バックボーンネットワークエンジニアが L3プロトコルを使ってL2リングを作り変えた話

～ついでに400G-ZRでWDMレスで長距離飛ばしちゃったよ～

2024年7月3日
JANOG54 Meeting in NARA

株式会社コミュニティネットワークセンター
大日方 周太
神谷 健太

今回共有、議論したいこと

今回我々はL2リングネットワークの更改に際し、既存のネットワークデザインをイチから見直し、新しい構成に刷新しました。

同時に400G-ZRを使いWDMを使わずに長距離伝送にもチャレンジしました。

これまで抱えていた課題をどうやって解決したのか、また更改後の現在の運用状況について共有した上で、

・更改時のネットワークデザインってどうしてる？
・新技術の導入はどういうタイミングで検討する？

などについて話し合いができればと思います。

※複製・転載・流用・転売・複写等することを固く禁じます

自己紹介

氏名：大日方　周太（おびなた　しゅうた）
所属：株式会社コミュニティネットワークセンター
　　技術本部 スペシャリスト

◆主な担当
・バックボーンネットワークの設計、構築、運用
・広域L2網/伝送の設計、構築、運用
・ピアリング交渉　(その他特命案件)

◆趣味
・スノーボード、ドライブ、ランニング

◆JANOG歴
・JANOG42(津)　　：初登壇(ライトニングトーク)
・JANOG48(大垣)　　：初プログラム登壇(グループ会社2人で)
・JANOG50(函館)　　：初ひとりでプログラム登壇
・JANOG51(富士吉田)　：他社の方3人と4人で登壇
・JANOG54(奈良)　　：今回のプログラム登壇(後輩と2人で)

※複製・転載・流用・転売・複写等することを固く禁じます

自己紹介

氏名：神谷　健太（かみや　けんた）
所属：株式会社コミュニティネットワークセンター
　　技術本部 コアネットワークグループ

◆主な担当
・バックボーンネットワークの設計、構築、運用
・広域L2網/伝送の設計、構築、運用
・ピアリング交渉等・・・

◆趣味
・旅行

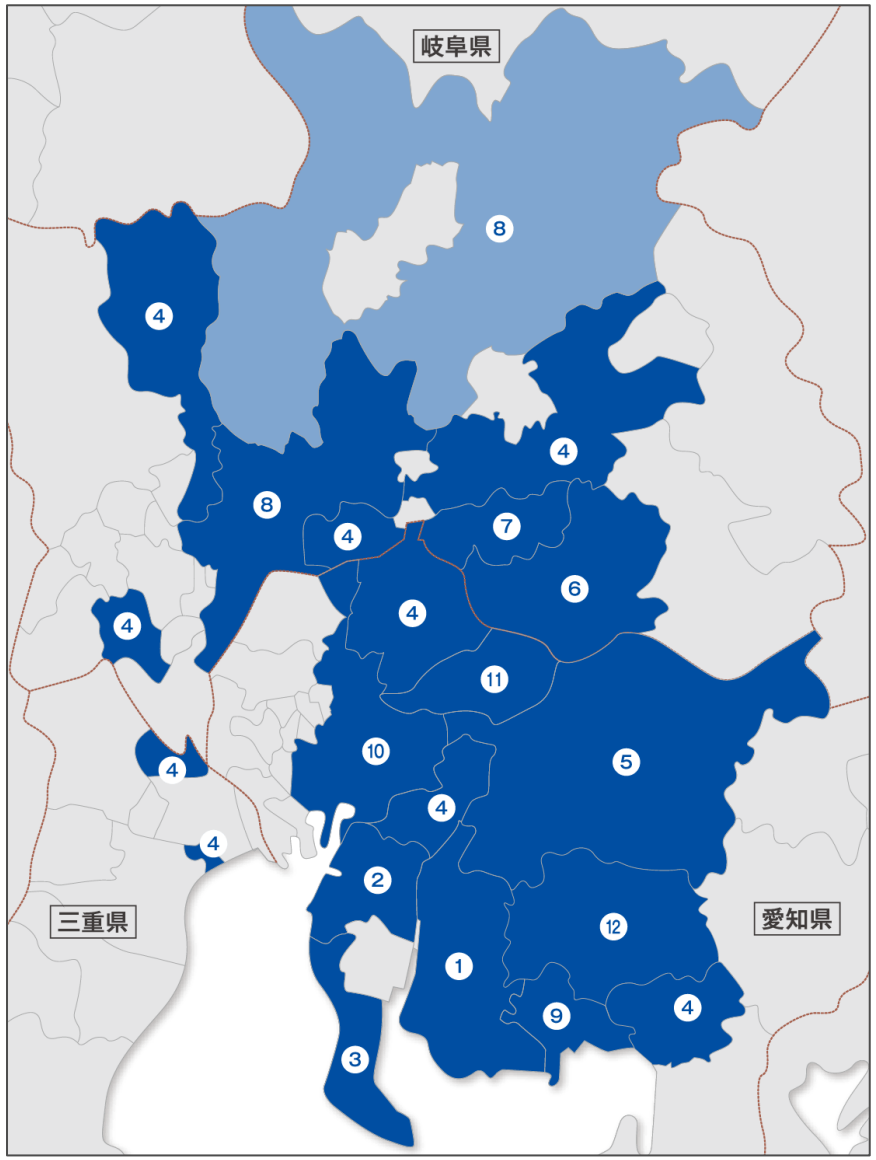
◆JANOG歴
・JANOG27(金沢)：リモート初参加
　　（学生時代にUstreamのライブ中継を視聴）
・JANOG34(高松)：初現地参加
・JANOG54(奈良)：今回のプログラム登壇

※複製・転載・流用・転売・複写等することを固く禁じます

㈱コミュニティネットワークセンターについて

東海地方のCATV局**12社**のグループ会社を
中心に、下記のサービスを展開しています。

- 放送配信サービス
- 通信サービス
- CATV局向けソリューション、技術支援
- 番組・物品共同調達



CNCi グループ

KATCH medias TAC CCNC CCNet ひまわりネットワーク Oribe NET WORK CTK ケーブルテレビ可児

CCN 三河湾ネットワーク株式会社 STARCAT Green City mics ミクスネットワーク

※複製・転載・流用・転売・複写等することを固く禁じます

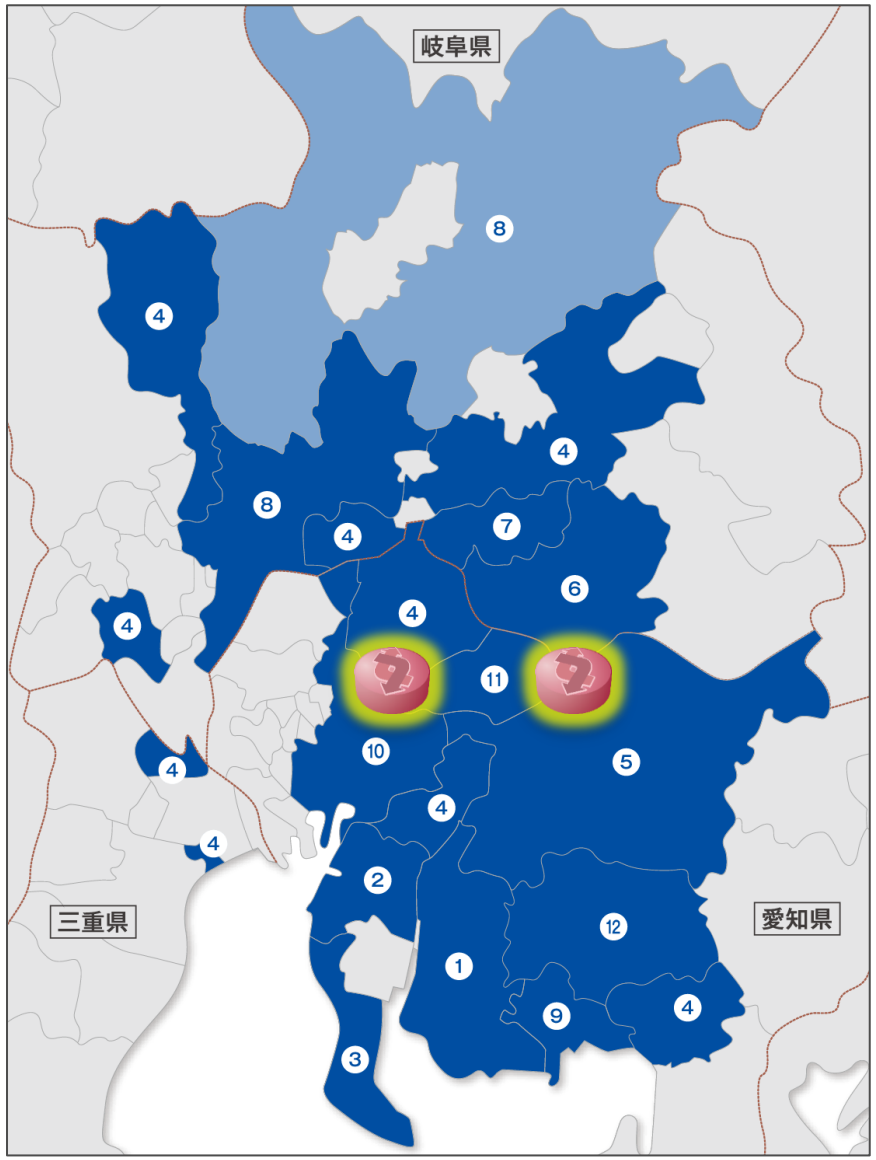
Agenda

- ネットワーク構成とL2リングの用途
- 抱えていた課題
- 更改対象のL2リングについて
- 更改後の構成や運用状況、将来設計、残された課題
- 400G-ZRの採用について
- 質疑、議論

※複製・転載・流用・転売・複写等することを固く禁じます

CNCIグループのネットワーク構成

- 理想
  →大きなBGPルータを真ん中に置いて
  　すべてのグループを収容したい



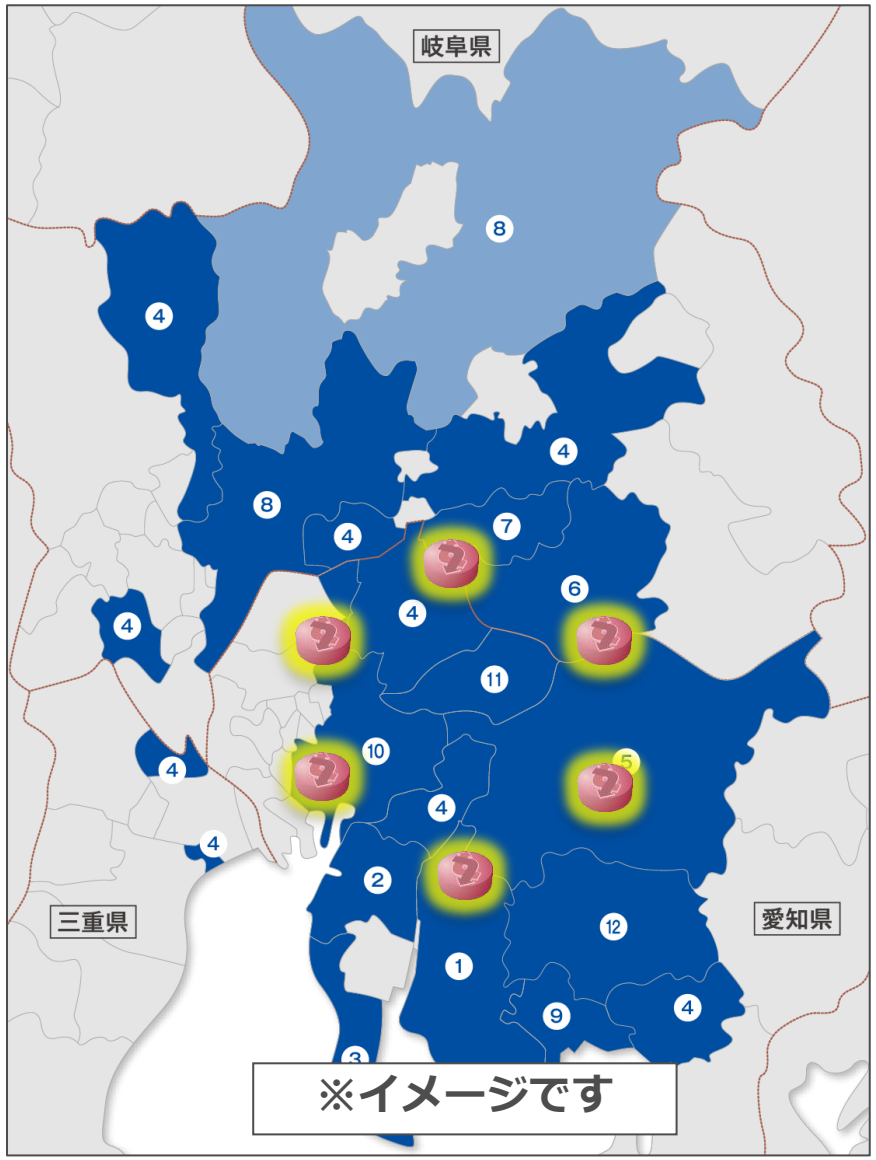
※複製・転載・流用・転売・複写等することを固く禁じます

CNCIグループのネットワーク構成

- 理想
  →大きなBGPルータを真ん中に置いて
  すべてのグループを収容したい

- 現実
  →エリアが広すぎる
  1局集中のリスクは回避したい

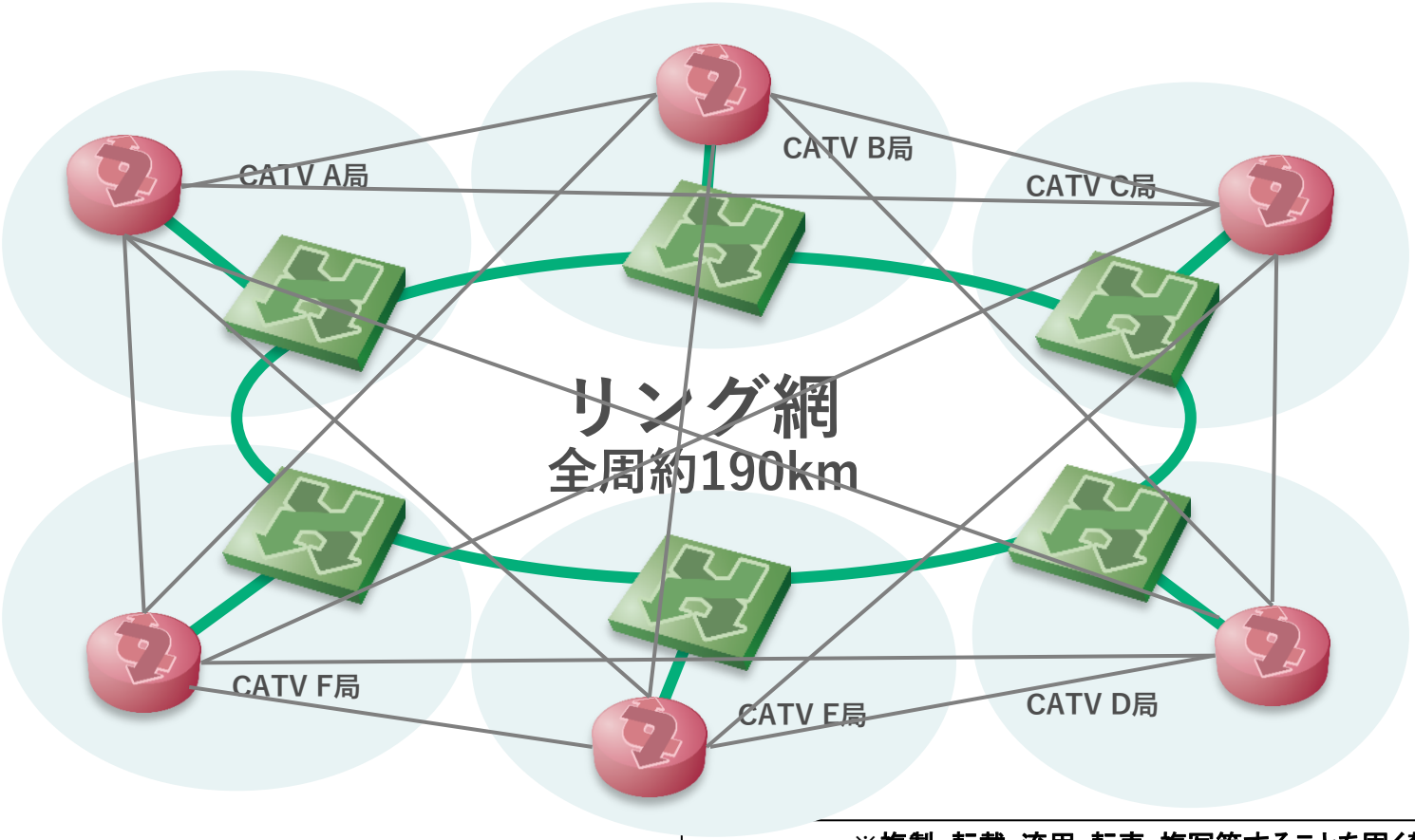
  →<u>BGPルータを分散配置している</u>

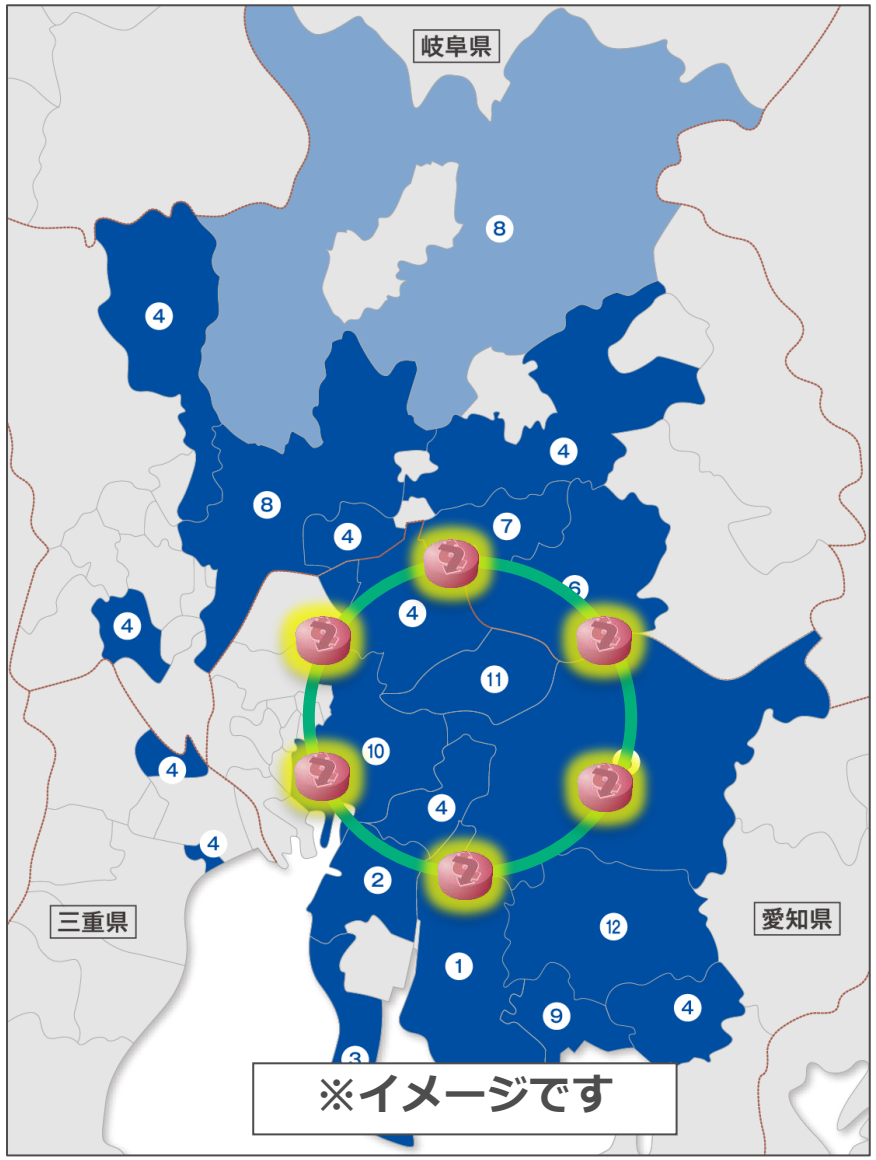


※複製・転載・流用・転売・複写等することを固く禁じます

CNCIグループのネットワーク構成

BGPルータ同士を接続するL2網
(iBGPをフルメッシュで接続)
→L2リングの主な用途

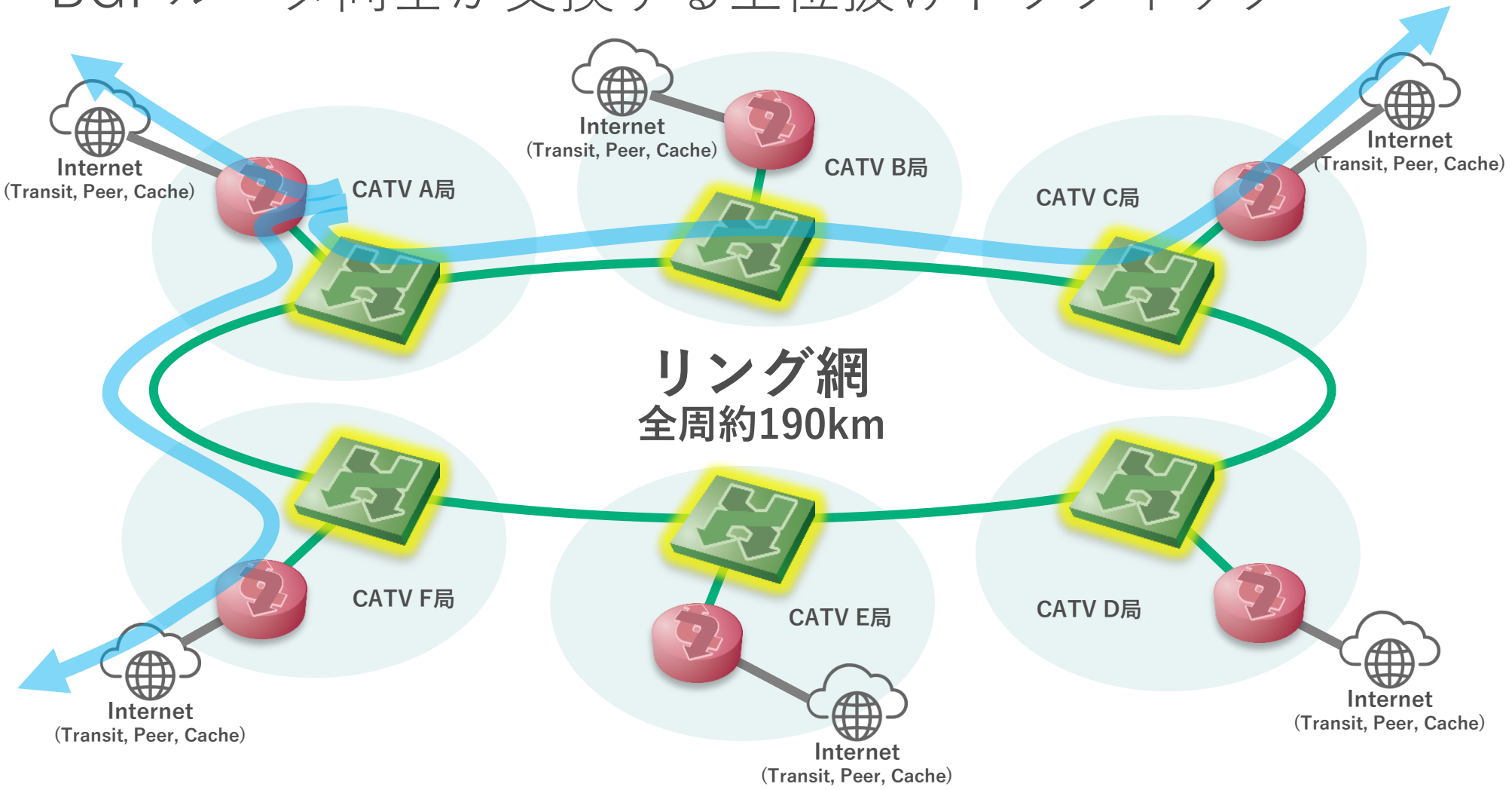




※複製・転載・流用・転売・複写等することを固く禁じます

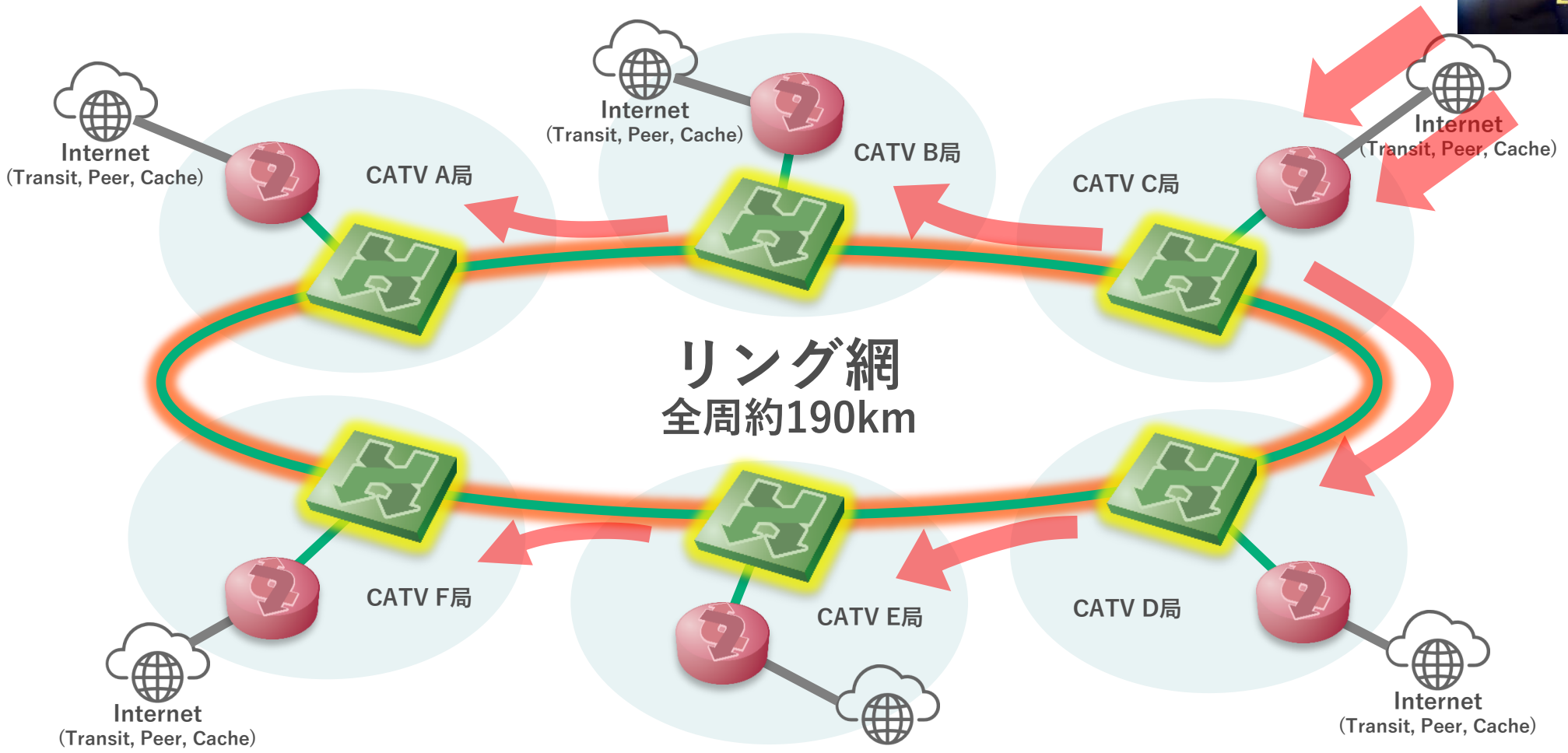
CNCIグループのネットワーク構成

L2リングに流れるトラフィック
→BGPルータ同士が交換する上位抜けトラフィック



【課題①】トラフィックの増加による帯域不足

近頃のイベントトラフィックによる帯域の圧迫
→ゲームアップデートやスポーツ配信など

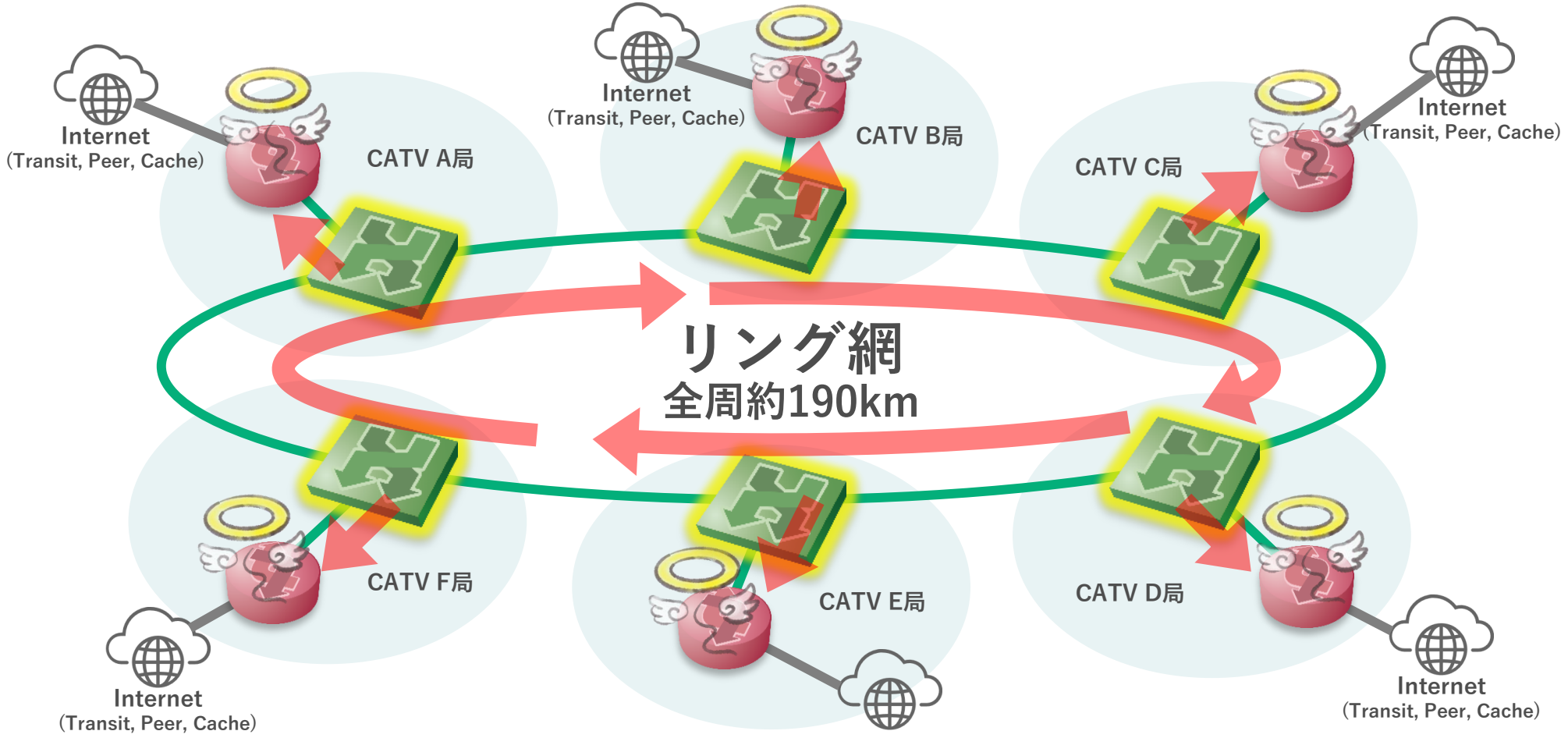


※複製・転載・流用・転売・複写等することを固く禁じます

引用元：AmazonPrimeVideo［www.amazon.co.jp］

【課題②】ループのリスク

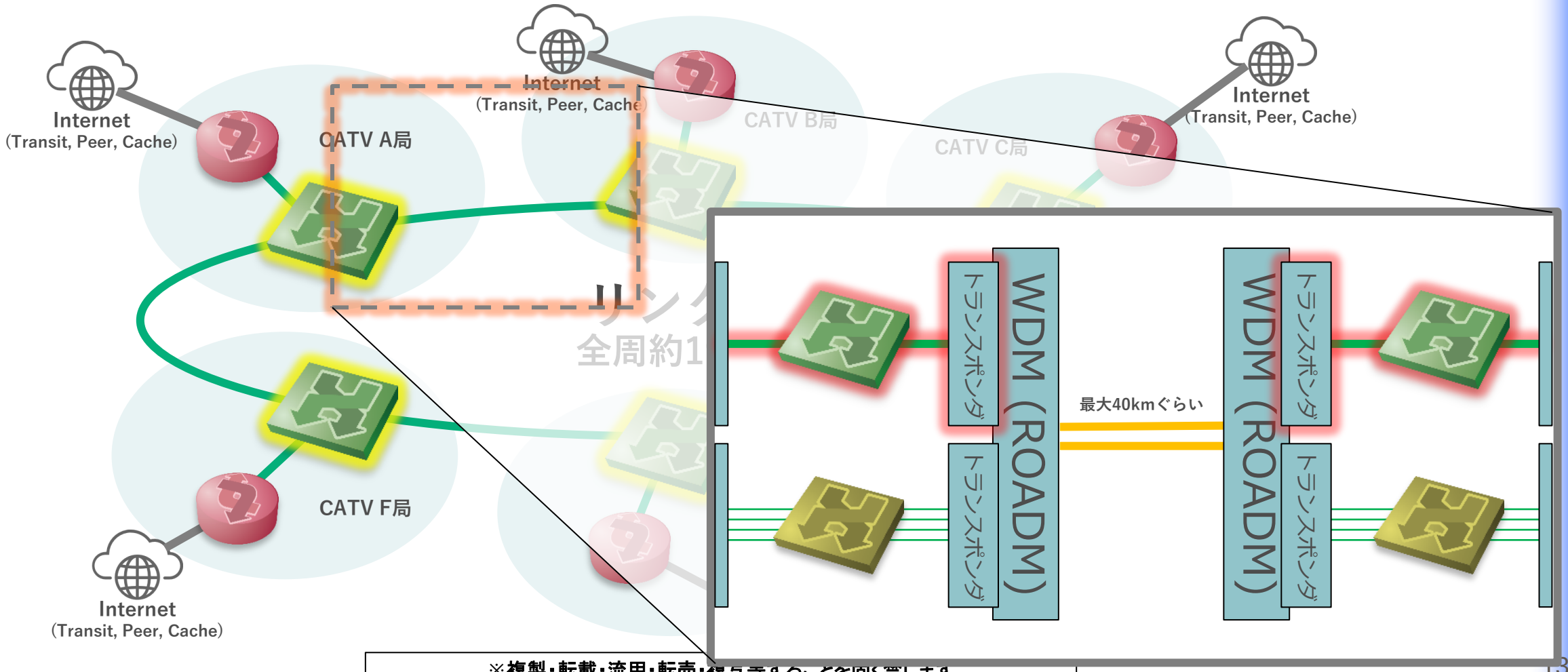
L2である以上、設定ミスやソフトウェアのバグでループ
→最悪、ルータがハングアップする可能性も。。。



※複製・転載・流用・転売・複写等することを固く禁じます

【課題③】局間接続にかかるコスト

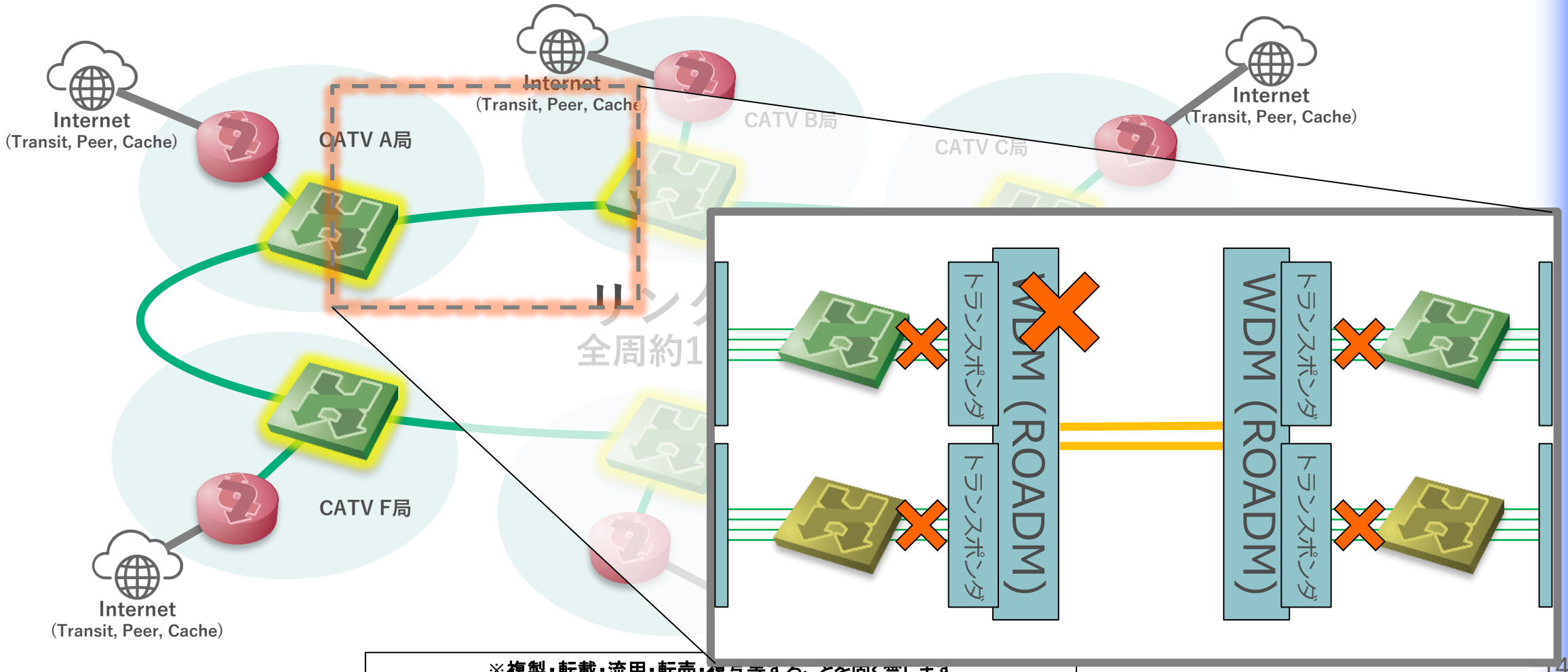
・局間にはWDMが入っているので、増速にはWDM側の増強も必要
→L2だけではなくトランスポンダも入れ替え



※複製・転載・流用・転売・複写等することを固く禁じます

【課題④】WDMが壊れたときの冗長性

・同じWDMに複数面のL2スイッチが収容されている
→WDMが壊れると複数面のリングに影響



※複製・転載・流用・転売・複写等することを固く禁じます

課題を解決したいけど・・・

※イメージです

オレ忙しいから、100Gとか400Gとか広帯域で、最近流行ってるVXLAN？とかSR？とか使っちゃって、かっこいいにリングに更改しといて～もちろん安く仕上げてね😈

は、はい。
分かりました…

※複製・転載・流用・転売・複写等することを固く禁じます

更改するにあたり個人的に気にしたこと

- できるだけ更改前と同じように運用できること
- できるだけ将来的な拡張の余地があること
- できるだけ無難でシンプルなこと
  - 学習コストが低い
  - 運用が複雑にならない
  - トラブルシュートが複雑にならない

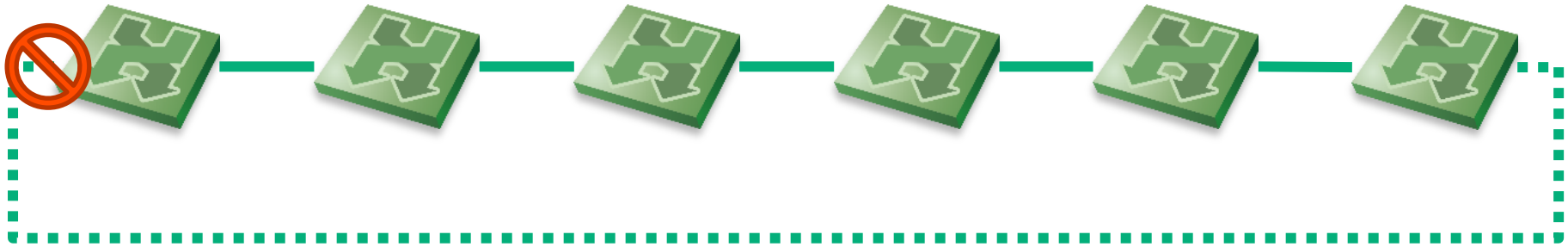
※複製・転載・流用・転売・複写等することを固く禁じます

更改対象のL2リングについて

※複製・転載・流用・転売・複写等することを固く禁じます

リングプロトコルの良さ①

- シンプルでイメージがしやすい
  - L2スイッチが数珠つなぎになっている
  - 障害時はブロッキングポートが開放されMACテーブルがクリアされる

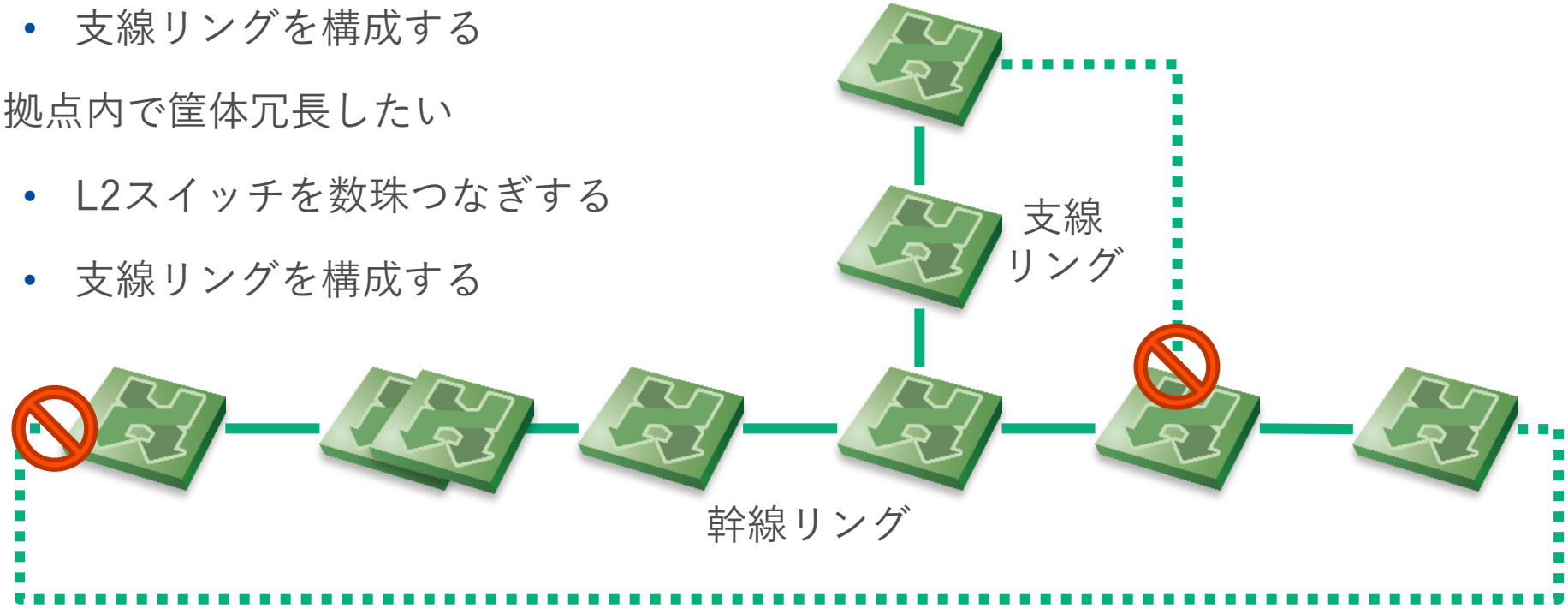


※複製・転載・流用・転売・複写等することを固く禁じます

リングプロトコルの良さ②

◆ シンプルな拡張性がある

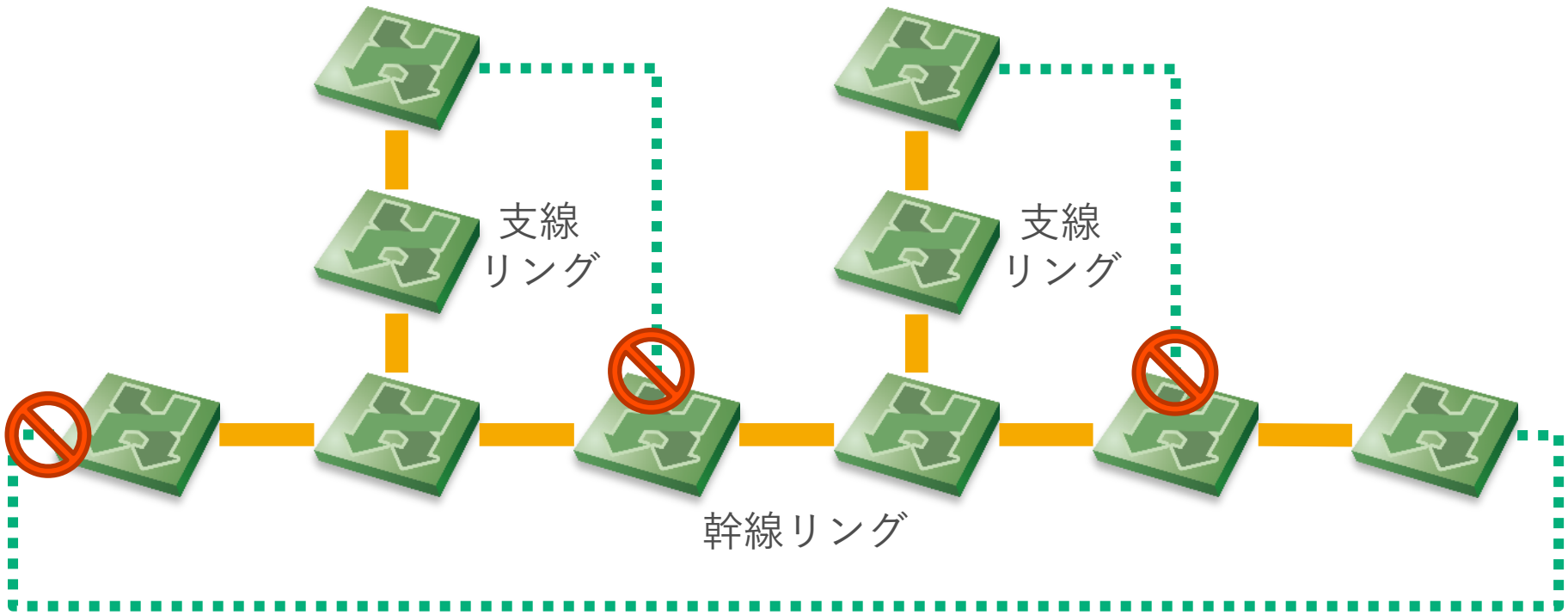
- 別の地域へL2網を伸ばしたい
  - 支線リングを構成する
- 拠点内で筐体冗長したい
  - L2スイッチを数珠つなぎする
  - 支線リングを構成する



※複製・転載・流用・転売・複写等することを固く禁じます

リングプロトコル網のトラフィックの例①

◆ 通常時

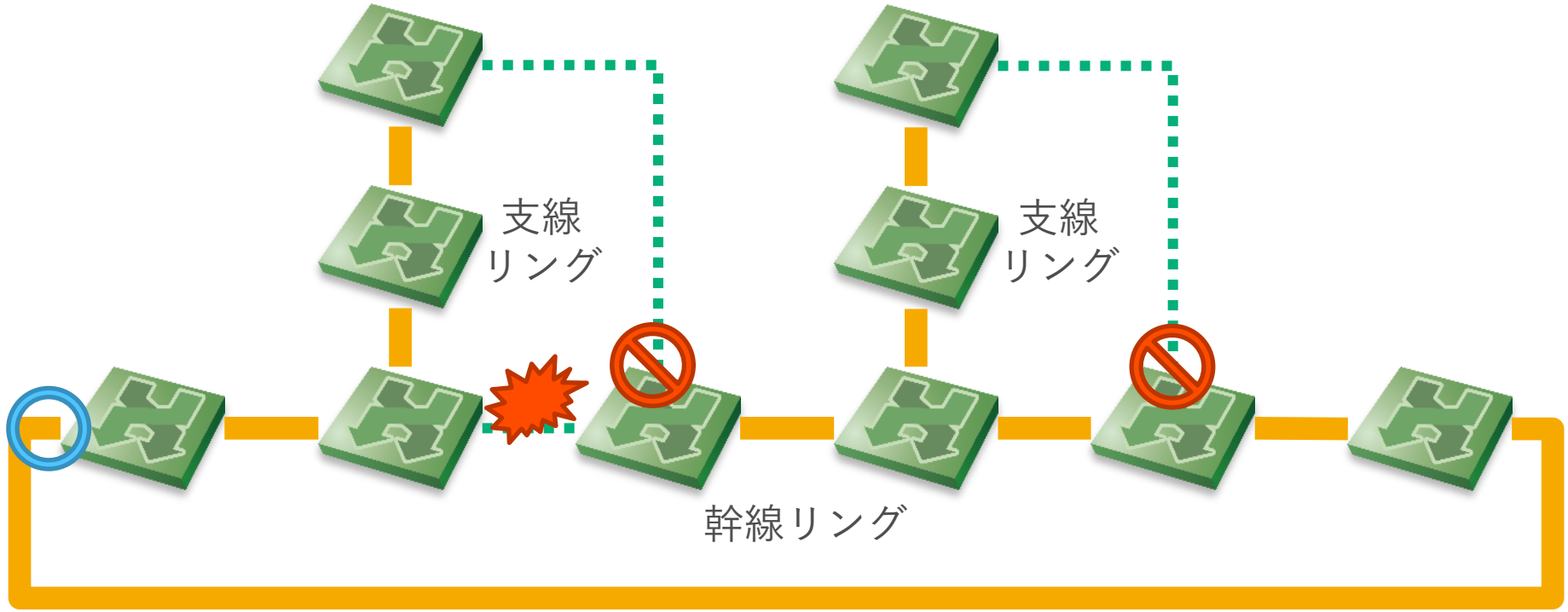


※複製・転載・流用・転売・複写等することを固く禁じます

リングプロトコル網のトラフィックの例②

◆ 幹線リング障害時

- 幹線リングのブロッキングポートが開放される

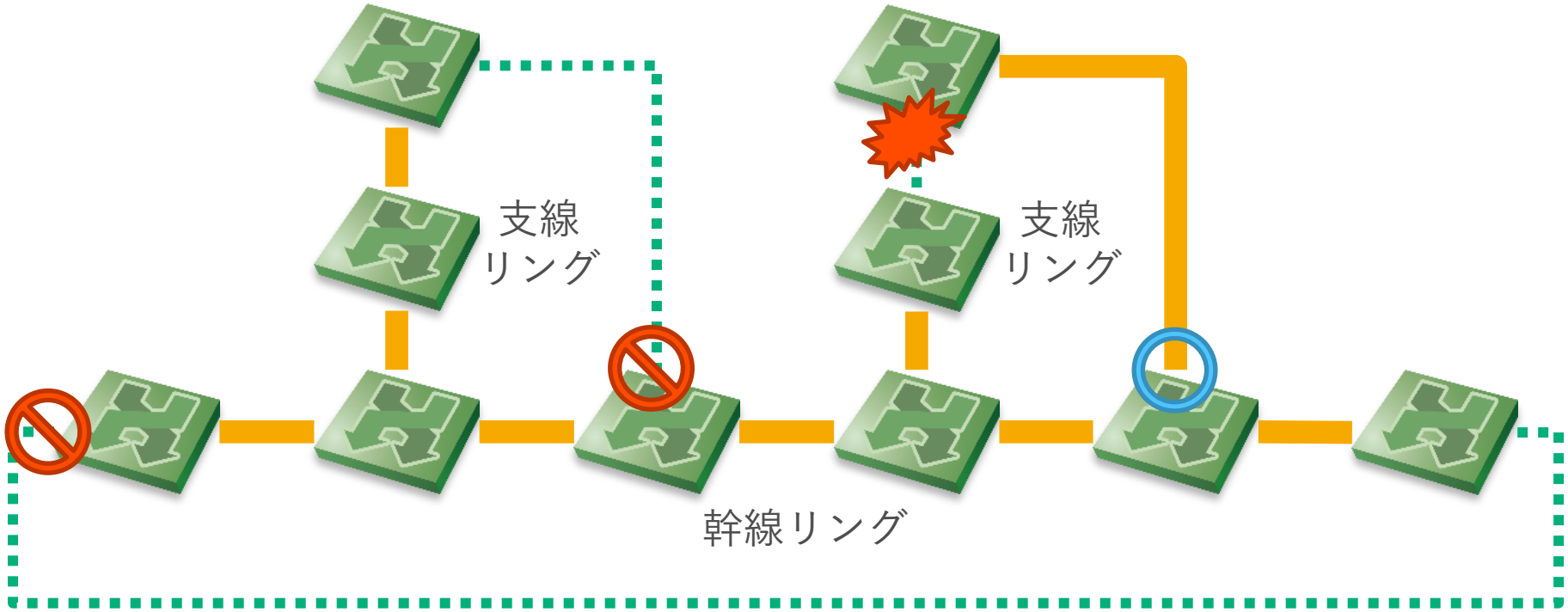


※複製・転載・流用・転売・複写等することを固く禁じます

リングプロトコル網のトラフィックの例③

◆ 支線リング障害時

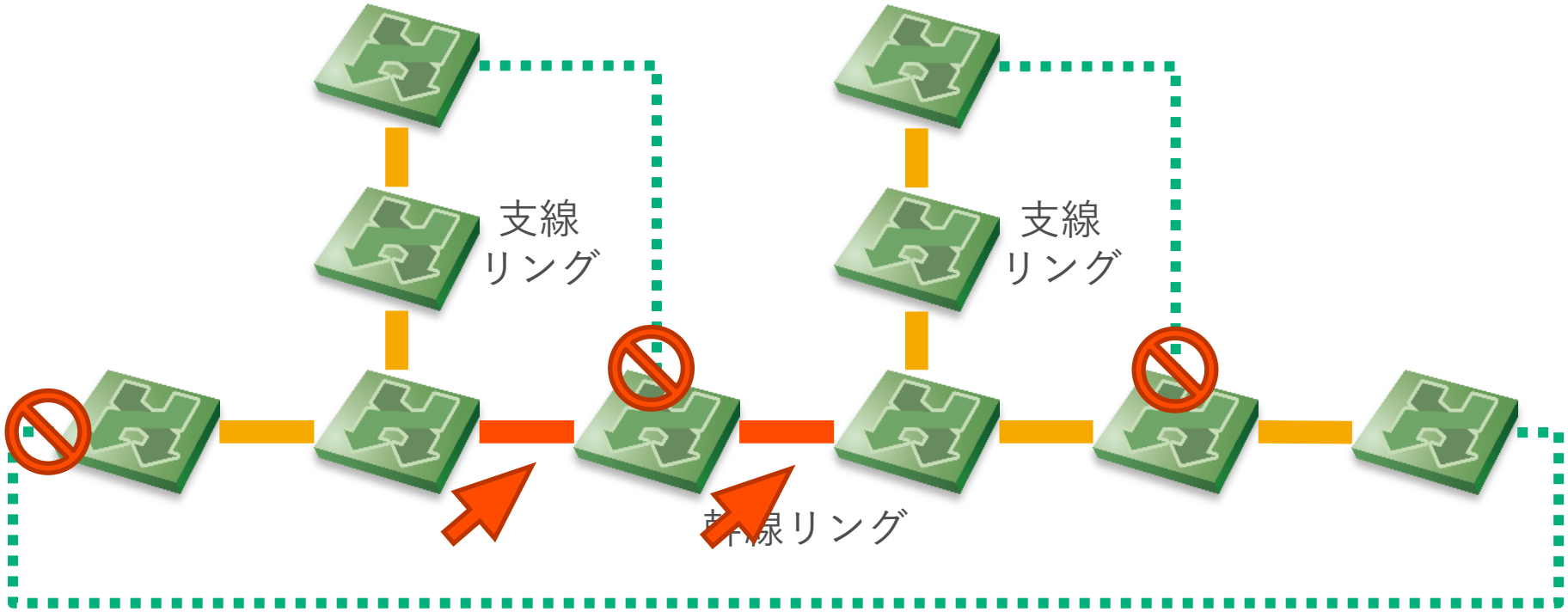
- 支線リングのブロッキングポートが開放される



※複製・転載・流用・転売・複写等することを固く禁じます

リングプロトコル網のトラフィックの特徴①

◆ リングの中央付近にトラフィックが集中しやすい

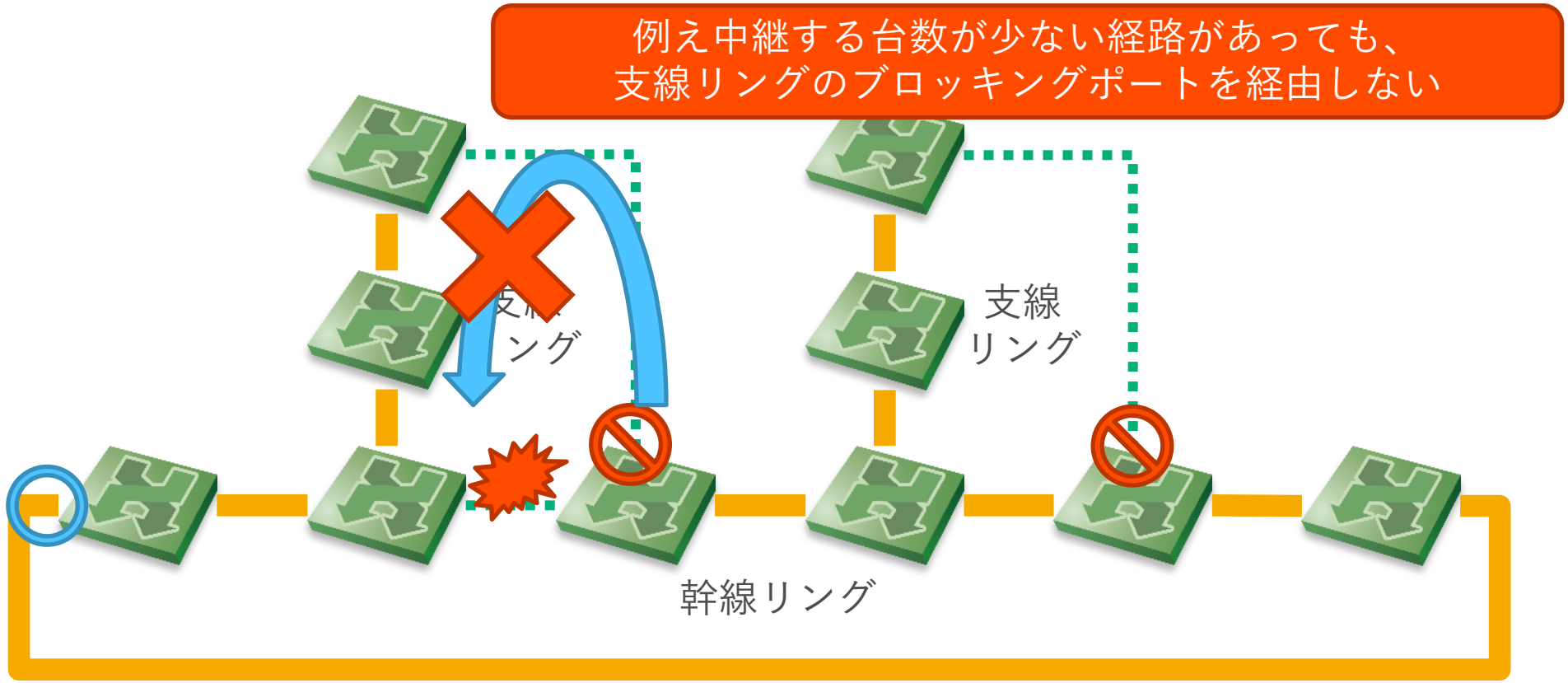


通常時このあたりにトラフィックが集中しやすい

※複製・転載・流用・転売・複写等することを固く禁じます

リングプロトコル網のトラフィックの特徴②

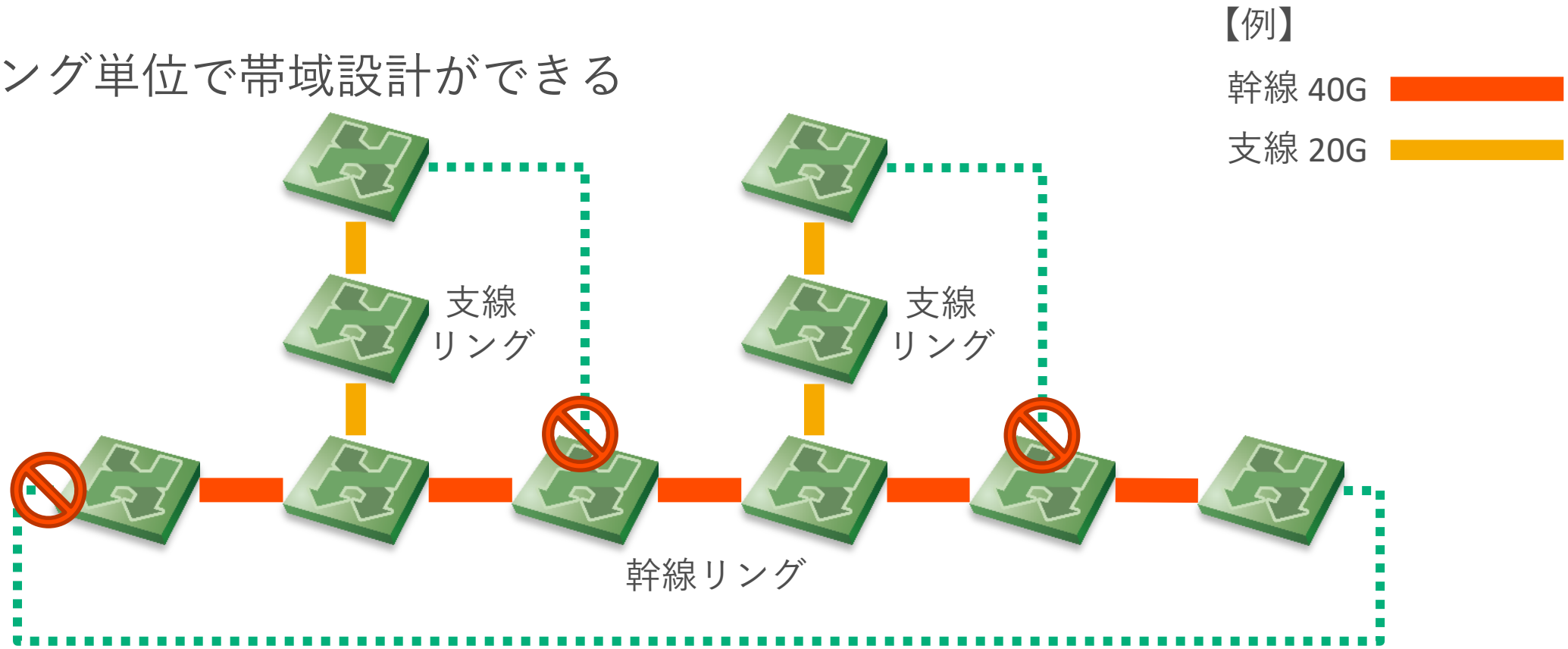
◆ 幹線リングの障害時に支線リングにトラフィックが迂回しない



※複製・転載・流用・転売・複写等することを固く禁じます

リングプロトコル網の帯域設計の特徴

- 幹線リングに近づくほど帯域が広い
- リング単位で帯域設計ができる

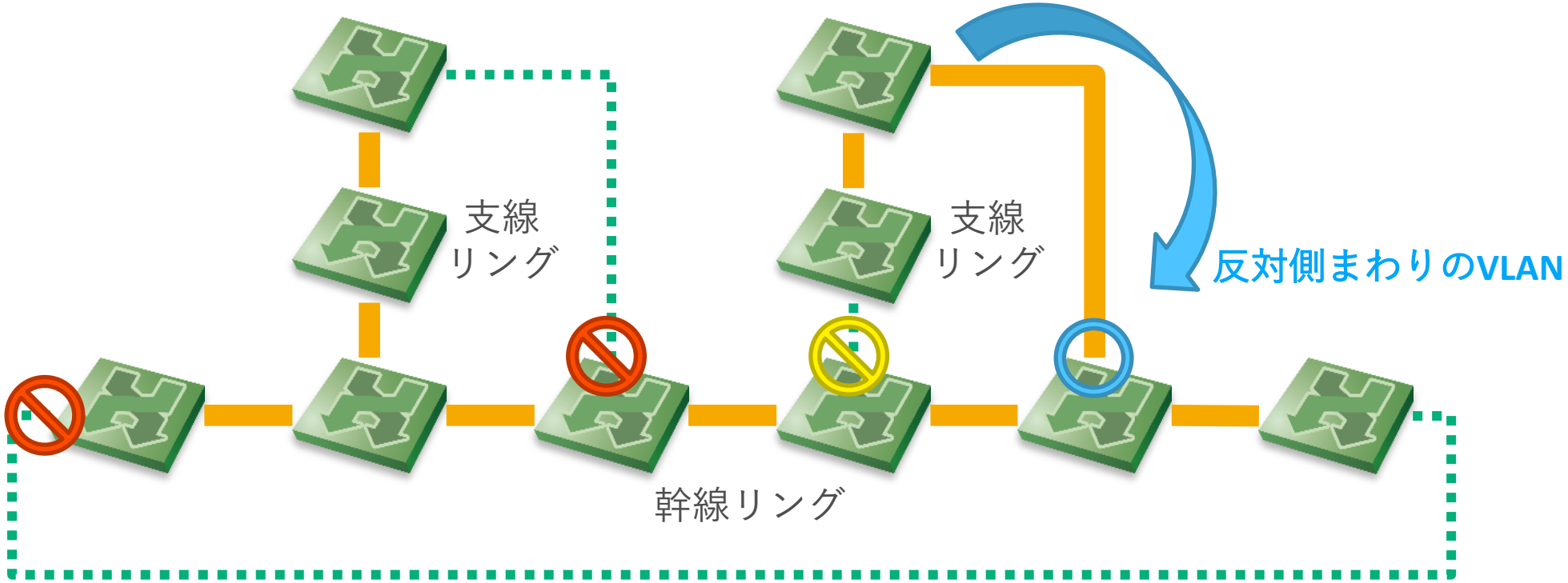


※複製・転載・流用・転売・複写等することを固く禁じます

リングプロトコル網のトラフィック分散

◆ 指定したVLANのみブロッキングポートを通過させる

- 反対側（支線リングの場合はリングの根元）のポートがブロッキングポートになる



※複製・転載・流用・転売・複写等することを固く禁じます

リングプロトコル網の課題

◆ 運用上の主な課題

- トラフィックエンジニアリング
  - トラフィック分散が部分的にしかできない
  - トラフィック迂回をするにはリングを切るしかない（回線支障等の対応など）
- ループの恐怖に怯えながらの運用
  - バグや設定ミスによるトラフィックの輻輳

◆ 更改面での主な課題（既存踏襲をする場合）

- 更改対象としてL2スイッチとWDMの更改が必要（高コスト）
- メーカ独自プロトコルのため選択肢が少ない（メーカのラインナップやロードマップに依存）

※複製・転載・流用・転売・複写等することを固く禁じます

更改後のネットワークについて

※複製・転載・流用・転売・複写等することを固く禁じます

EVPN/VXLANの採用

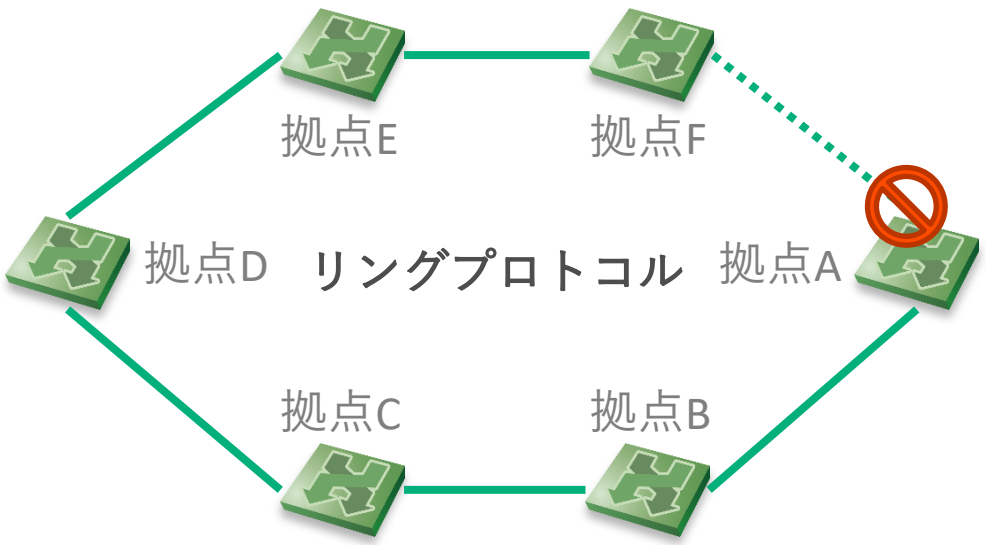
CNCi

- EVPN/VXLANはL3ネットワーク上にL2ネットワークを構築する技術
  - EVPN：MACアドレスの経路情報をBGPで交換する
  - VXLAN：L3ネットワーク上にL2ネットワークを構築するトンネル技術
  - EVPN/VXLAN：EVPNがコントロールプレーン、VXLANがデータプレーンのネットワーク
    - MACアドレス経路のネクストホップがVXLANのカプセル化を行うVTEPのアドレスになる
  - 標準化されているプロトコルである
  - 大手通信事業者での豊富な採用実績がある
    - 過去のJANOGでも採用実績が多数紹介されています
  - 東海地域の通信事業者として普段からBGPを運用している

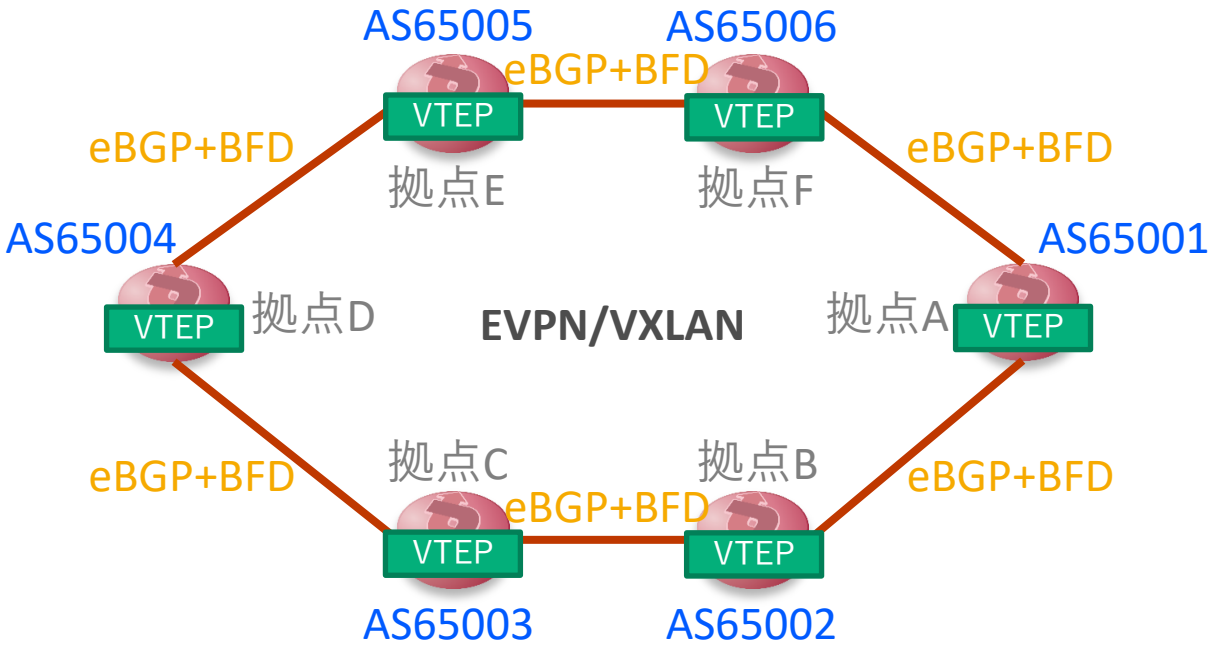
※複製・転載・流用・転売・複写等することを固く禁じます

更改前後のネットワークの概要図

更改前



更改後

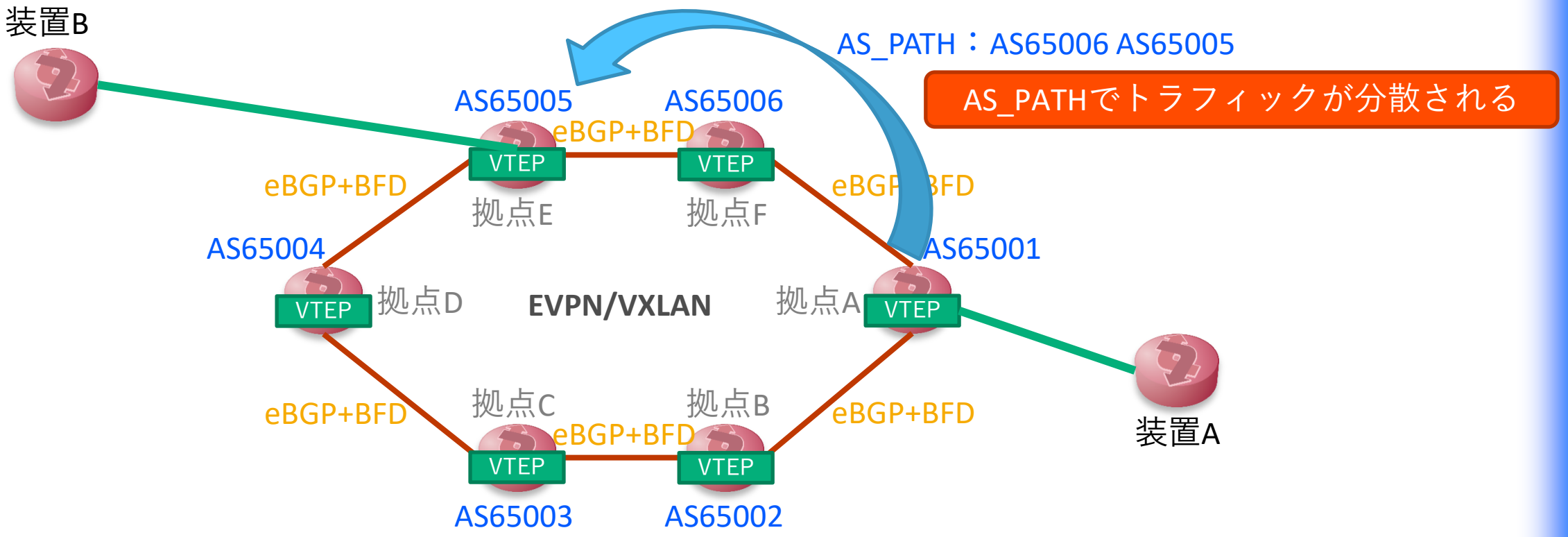


VTEPのアドレスとEVPNの情報をBGPで経路交換

※複製・転載・流用・転売・複写等することを固く禁じます

EVPN/VXLAN網のトラフィックの例①

◆ eBGPの経路情報によりトンネル先のVTEPまでは AS_PATHの最短経路

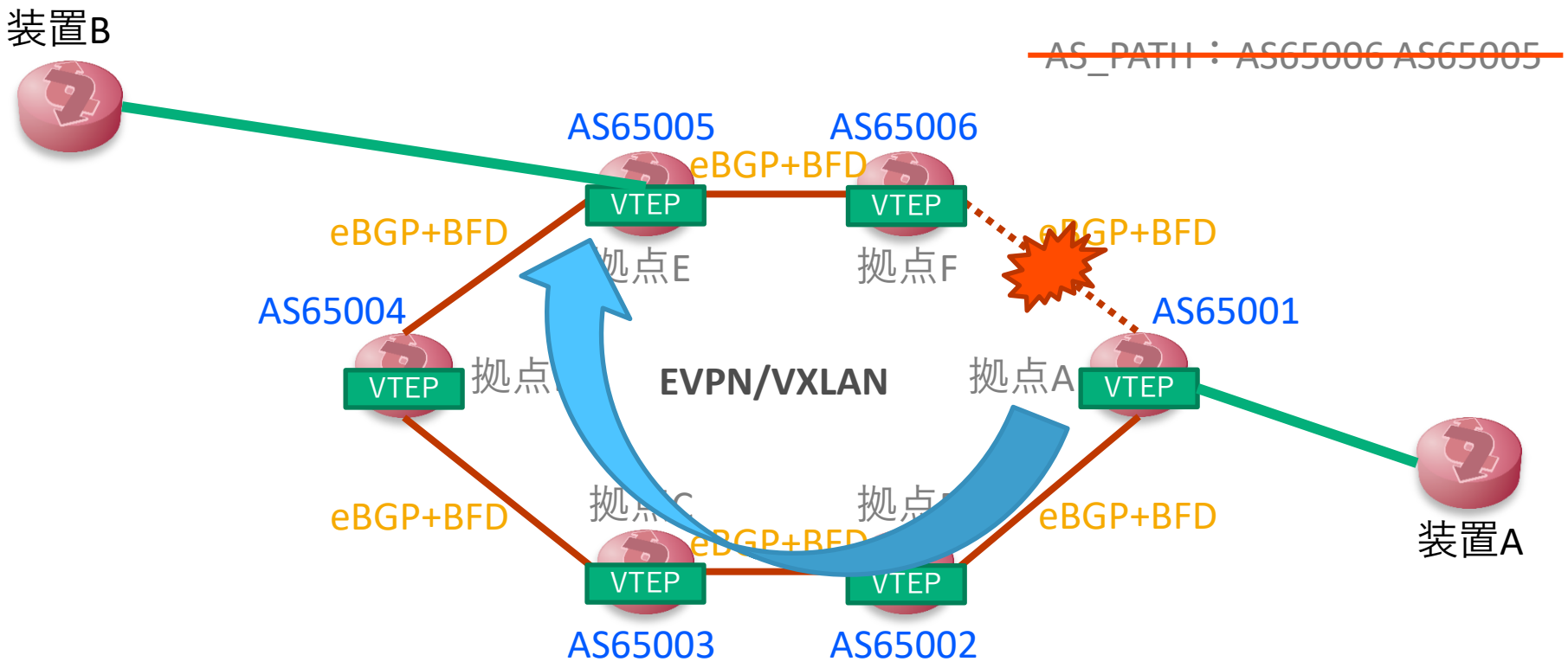


AS_PATH：AS65002 AS65003 AS65004 AS65005

※複製・転載・流用・転売・複写等することを固く禁じます

EVPN/VXLAN網のトラフィックの例①

◆ 障害時

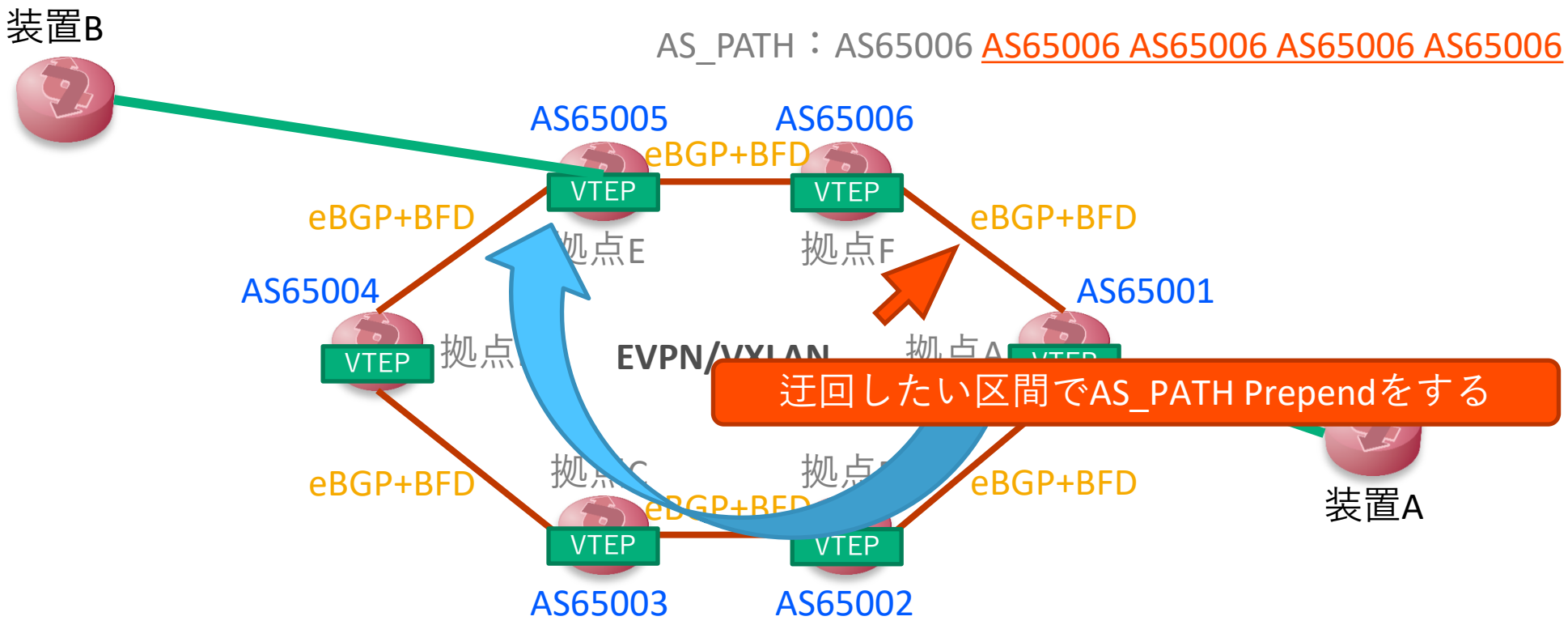


AS_PATH：AS65002 AS65003 AS65004 AS65005

※複製・転載・流用・転売・複写等することを固く禁じます

EVPN/VXLAN網のトラフィックの迂回

◆ AS_PATH Prependで冗長性を確保したままトラフィックを迂回



AS_PATH：AS65002 AS65003 AS65004 AS65005

※複製・転載・流用・転売・複写等することを固く禁じます

設計時のあれこれ

- ECMPにならないようにリングの1カ所にだけ常時Prependを入れた
  - リングが偶数拠点のため反対側の拠点までAS_PATHが等コストになってしまうため
- BGP Additional Pathsを有効にした
  - バックアップ経路が有効であることを確認するため
- 検証時に固定のIPv6 Link-localによるPeerを設定していたが不採用とした
  - BGPピアのアドレス設計が不要になるがtracerouteができなかったため
    - 一般的にICMP Time Exceededは入力インターフェースのアドレスを使用するらしい
    - ネットワークの規模や障害時のトラブルシュートを想定し不採用とした
    - アドレス設計を機械的に行えるようにしつつtracerouteができるIPv6アドレスを使用
    - 同じ理由でBGP Unnumberedも検証せずに不採用とした

※複製・転載・流用・転売・複写等することを固く禁じます

課題

◆ マルチキャストに課題が残る

- EVPN/VXLANでマルチキャストは複製される（HER：Head End Replication）
- L2スイッチが数珠つなぎのリングプロトコル網では気にしなくてよい要素
- ケーブルテレビ局なのでマルチキャストで映像配信を行うことも

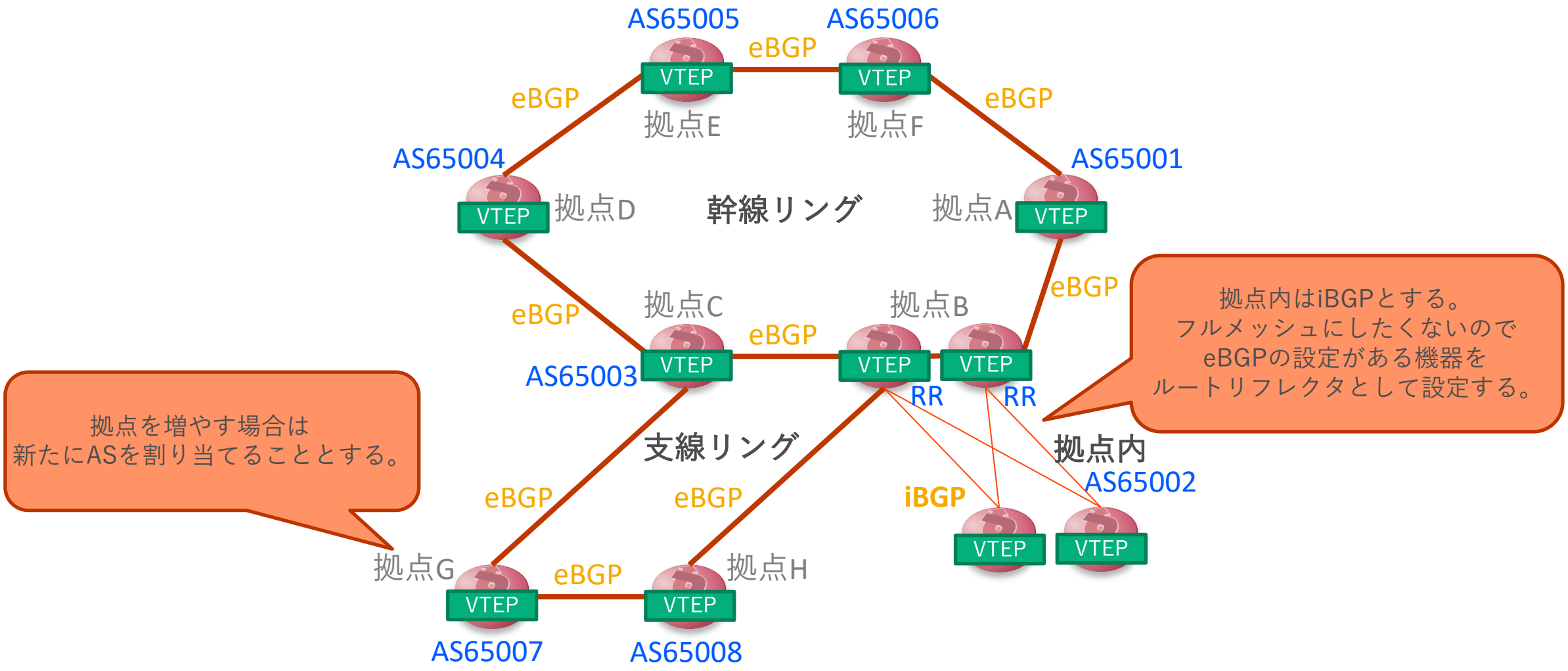
※複製・転載・流用・転売・複写等することを固く禁じます

将来的な拡張性についての検証

- リングプロトコル網に近い拡張性を確認したい
  - 拠点を増やすことができるか？
  - 拠点内で機器を増やすことができるか？
  - どのような拡張方法となるのか想定した上で更改したい
- 検証にはコンテナ環境を使用
  - （参考）JANOG51で紹介されていたcontainerlabを使用
    - https://www.janog.gr.jp/meeting/janog51/lab/

※複製・転載・流用・転売・複写等することを固く禁じます

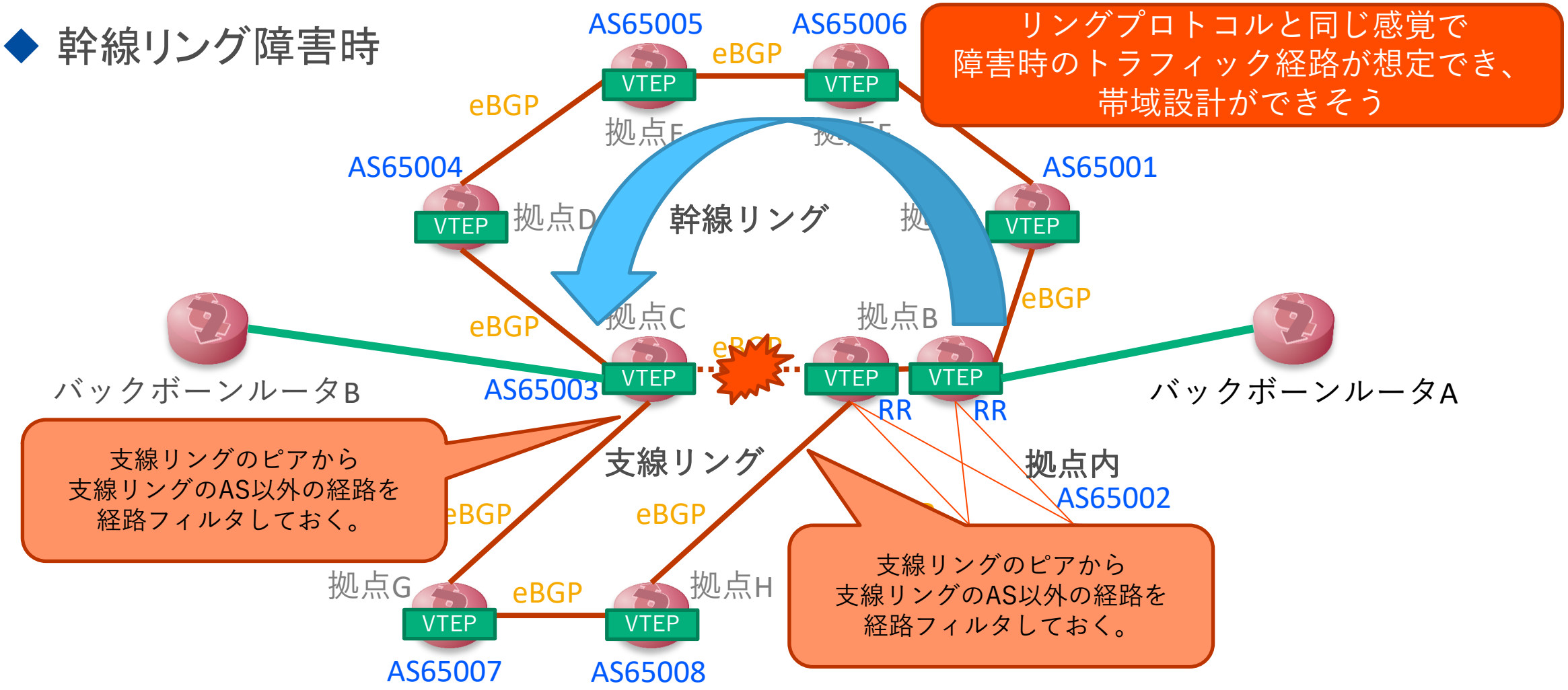
検証したネットワークの概要図



※複製・転載・流用・転売・複写等することを固く禁じます

検証したネットワークのトラフィックの例

◆ 幹線リング障害時



リングプロトコルと同じ感覚で
障害時のトラフィック経路が想定でき、
帯域設計ができそう

支線リングのピアから
支線リングのAS以外の経路を
経路フィルタしておく。

支線リングのピアから
支線リングのAS以外の経路を
経路フィルタしておく。

※複製・転載・流用・転売・複写等することを固く禁じます

将来的な拡張性についての検証結果

- 拠点を増やすことができるか？
  - 新しいASを拠点に割り当てる
- 支線リングにあたる拠点を増やすことができるか？
  - 新しいASを拠点に割り当てる
  - 支線リングのピアからは支線リング以外の経路をフィルタする
- 拠点内で機器を増やすことができるか？
  - iBGPとルートリフレクタでネットワークを構成する
- その他
  - EVPN/VXLAN All Active MultihomingでMLAGも標準プロトコルで使えるようになった

※複製・転載・流用・転売・複写等することを固く禁じます

比較検討したその他の構成

◆ eBGP+VXLAN

- VXLANのカプセル化を行うVTEPのアドレスをeBGPで経路広告する構成
  - 別のネットワークで採用し運用実績あり
- EVPNの代わりにL2ネットワークを構成するVTEPのフラッドリストを静的に設定する
- スイッチを増設する分だけフラッドリストに追記しなければならないという課題

◆ EVPN/MPLS

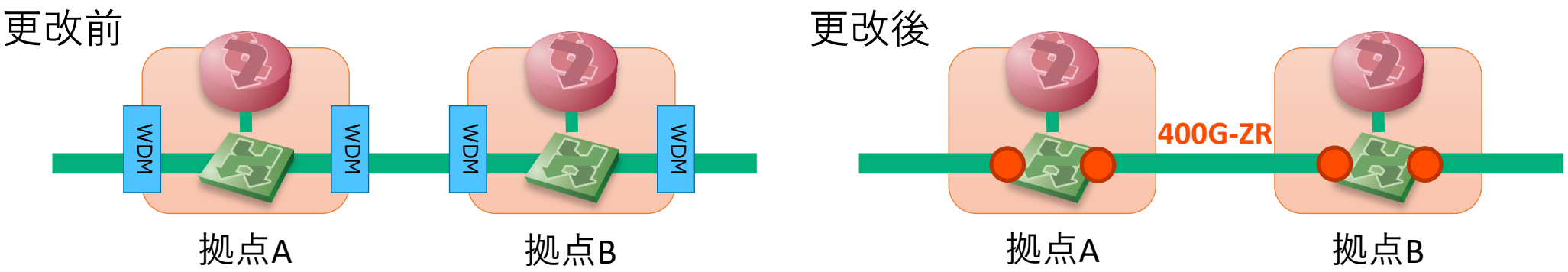
- MPLSの運用経験ゼロで学習コストが非常に高い
- 検証環境を構築したが今までのリング網の運用感覚と大きく異なると感じた
- VPWSという仮想回線の機能が便利そう

※複製・転載・流用・転売・複写等することを固く禁じます

400G-ZRオプティクスの採用

◆ 400G-ZRは120kmの長距離伝送が可能な技術

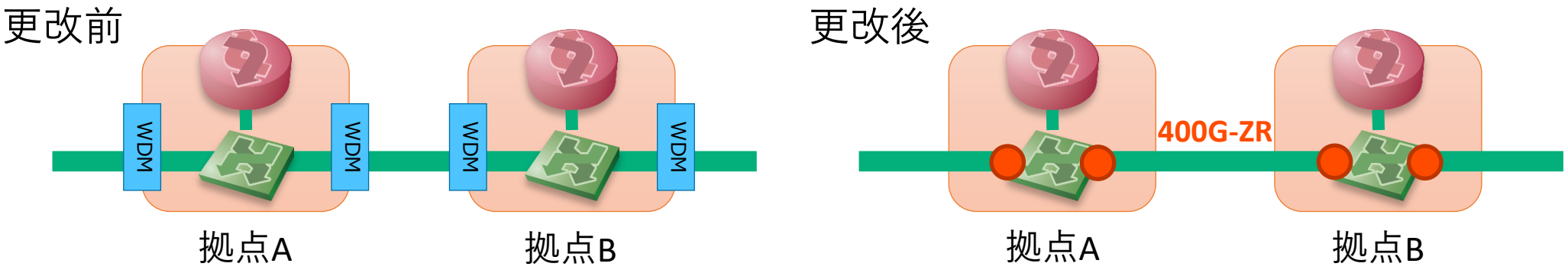
- 本件の最長区間38.1km(損失16.9dB)でもかなり余裕をもって伝送が可能
  - お借りした検証機にて該当の拠点間で使用感を確認済み(2021年9月頃)
- 別の区間34.7km(損失8.3dB)ではAMPモジュール無しで伝送が可能
- （参考）採用した技術の詳細はJANOG51のプログラムで紹介されています
  - https://www.janog.gr.jp/meeting/janog51/shinkansen/



※複製・転載・流用・転売・複写等することを固く禁じます

400G-ZRオプティクス採用のメリット

- スイッチとWDMの両方を更改するよりもコストを抑えることができた
- 障害ポイントを減らすことができた
  - 障害時の切り分けで発生する運用負荷が軽減
    - 従来）拠点間のリンクダウン発生時にスイッチとWDMの各サポートに問い合わせを実施
    - 従来）サポートデータを対向のスイッチとWDMで合計4台から取得
    - 従来）双方の窓口から「問題ない」と回答があるとトラブルシュートが長引くことも



※複製・転載・流用・転売・複写等することを固く禁じます

事例紹介まとめ

- 【課題①】トラフィックの増加による帯域不足
  - 広帯域な規格である400Gを採用することで解決
  - EVPN/VXLANを採用することで通常時のトラフィックも分散
    - ブロッキングポートのトポロジーからAS_PATHによる最短経路になった
- 【課題②】ループのリスク
  - L3ネットワーク上にL2ネットワークを構築するEVPN/VXLANを採用することで解決
    - 将来的な拡張性をある程度考慮した構成を採用することができた
    - 普段使用しているBGPがベースであるため学習コストを抑えた
- 【課題③】局間接続にかかるコスト
  - 個別の伝送装置(WDM)の代わりに400G-ZRモジュールを採用することで解決
- 【課題④】WDMが壊れたときの冗長性
  - 個別の伝送装置(WDM)の代わりに400G-ZRモジュールを採用することで解決
    - 個別の伝送装置(WDM)を使用した既設ネットワークと本ネットワークの冗長構成

※複製・転載・流用・転売・複写等することを固く禁じます

議論・質疑応答

- 既存踏襲vs新技術導入に関するメリット・デメリット
  - 新技術を導入する際はこんな観点で気を付けていますとか、どうしましたか？とか
  - 今回紹介した事例に関する気になるあれこれとか
- 実際の新技術の導入事例やその他アイデアの共有（会場から）
  - 会場からの新技術導入時の苦労話とか
  - 今回紹介した事例に対するその他のアイデアとか
- 例えば、大規模なL2ネットワークの更改するとしたら？
  - 今回のL2リングはEVPN/VXLANに置き換えることができたが、もっと大きなL2リングが…
  - VLANがいっぱいなネットワーク等、L2ネットワークってなかなか変えられない
- 例えば、L3ネットワークにおけるマルチキャストはどこまで最適化するか？
  - L2ネットワークだったら簡単に通せたマルチキャストについて
  - ケーブルテレビ事業者なのでマルチキャストの映像配信に使用しているL2リングがあったり…

※複製・転載・流用・転売・複写等することを固く禁じます

ご清聴ありがとうございました。