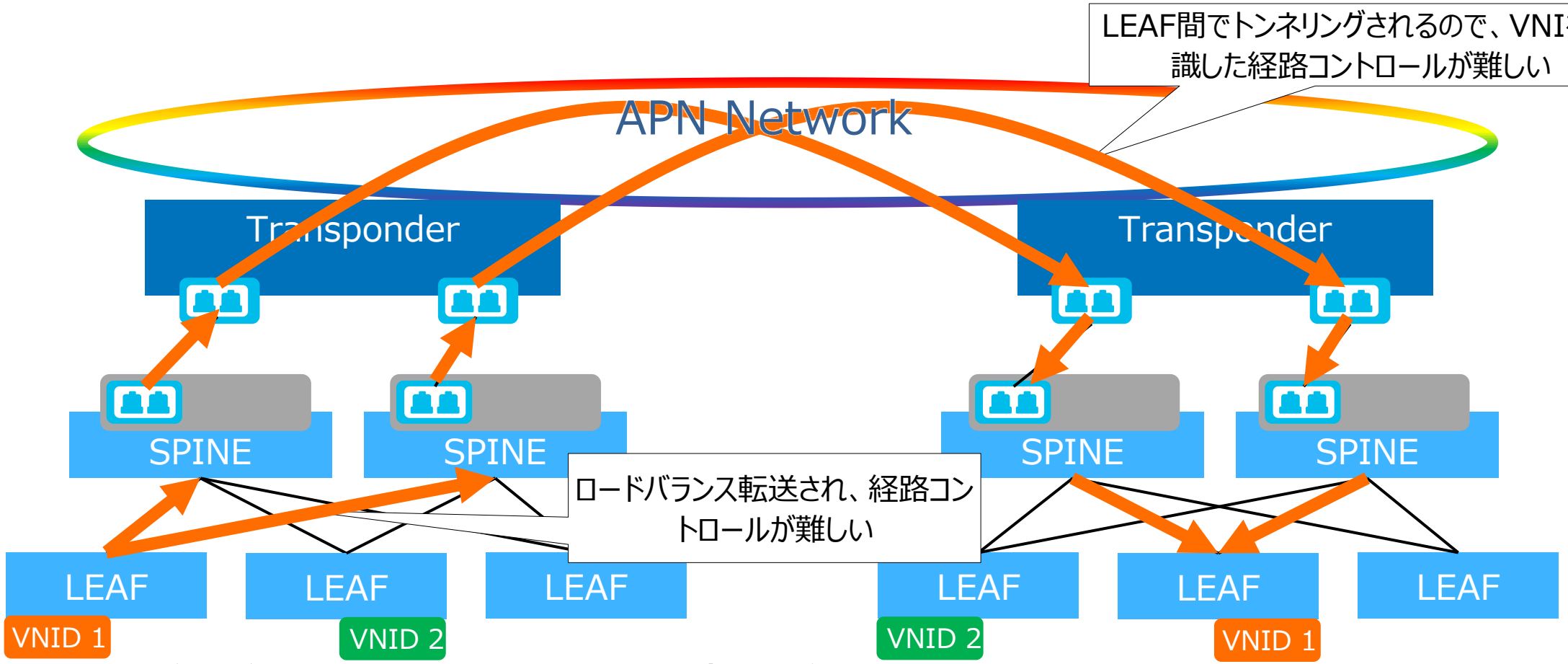
VXLANのVNIと波長を組み合わせる

VNID(L2, L3)をDCスイッチの物理インターフェースとリンクさせる必要がある



VXLANファブリックのデメリット

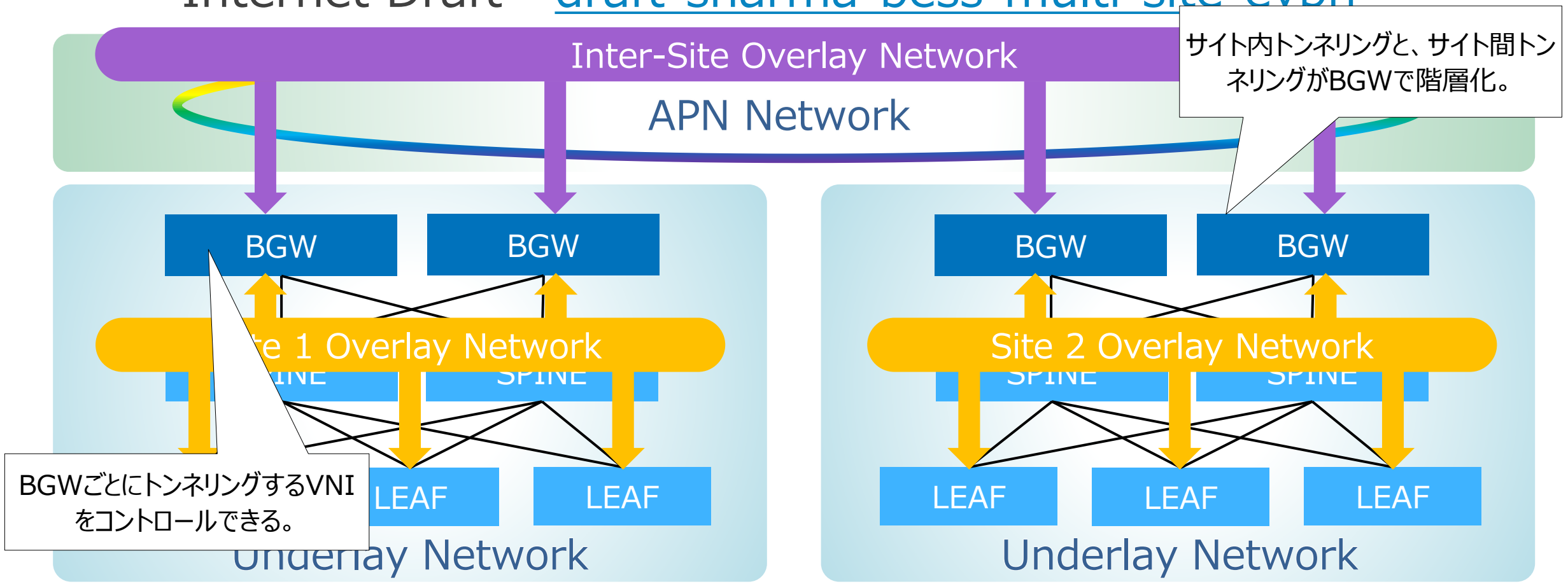
VXLANの特徴がデメリットに



経路や通過するノードの積極的なコントロールには不向き

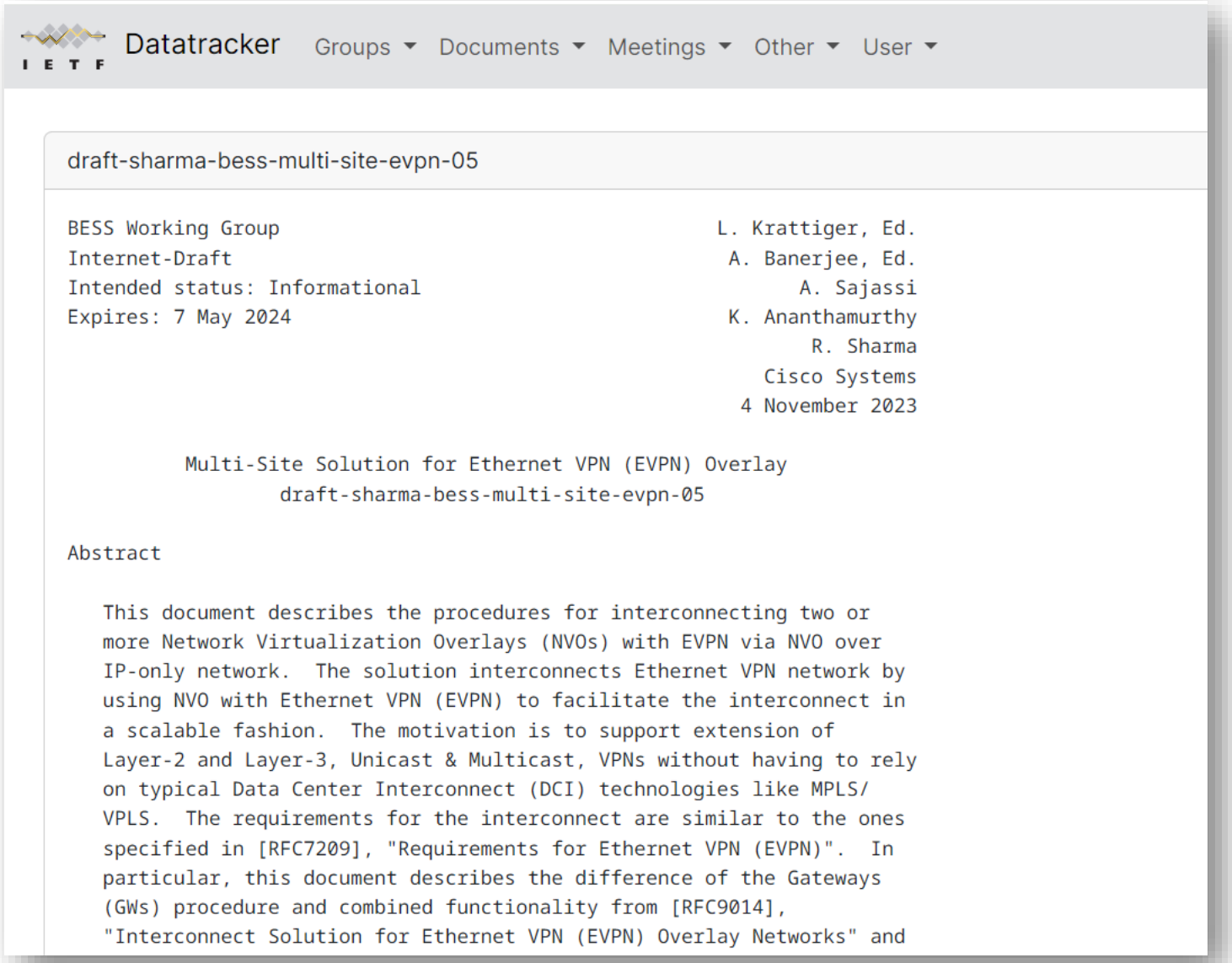
Multi-site Architectureに着目した

Internet Draft “draft-sharma-bess-multi-site-evpn”



BGWで階層化されるので、その境界でコントロールが可能

draft-sharma-bess-multi-site-evpn



Datatracker Groups Documents Meetings Other User

draft-sharma-bess-multi-site-evpn-05

BESS Working Group — L. Krattiger, Ed.
Internet-Draft — A. Banerjee, Ed.
Intended status: Informational — A. Sajassi
Expires: 7 May 2024 — K. Ananthamurthy
R. Sharma
Cisco Systems
4 November 2023

Multi-Site Solution for Ethernet VPN (EVPN) Overlay
draft-sharma-bess-multi-site-evpn-05

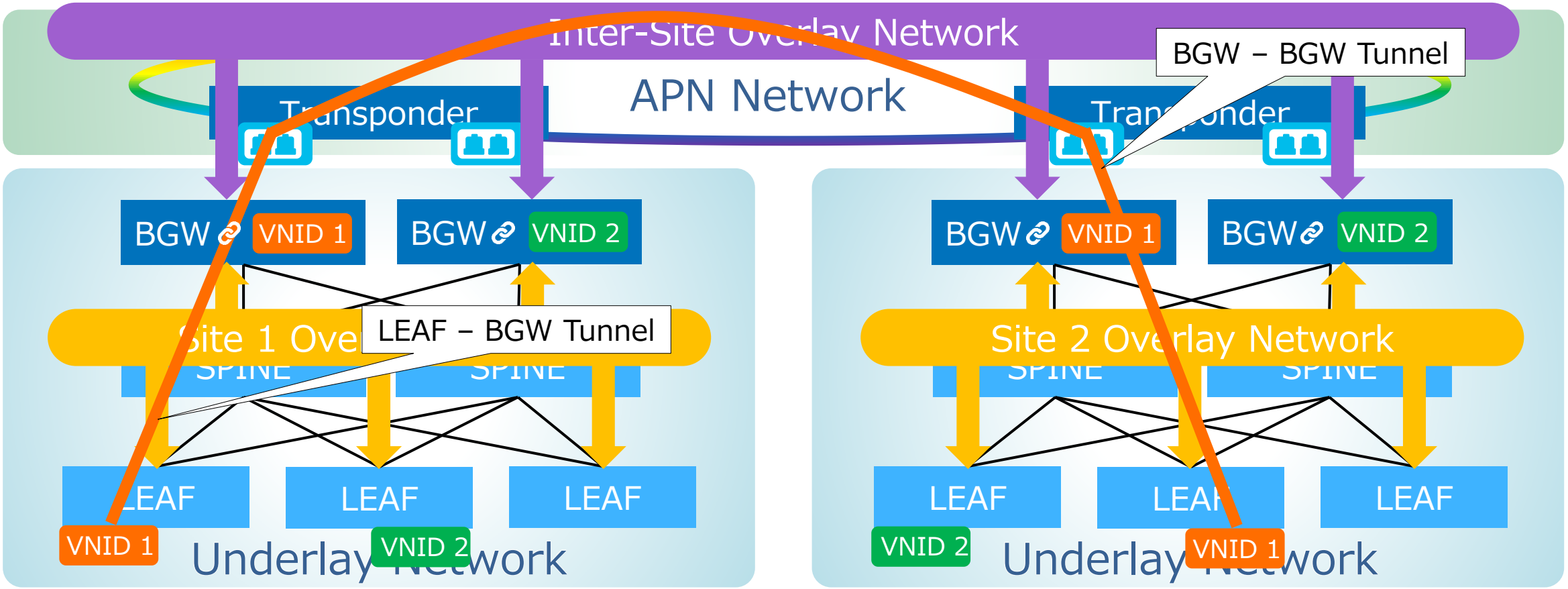
Abstract

This document describes the procedures for interconnecting two or more Network Virtualization Overlays (NVOs) with EVPN via NVO over IP-only network. The solution interconnects Ethernet VPN network by using NVO with Ethernet VPN (EVPN) to facilitate the interconnect in a scalable fashion. The motivation is to support extension of Layer-2 and Layer-3, Unicast & Multicast, VPNs without having to rely on typical Data Center Interconnect (DCI) technologies like MPLS/ VPLS. The requirements for the interconnect are similar to the ones specified in [RFC7209], "Requirements for Ethernet VPN (EVPN)". In particular, this document describes the difference of the Gateways (GWs) procedure and combined functionality from [RFC9014], "Interconnect Solution for Ethernet VPN (EVPN) Overlay Networks" and

Multi-site Architectureと波長の組合わせ

NTT西日本

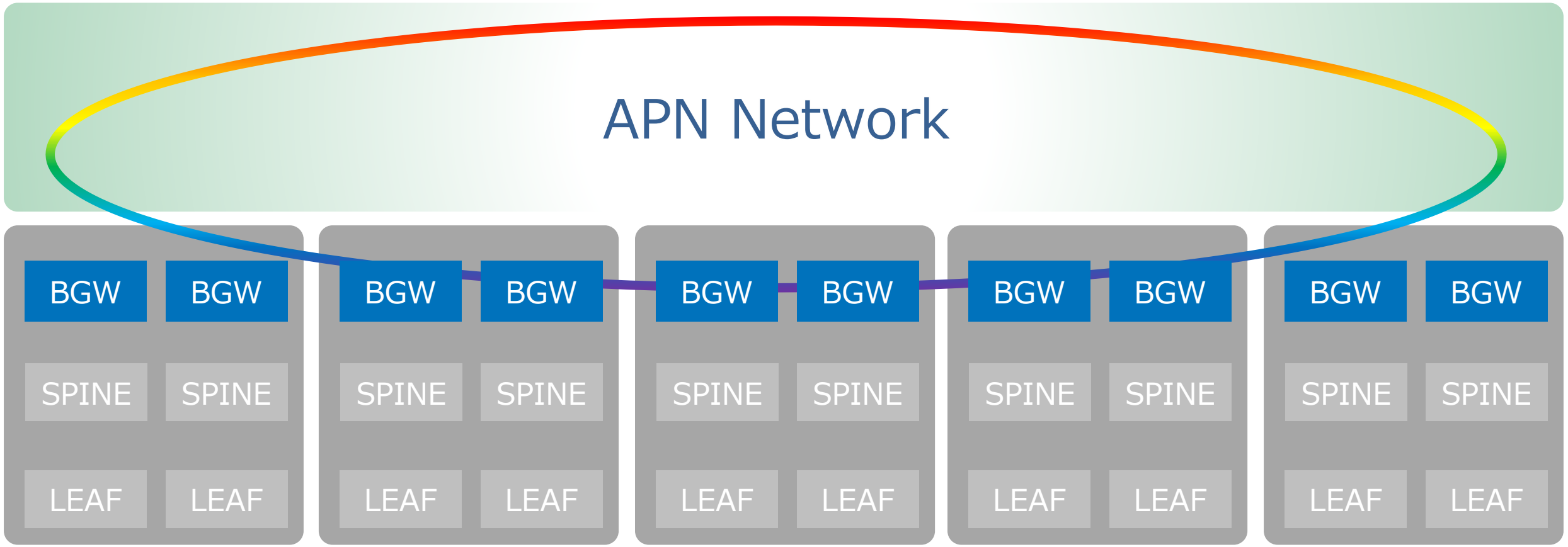
BGWで特定のVNIDをインポートすることで、その先の波長にIDを与える



DCIを仮想化し、テナントごとのHCIクラスタを完全に分離して構成できる

リング光パストポロジーでの光パス

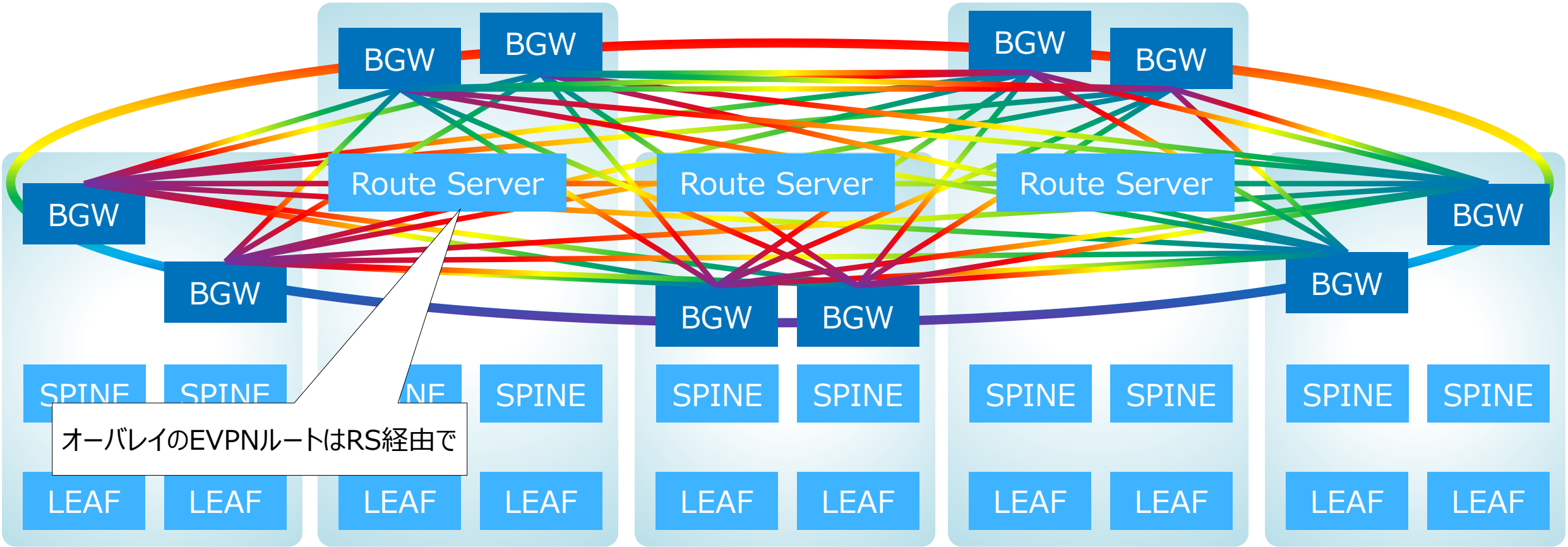
リング型パストポロジーでBGW間の光パスはどう設定するか



フルメッシュ？　ハブ&スポーク？

APNリング/DCI Full-mesh構成

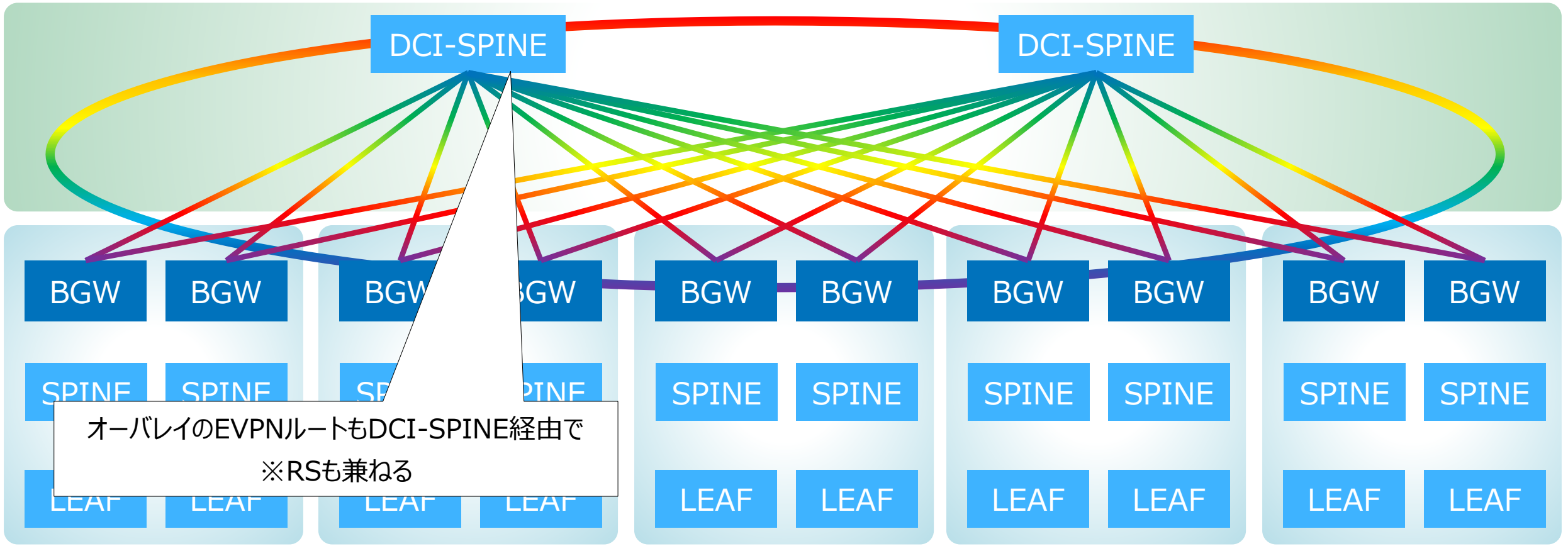
異なる方路で各サイトのBGWをフルメッシュで接続する（高性能）



BGW-BGWで直接接続するので低遅延だが、必要波長数が膨大になる

APNリング/DCI CLOS構成

異なる方路に配置されたDCI-SPINE経由で接続する（低コスト）



必要波長数を節約できるが、DCI-SPINE通過とO/E, E/O変換の遅延が乗る

検証しました

▷ テナントごとのMACアドレスラーニング

- エンドホストのMACアドレス、IPアドレスをEVPNでどう学習するか。

▷ ブリッジング

- 学習したMACアドレス向けのトラフィックをどうブリッジングするか。

▷ 伝送区間DOWN時の切り替わり性能

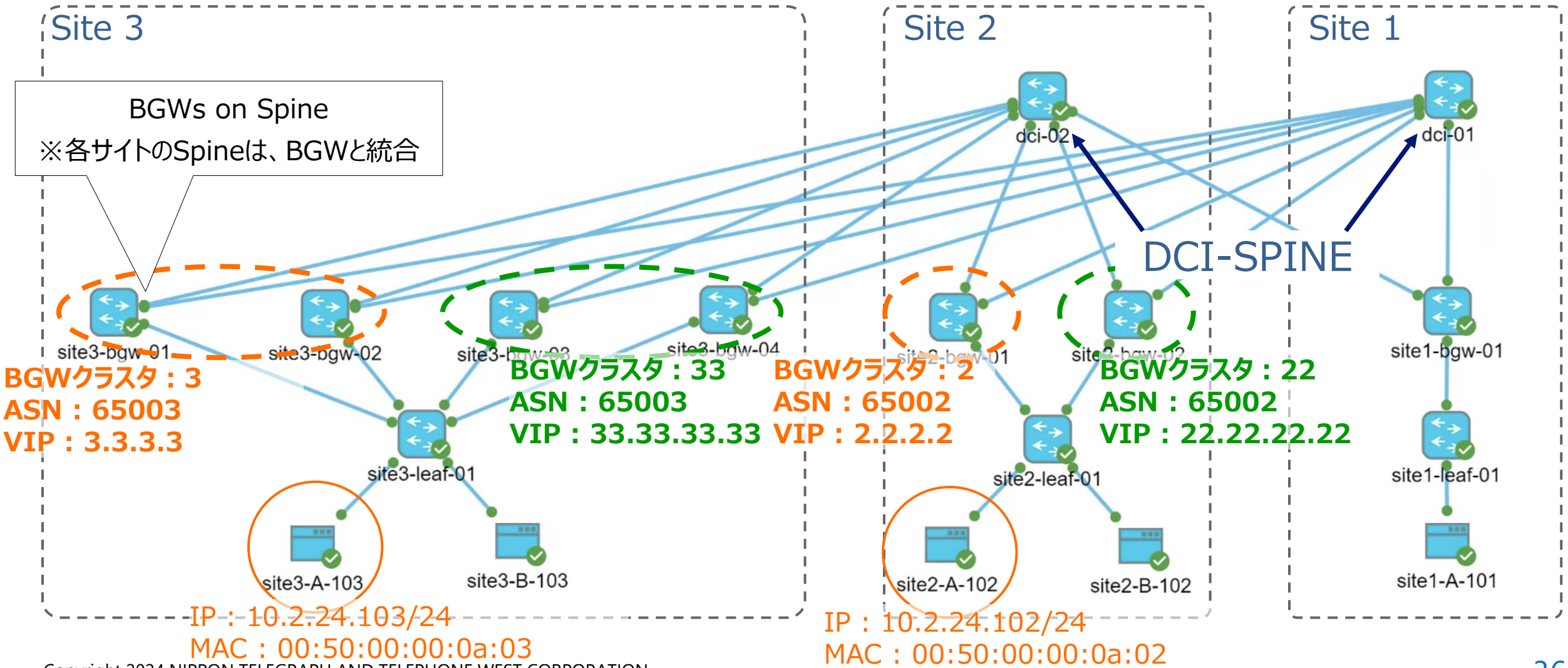
- 伝送区間に障害が起きた場合の切り替わり時間

▷ レイテンシー

- DCI CLOS、DCI Full-meshのレイテンシー

検証構成

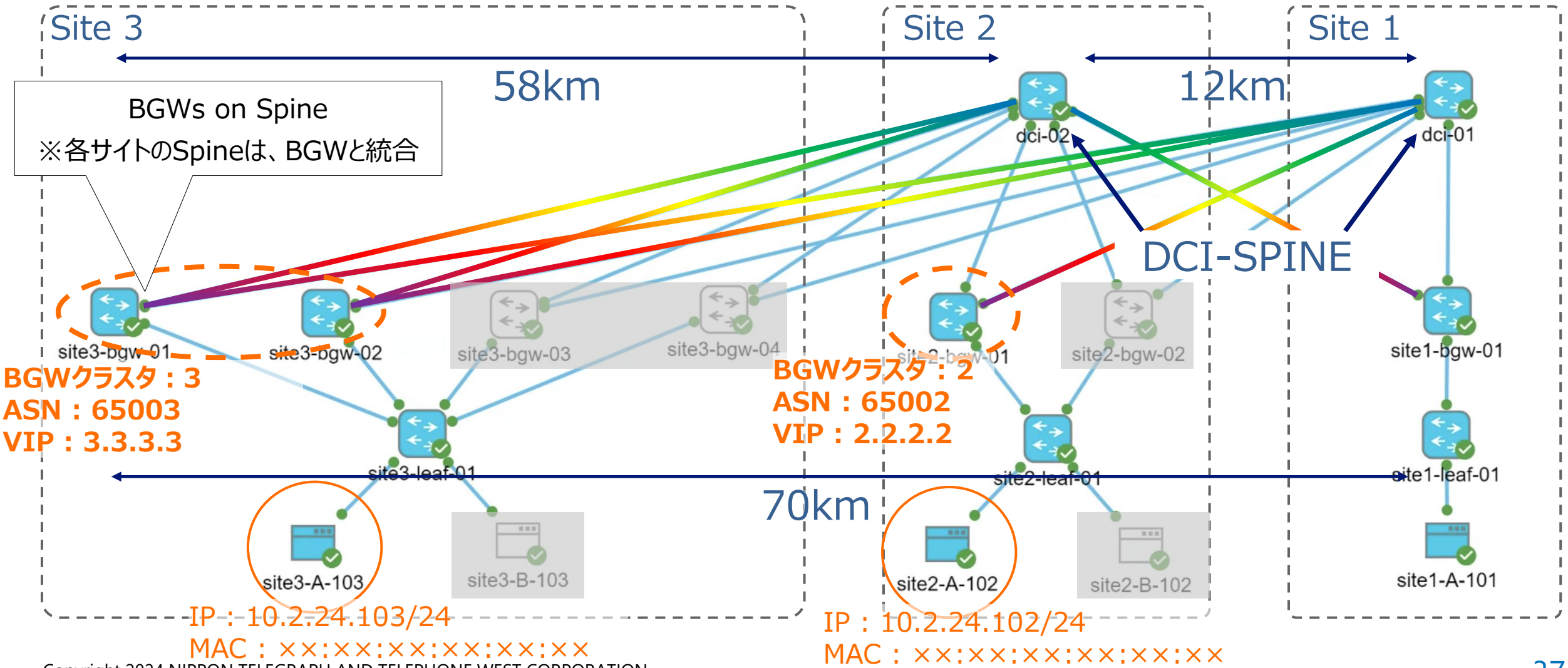
シミュレーション環境



検証構成

物理環境

全リンク　100G



検証しました

▷ テナントごとのMACアドレスラーニング

- エンドホストのMACアドレス、IPアドレスをEVPNでどう学習するか。

▷ ブリッジング

- 学習したMACアドレス向けのトラフィックをどうブリッジングするか。

▷ 伝送区間DOWN時の切り替わり性能

- 伝送区間に障害が起きた場合の切り替わり時間

▷ レイテンシー

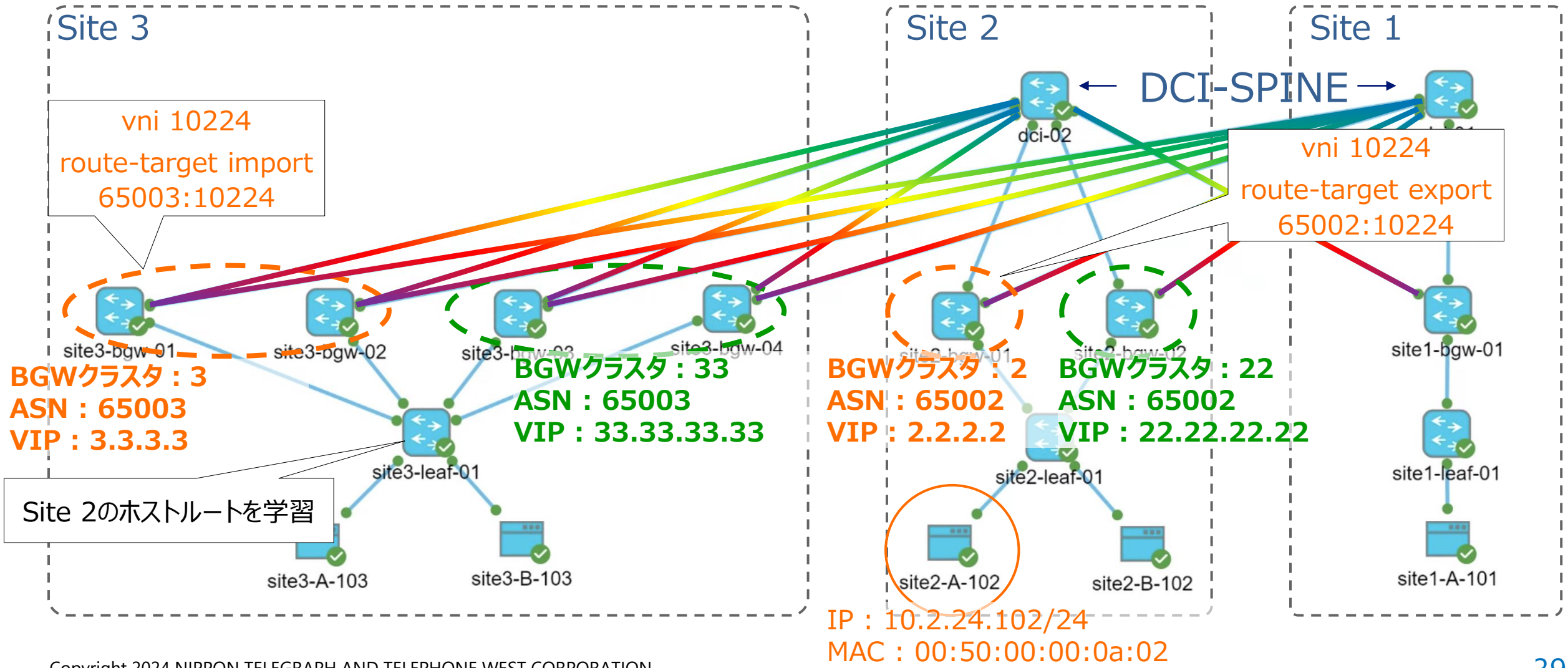
- DCI CLOS、DCI Full-meshのレイテンシー

テナントごとのMACアドレスラーニング

NTT西日本

ホストのMACアドレスとIPアドレスの学習

テナントA
テナントB



テナントごとのMACアドレスラーニング

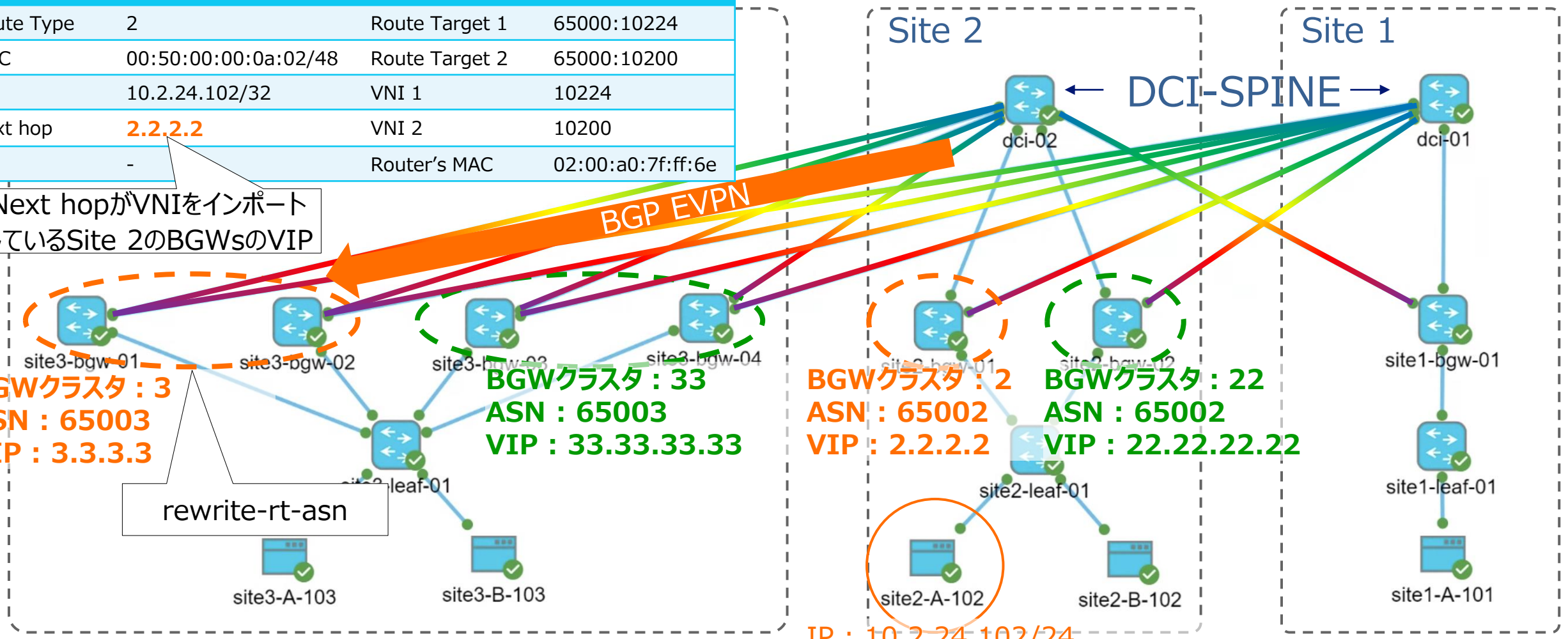
NTT西日本

DCI-SPINEからBGWへののアップデート

テナントA
テナントB

EVPN NLRI			
Route Type	2	Route Target 1	65000:10224
MAC	00:50:00:00:0a:02/48	Route Target 2	65000:10200
IP	10.2.24.102/32	VNI 1	10224
Next hop	2.2.2.2	VNI 2	10200
-	-	Router's MAC	02:00:a0:7f:ff:6e

Next hopがVNIをインポートしているSite 2のBGWsのVIP



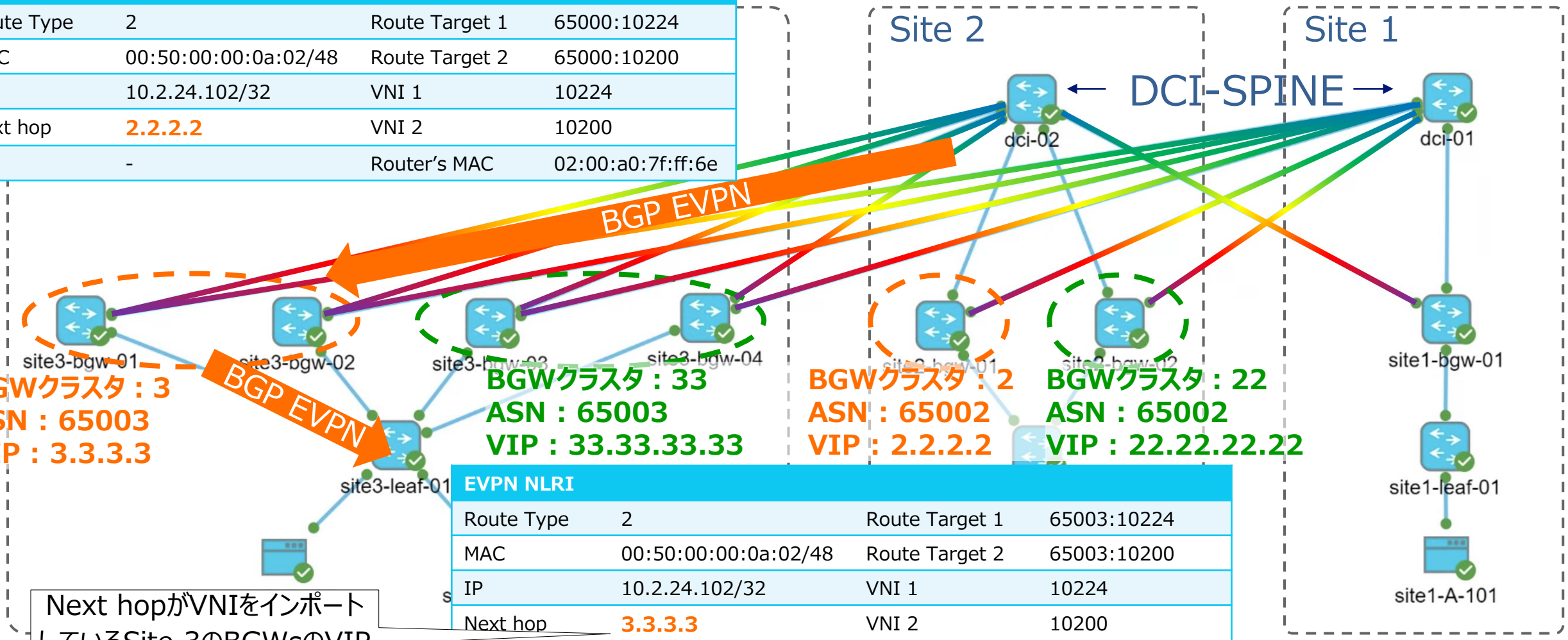
テナントごとのMACアドレスラーニング

NTT西日本

BGWからLEAFへのアップデート

テナントA
テナントB

EVPN NLRI			
Route Type	2	Route Target 1	65000:10224
MAC	00:50:00:00:0a:02/48	Route Target 2	65000:10200
IP	10.2.24.102/32	VNI 1	10224
Next hop	2.2.2.2	VNI 2	10200
-	-	Router's MAC	02:00:a0:7f:ff:6e



EVPN NLRI			
Route Type	2	Route Target 1	65003:10224
MAC	00:50:00:00:0a:02/48	Route Target 2	65003:10200
IP	10.2.24.102/32	VNI 1	10224
Next hop	3.3.3.3	VNI 2	10200
-	-	Router's MAC	02:00:a0:7f:ff:78

Next hopがVNIをインポートしているSite 3のBGWsのVIP

検証しました

▷ テナントごとのMACアドレスラーニング

- エンドホストのMACアドレス、IPアドレスをEVPNでどう学習するか。

▷ ブリッジング

- 学習したMACアドレス向けのトラフィックをどうブリッジングするか。

▷ 伝送区間DOWN時の切り替わり性能

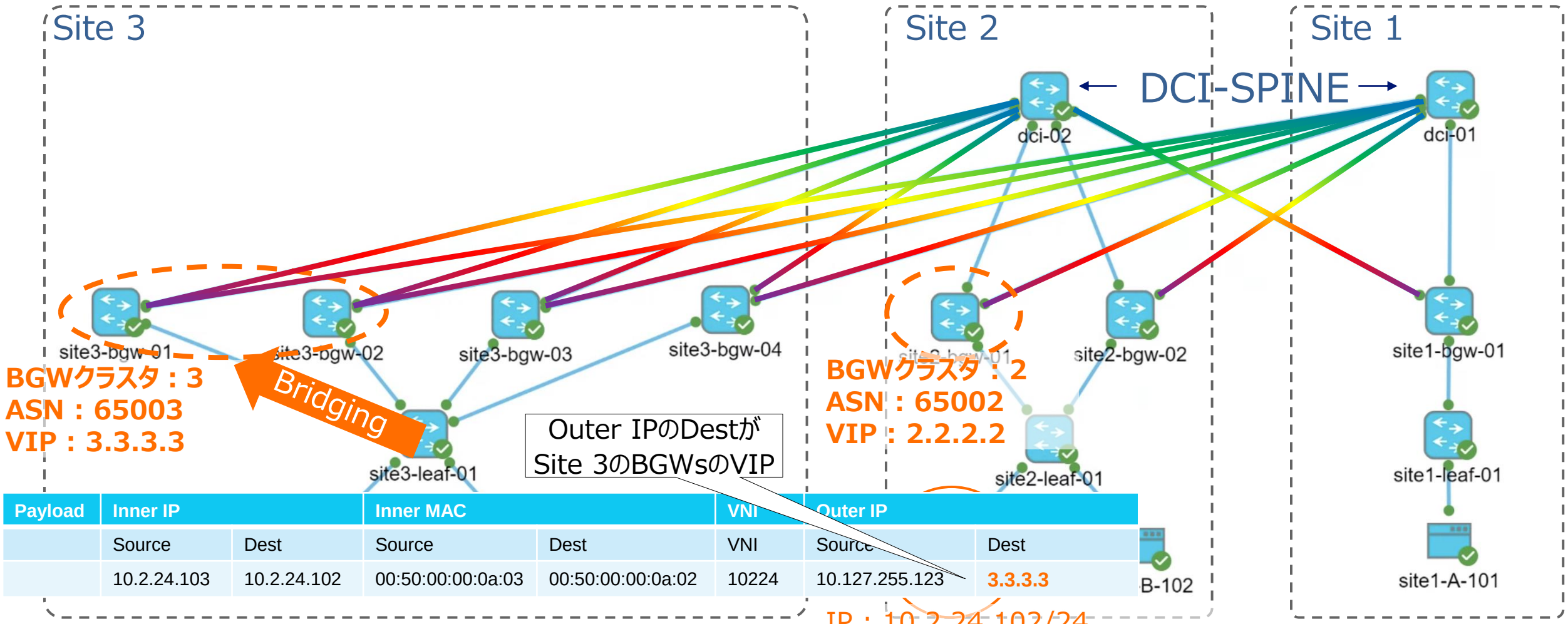
- 伝送区間に障害が起きた場合の切り替わり時間

▷ レイテンシー

- DCI CLOS、DCI Full-meshのレイテンシー

ブリッジング

ローカルLeaf → ローカルBGW

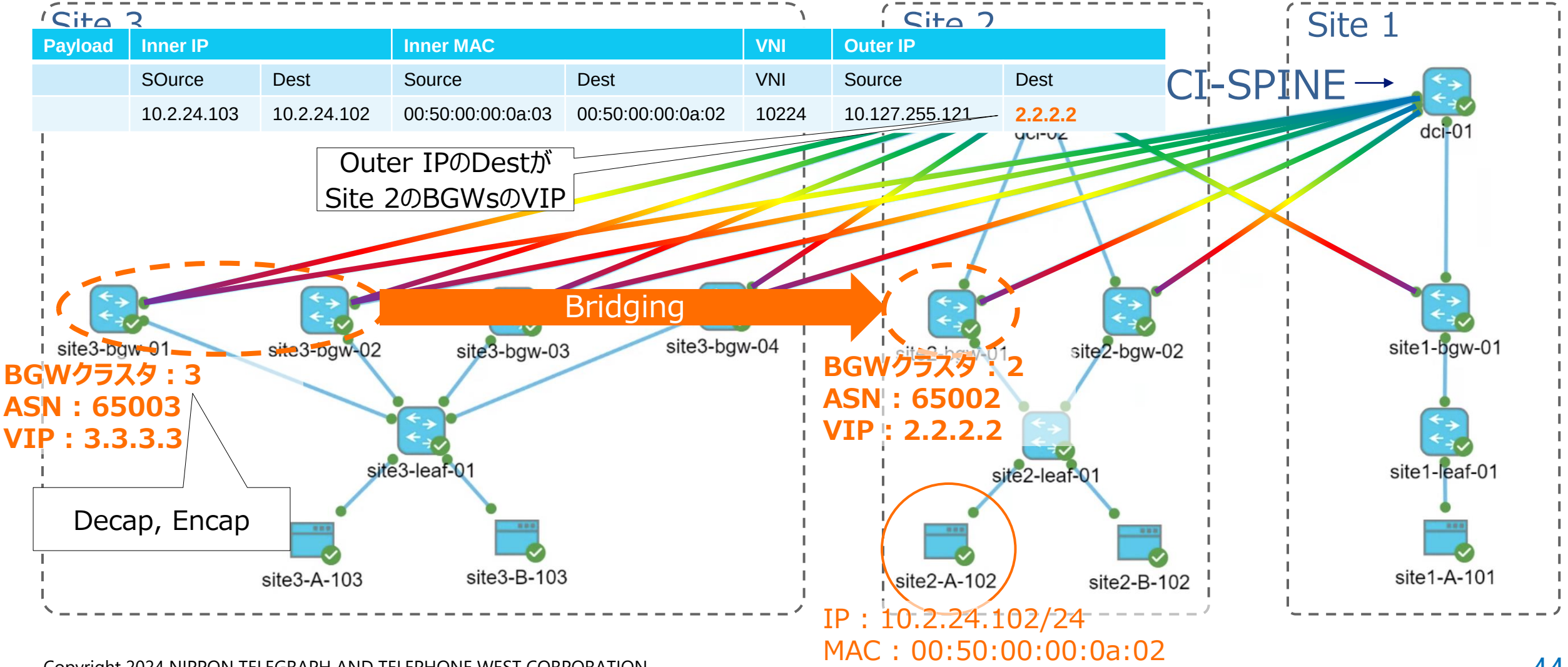


Payload	Inner IP		Inner MAC		VNI	Outer IP	
	Source	Dest	Source	Dest	VNI	Source	Dest
	10.2.24.103	10.2.24.102	00:50:00:00:0a:03	00:50:00:00:0a:02	10224	10.127.255.123	3.3.3.3

ブリッジング

ローカルBGW → リモートBGW

Payload	Inner IP		Inner MAC		VNI	Outer IP	
	SOurce	Dest	Source	Dest	VNI	Source	Dest
	10.2.24.103	10.2.24.102	00:50:00:00:0a:03	00:50:00:00:0a:02	10224	10.127.255.121	2.2.2.2



検証しました

▷ テナントごとのMACアドレスラーニング

- エンドホストのMACアドレス、IPアドレスをEVPNでどう学習するか。

▷ ブリッジング

- 学習したMACアドレス向けのトラフィックをどうブリッジングするか。

▷ 伝送区間DOWN時の切り替わり性能

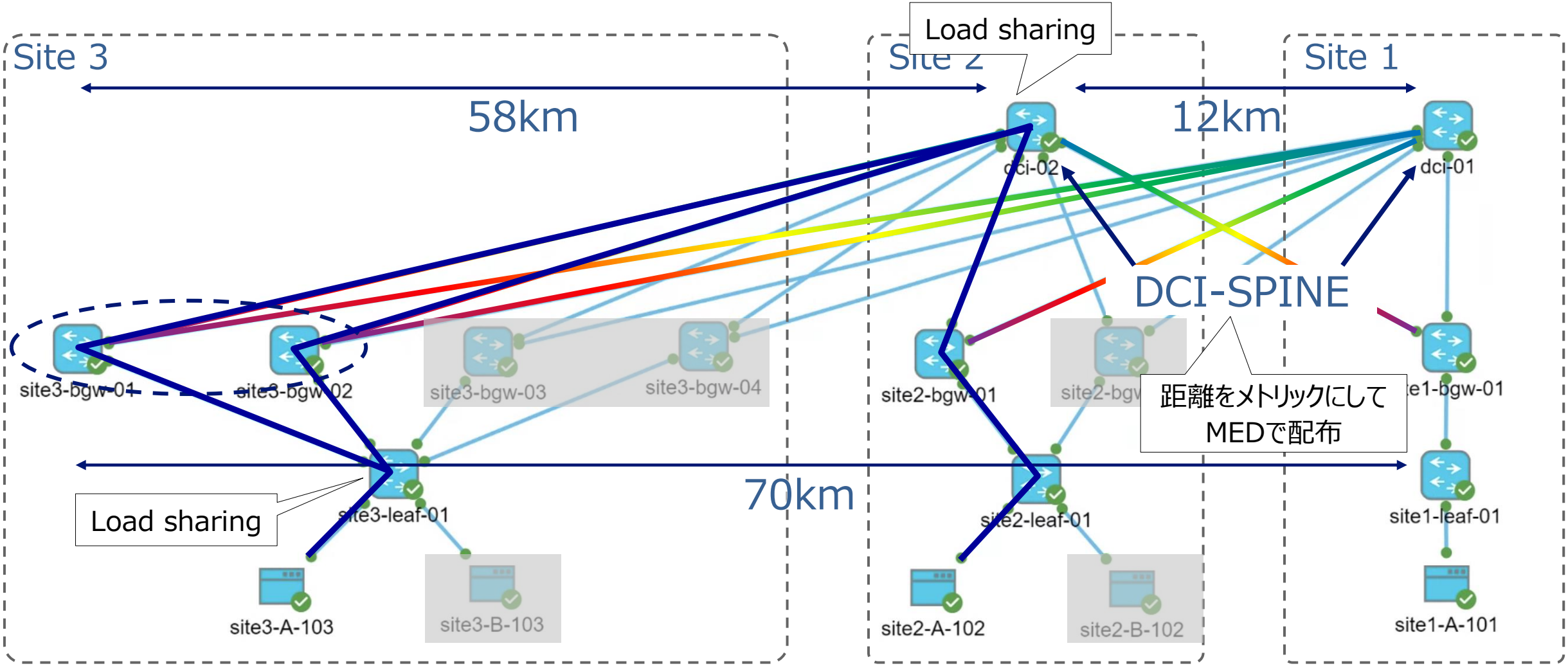
- 伝送区間に障害が起きた場合の切り替わり時間

▷ レイテンシー

- DCI CLOS、DCI Full-meshのレイテンシー

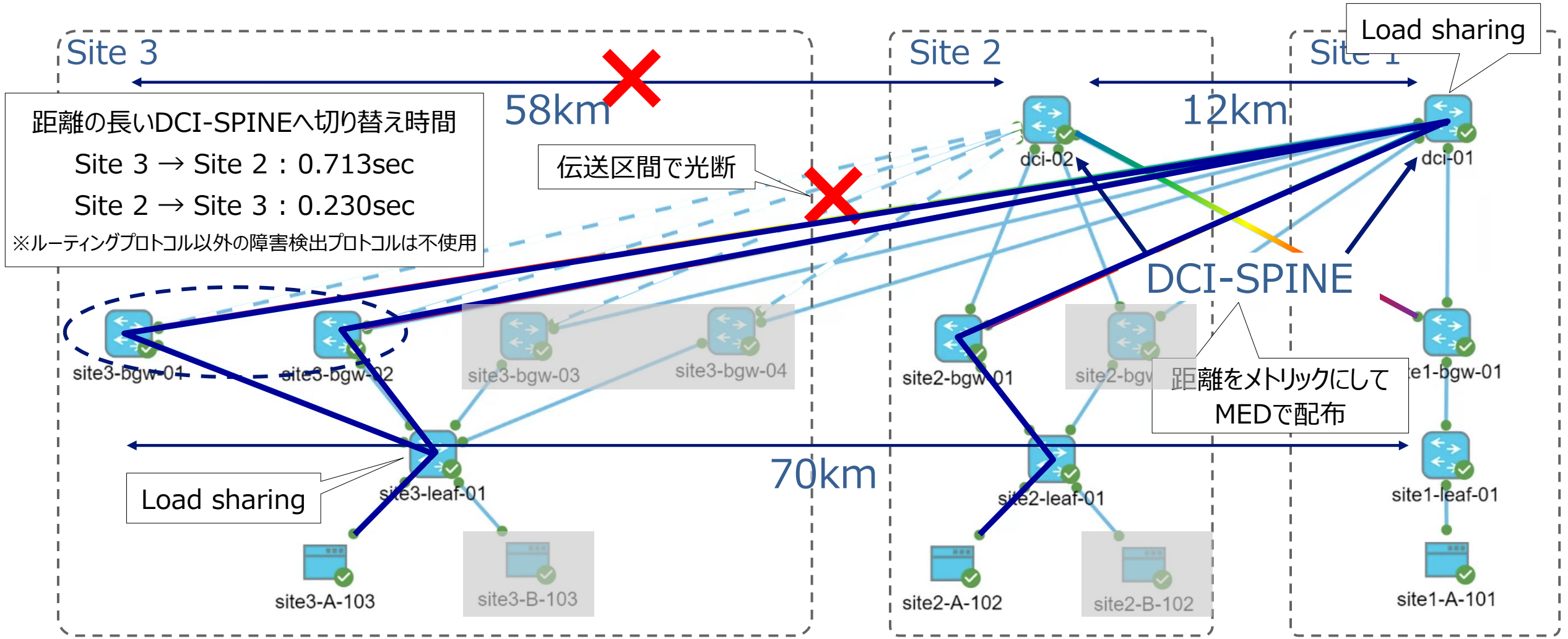
伝送区間DOWN時の切り替わり

DCI CLOSでの転送ルート



伝送区間DOWN時の切り替わり

DCI CLOSでの伝送区間down時の切り替え経路と時間



検証しました

▷ テナントごとのMACアドレスラーニング

- エンドホストのMACアドレス、IPアドレスをEVPNでどう学習するか。

▷ ブリッジング

- 学習したMACアドレス向けのトラフィックをどうブリッジングするか。

▷ 伝送区間DOWN時の切り替わり性能

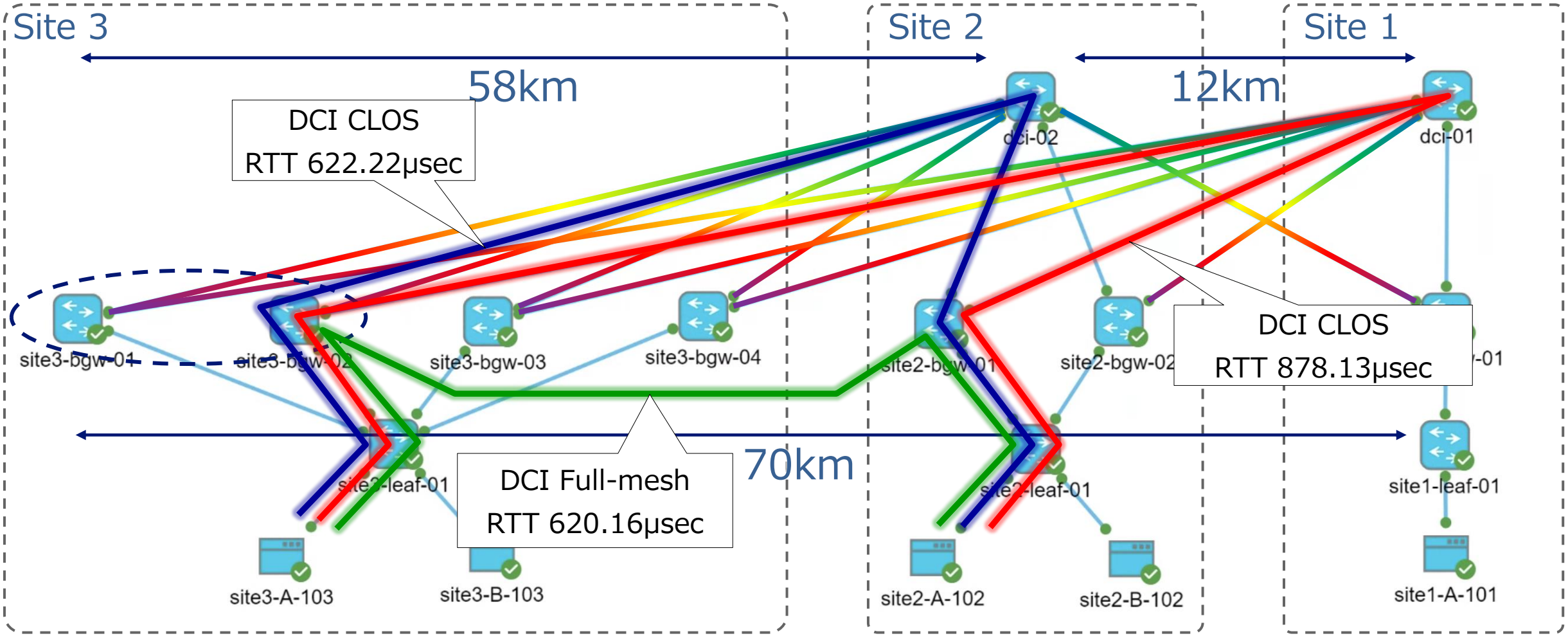
- 伝送区間に障害が起きた場合の切り替わり時間

▷ レイテンシー

- DCI CLOS、DCI Full-meshのレイテンシー

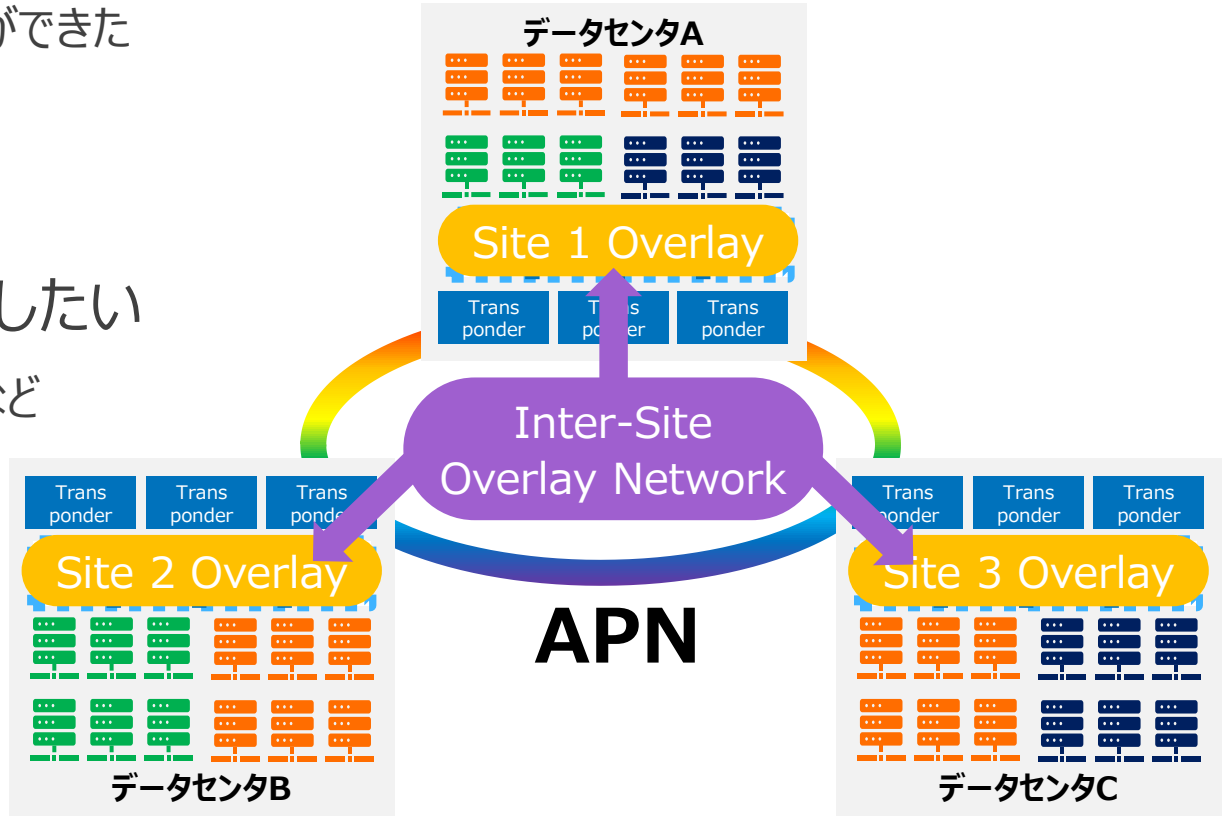
レイテンシー

各ルートでのRTT



まとめ

- EVPN Multi-site Architectureを活用し、APNとIPを統合するネットワークを検討した
  - 性能・コストを考慮し２パターンのトポロジーについて動作検証
  - 現状はいずれも期待通り機能し、基礎値を取得することができた
- DC間でのAPN活用に拡張性・柔軟性を付加
  - テナントごとに波長やHCIクラスタを分離しての提供など
- 今後はさらなるデータセンタの価値向上にトライしたい
  - 波長・コンピューティングを含むE2Eのリソースコントロールなど



議論したいこと

- 今回の発表の潜在的なニーズについてご意見ください！！
- 他のテクノロジーを活用した光波長の有効活用について
  ご意見ください！！
- DC間延伸のIP層どうやってますか？