JANOG57 事前資料

JANOG57 in OSAKA

イベントネットワークの最前線を語ろう
- JANOG57会場ネットワークの知見とこれから

京都大学 吉川知輝
さくらインターネット 米田悠人
さくらインターネット 江草陽太

目次

1. タイムスケジュール
2. 今回採用する議論形式
3. 議論テーマ募集
4. 会場ネットワーク構成紹介
5. 議論可能なポイント

タイムスケジュール

～5分：プログラムと議論形式の説明

5分~35分：JANOG57における会場ネットワーク構成と運用について紹介

- ネットワーク設計や運用上の課題
- 会場ネットワークの利用状況やトラフィックパターンについて

35分~70分：特定のテーマについての議論

- 詳細は次ページ参照

70分~80分：他プログラムと同様にフリーテーマでの議論・質問

- もっと深掘りたい内容等

今回採用する議論形式

JANOG57 in OSAKA

- 2/6～プログラム開始30分後まで議論を希望するテーマを募集
- 事前に登壇者が設定したテーマ+希望の多かったテーマを1テーマ5~10分ほどで議論
  - 可能なものに関しては実際の会場ネットワークの情報を開示
- 複数参加者がマイク前に立ち講演者と合わせて同時に議論
  - 途中で随時マイク前の参加者の変更も可能にすることで議論に誰でも参加可能に
- Google Drive上にアップロードいただき参加者側からも資料の投影が可能

議論テーマ募集

- 以下のURLにて議論したいテーマを募集します
- URL: https://theme-submission.janog57.ishikari-dc.jp
  - 議論テーマの投稿・投票(いいね)が可能
- テーマの中から特に議論が盛り上がりそうなもの・得票が多いものを選び当日議論テーマに設定します
- テーマを投稿された方は是非当日の議論にも参加ください！
- 「議論可能なトピック」もご参照の上、興味のあるテーマを投稿・投票ください！

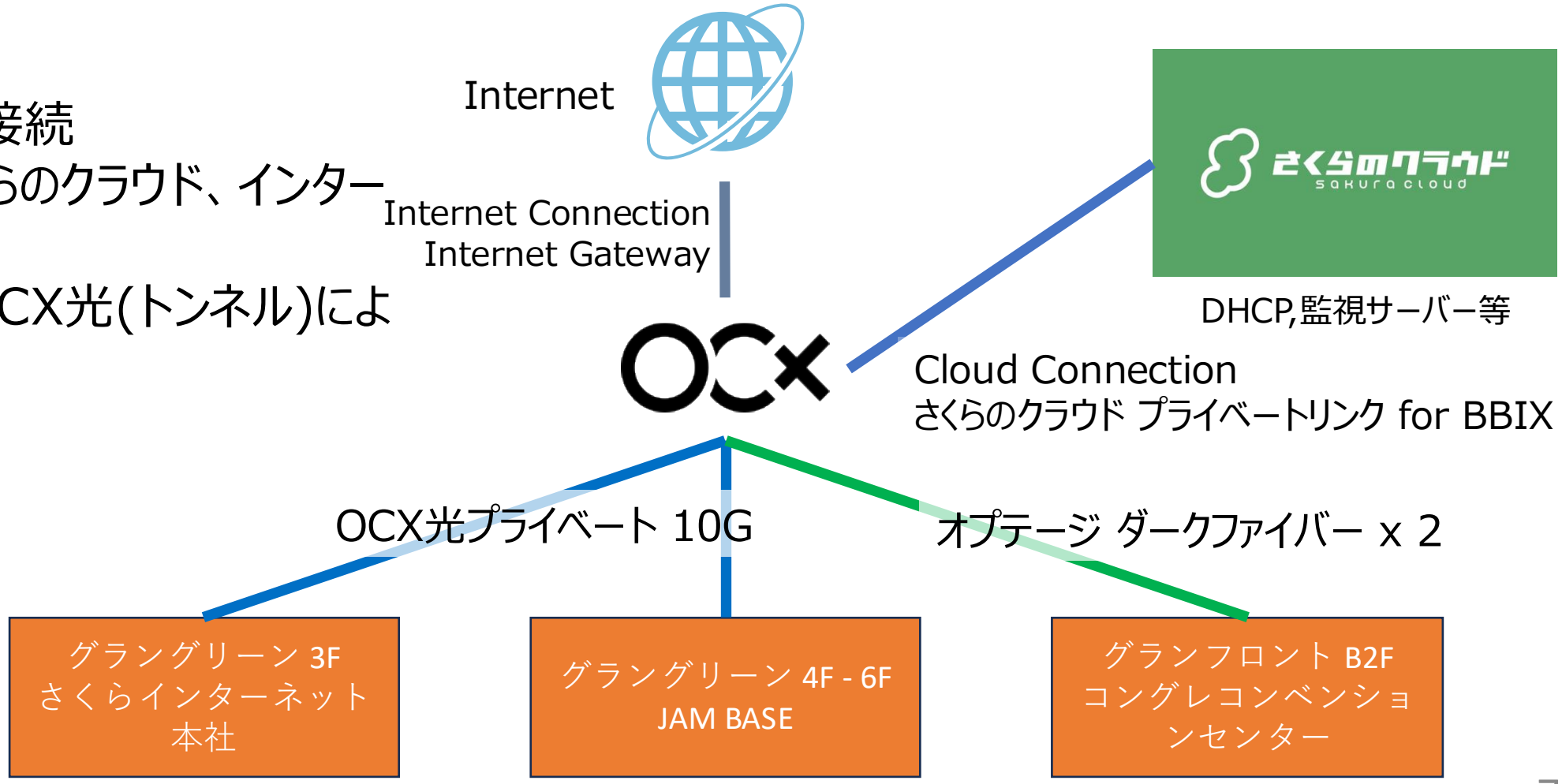
JANOG57 in OSAKA

会場ネットワーク構成紹介

2/5現在の構成です
本番当日は変更になる可能性があります

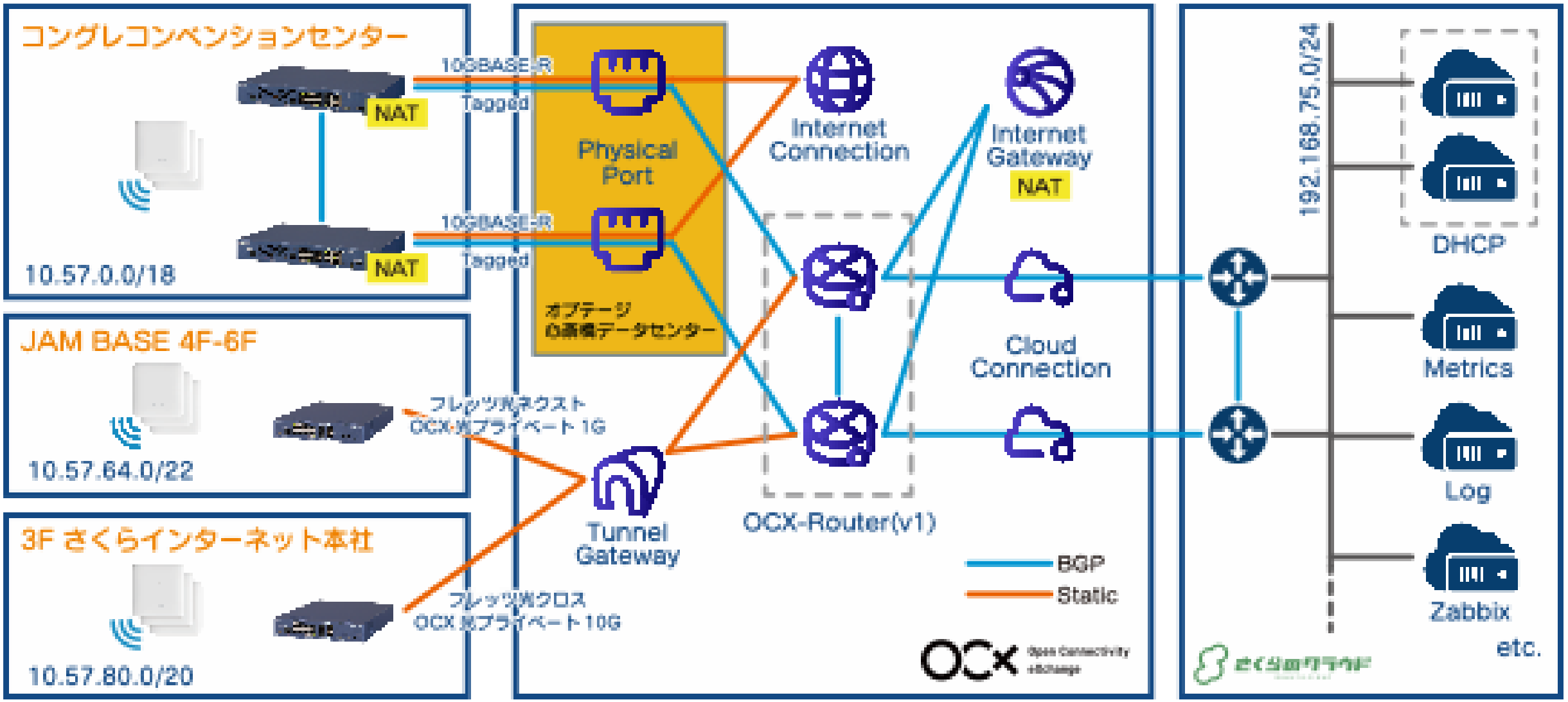
バックボーン

- 3会場分散
- 各会場をOCXに接続
- 拠点間接続、さくらのクラウド、インターネット接続を実現
- ダークファイバー/OCX光(トンネル)により接続



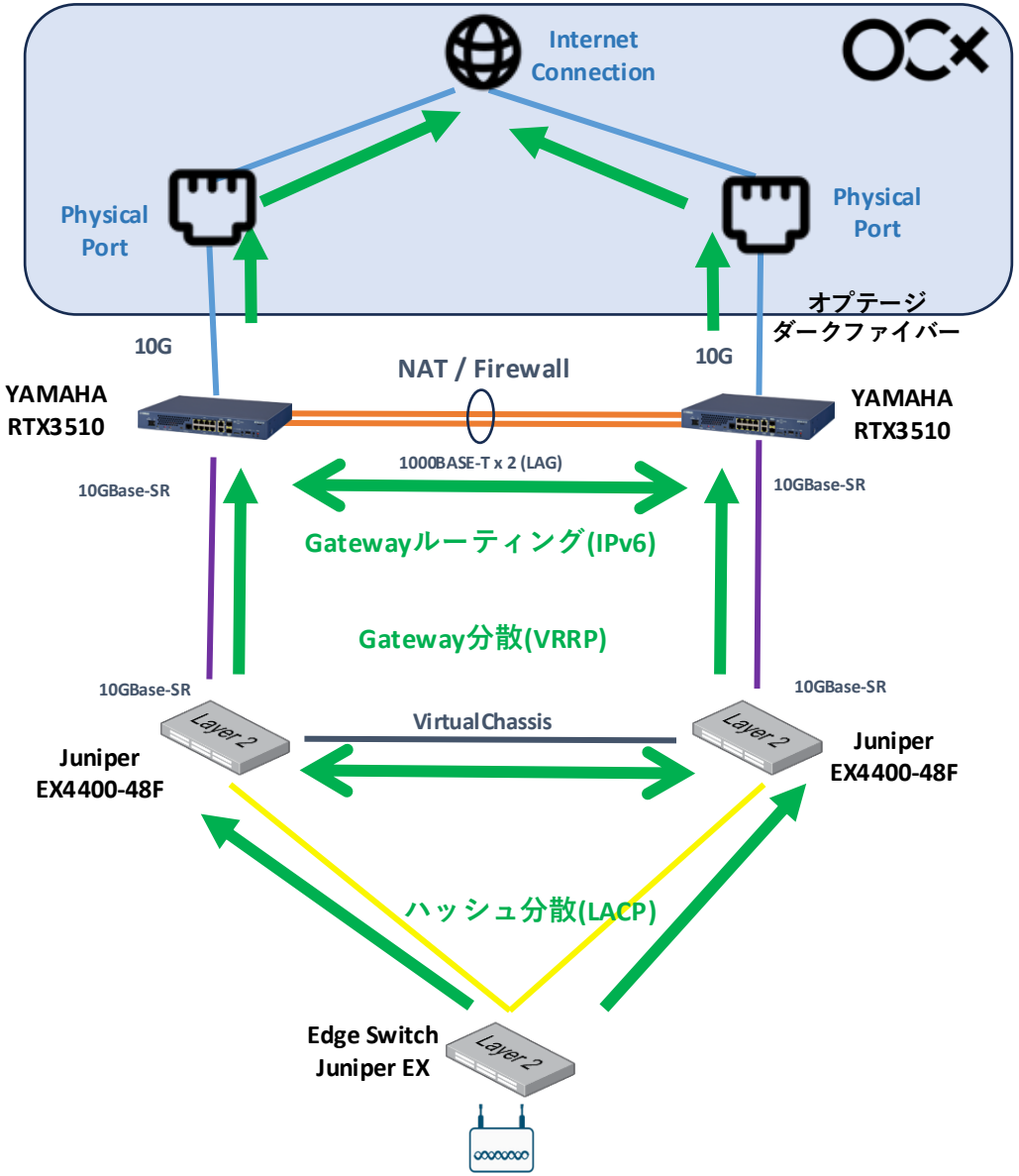
※IPv6 はフレッツ光クロス IPoE (BBIX)

バックボーン概要図


コングレコンベンションセンター
NAT
NAT
10.57.0.0/18
10GBASE-R
Tagged
10GBASE-R
Tagged
Physical Port
オプテージ
心斎橋データセンター
JAM BASE 4F-6F
10.57.64.0/22
フレッツ光ネクスト
OCX 光プライベート 1G
3F さくらインターネット本社
10.57.80.0/20
フレッツ光クロス
OCX 光プライベート 10G
Tunnel Gateway
Internet Connection
Internet Gateway
NAT
Cloud Connection
OCX-Router(v1)
BGP
Static
192.168.75.0/24
DHCP
Metrics
Log
Zabbix
etc.

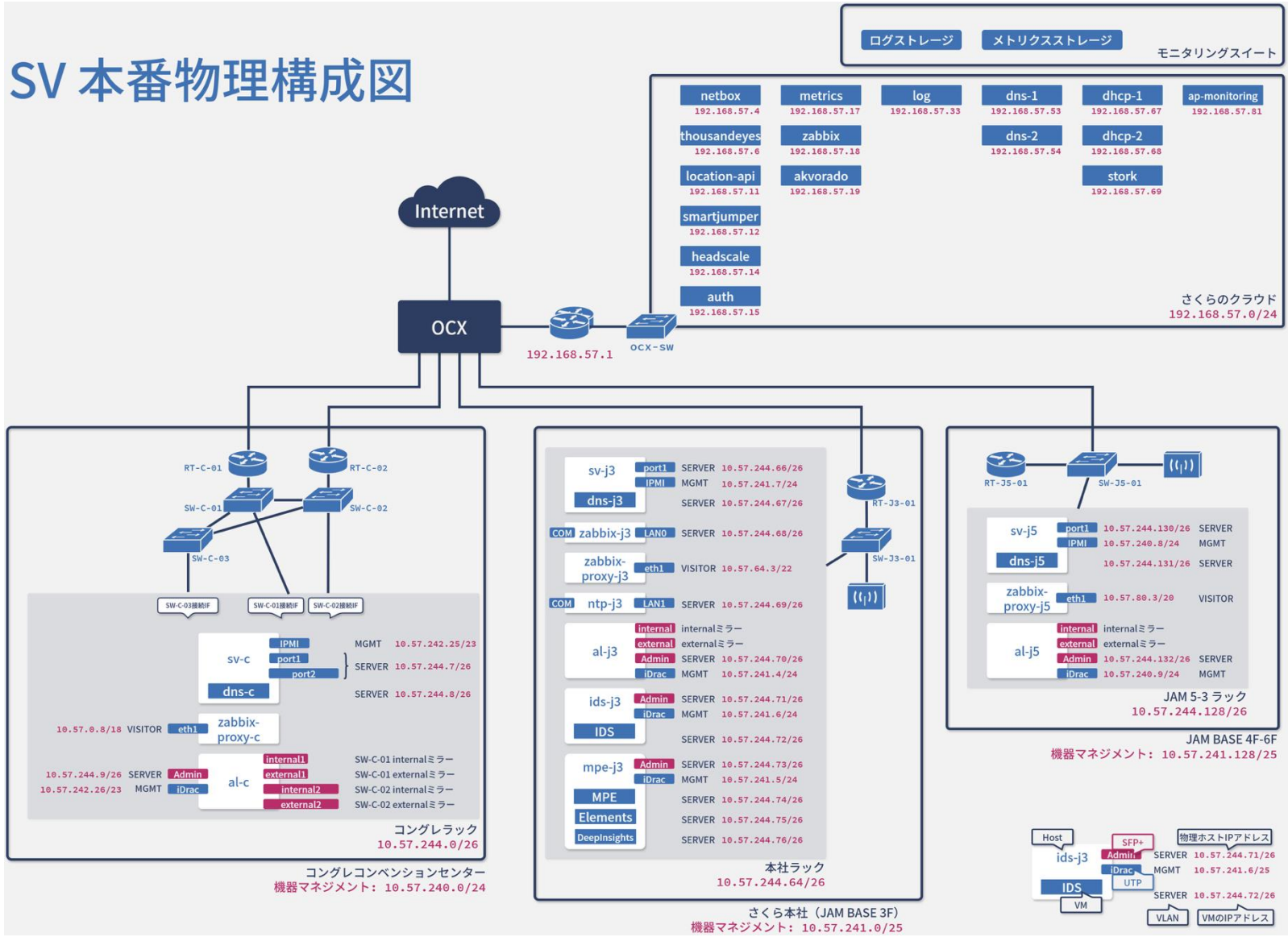
コングレ ネットワーク構成

- RTX3510 2台による冗長・負荷分散
- IPv4
  - ゲートウェイをVRRPによる冗長化
  - 配布するゲートウェイアドレスを変えることで負荷分散
- IPv6
  - マルチプレフィックスによる負荷分散
  - 相互監視+RA失効による冗長化
- コアスイッチ
  - Juniper EXスイッチ 28台
  - Virtual Chassisによる冗長化
  - 下位スイッチへLACPによる冗長化
- AP
  - Juniper Mist AP 80台



サーバー構成

- 物理サーバー＋さくらのクラウドを組み合わせたハイブリット構成
- DNS
  - 各会場に設置したサーバーを拠点間で順次参照
  - dnsdistによる負荷分散
  - Knot Resolver/Unbound
  - DoHに対応
- 監視
  - Prometheus
  - Grafana
  - Zabbix
  - Mist System



議論可能なトピック

議論可能なトピック

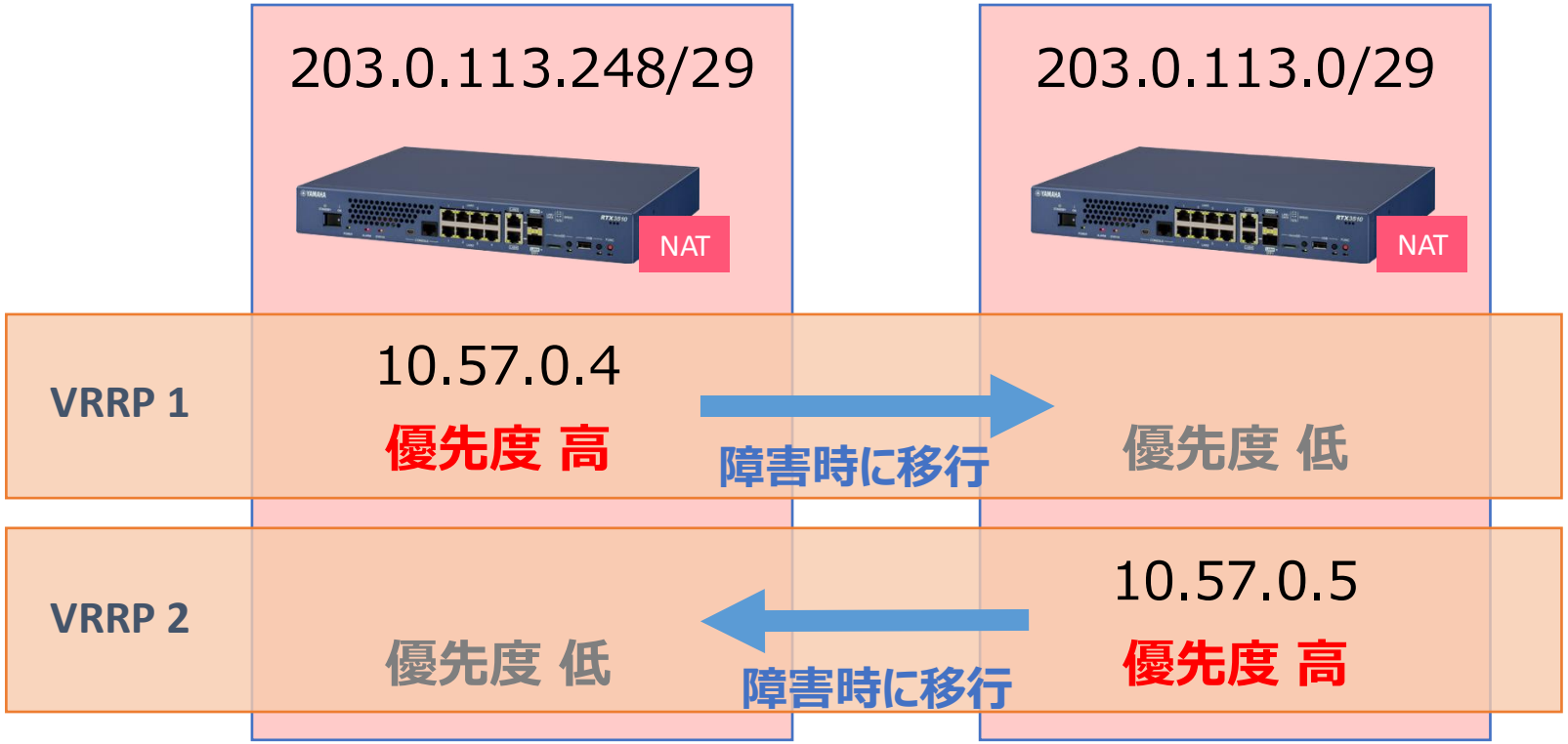
JANOG57 in OSAKA

議論テーマの候補をいくつか挙げます

議論したいテーマは是非Slidoへ投稿を！

- IPv4の冗長化とトラフィック分散
- IPv6の冗長化とトラフィック分散
- IPv6マルチキャスト設計
- DNS冗長化手法
- 構成管理ツール
- データを元にした帯域制限の必要性
- 長距離光ファイバ敷設の工夫

IPv4の冗長化とトラフィック分散



奇数MACアドレスに対し配布

偶数MACアドレスに対し配布

- VRRPにより冗長化したゲートウェイを2つ用意
- DHCPサーバーで2つの異なるゲートウェイを分散して提案
- 冗長化とトラフィック分散を達成
- 上流プロバイダとの接続の死活をVRRPステートに反映

クライアントのMACアドレス末尾の偶奇により配布するゲートウェイを分散

＋：2台のDHCPサーバーを独立させることができる

＋：ゲートウェイを確実に半々で分散することができる

＋：DHCPサーバーが片方落ちてもgatewayの分散は残る

－：トラシュー時どちらのgatewayを通ったかの確認が面倒

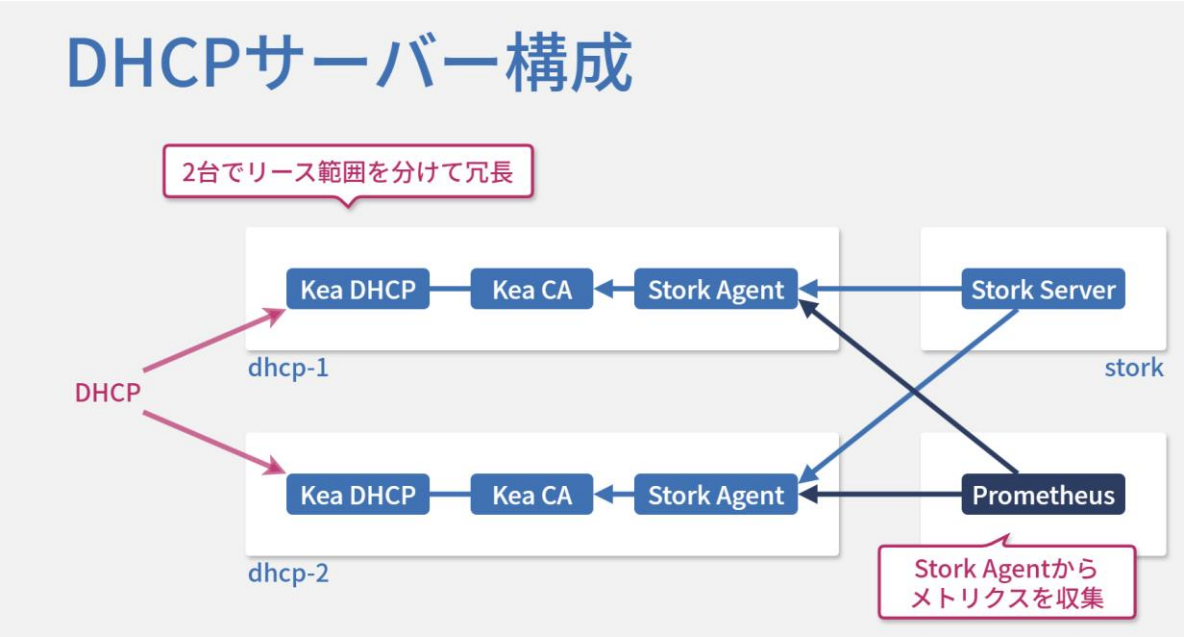
他に可能な方法

2台のDHCPで別のgatewayアドレスを通知

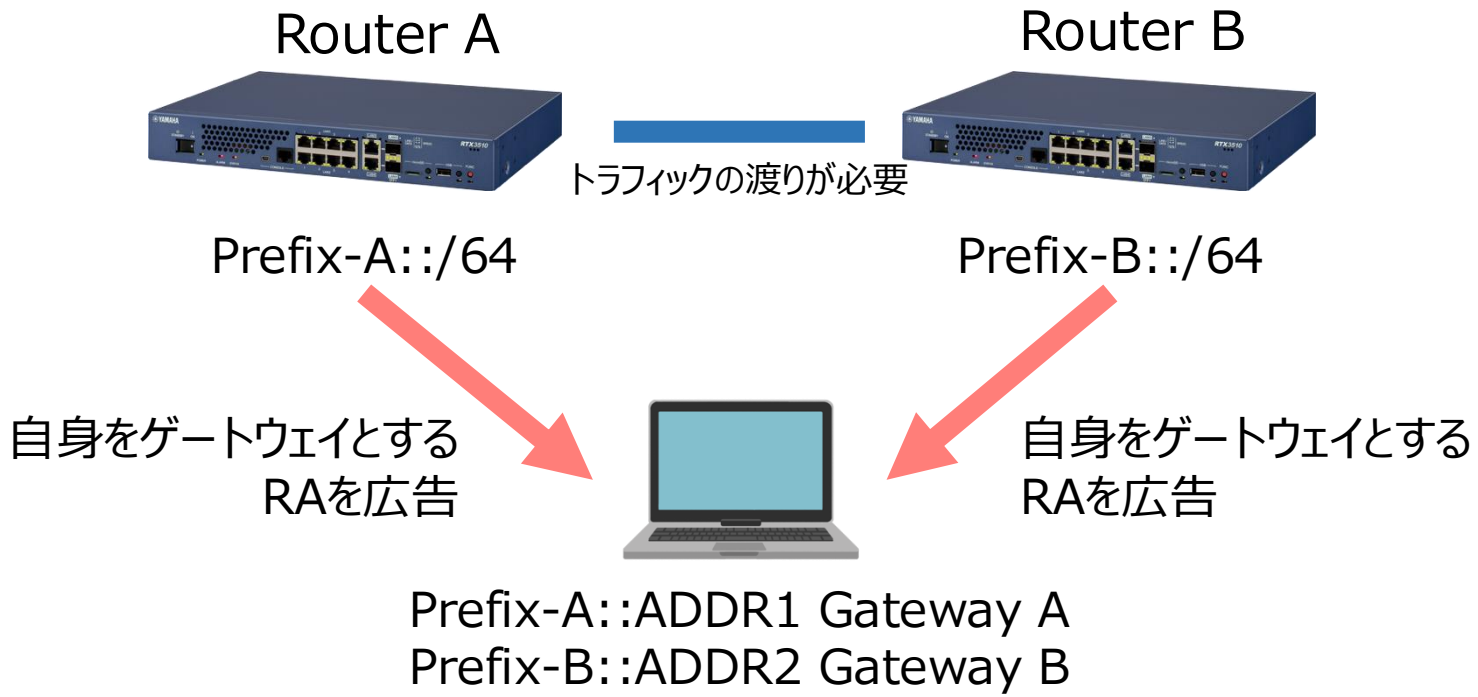
- 検証時に偏りが大きかった(7:3)

Kea HAのLoadBalancingモード

- 2台を独立させたかった



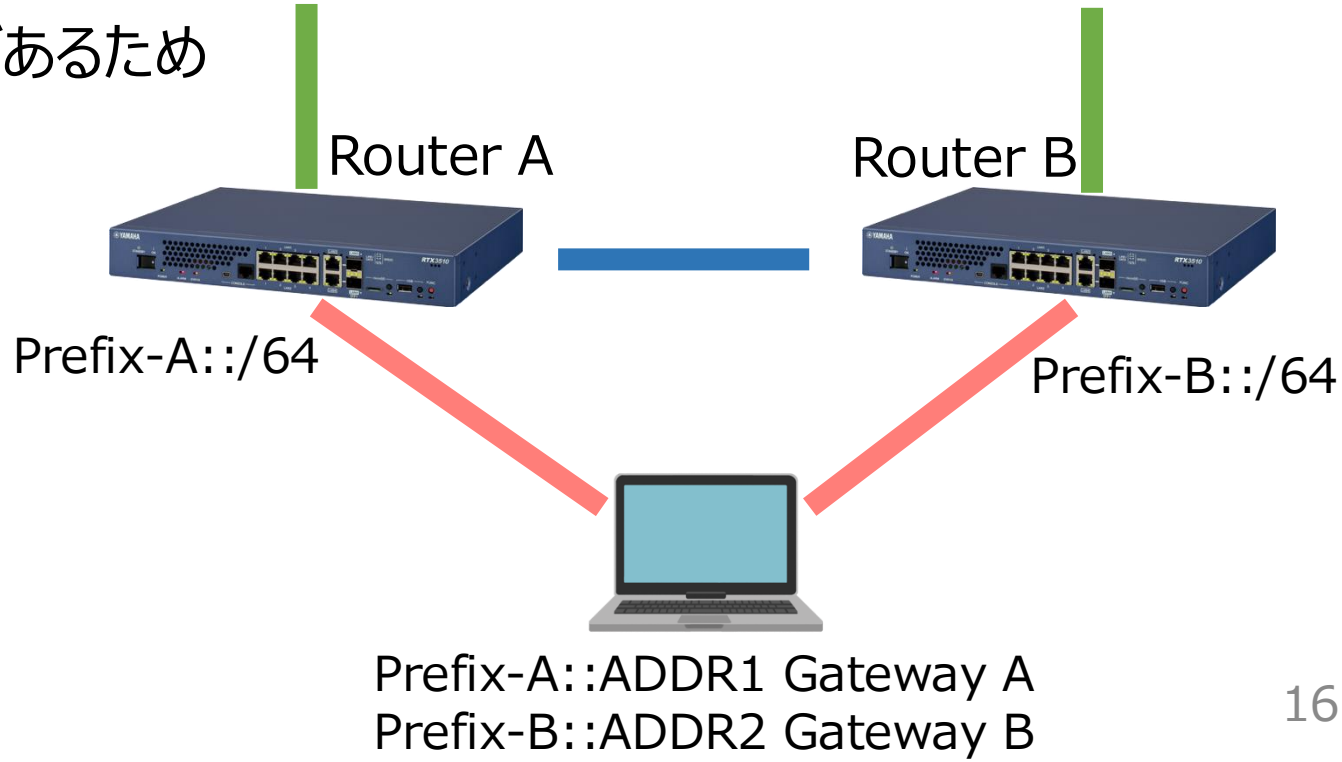
IPv6の冗長化とトラフィック分散



- IPv6のマルチホーミングによる冗長化とトラフィック分散
- 2台のルーターから異なるプレフィックスのRAを広告
- クライアントには通常時2つのプレフックスとゲートウェイを提供しSource Addressを選択させる
  →実際にどのくらいトラフィックが分散されるのかは当日公開予定
- 相手方に異常がある場合は自身のPrefixに加えLifetimeを0秒とした相手のPrefixと、自身のGateway Preferenceを高めたRAを送出
- Prefixとゲートウェイの組み合わせは必ずしも同一機器にはならない

課題: Prefixとゲートウェイの組み合わせは必ずしも同一機器にはならない

- IPv6 Source AddressとGatewayは独立して選択される
- Source AddressをPrefix-A::ADDR1、GatewayをBとすることが可能
- 上流プロバイダーがuRPF等のBCP38を実装している場合通信不能になる
- 上り下りで異なる経路を通る可能性があるため
  Stateful Firewallがかけられなくなる



解決策①: クライアントでSource Address Routing

- Router Aから割り当てられたPrefixをSource AddressとするときはGatewayを Router Aとする
- RFC8028に”SHOULD select”という記載はあるがクライアントに実装がないのが現状
- 問題視するInternet-Draftは複数提出されているがいずれも進展なし
- 小規模なネットワークでしか需要がないため進まないのでは

解決策②: ルーターでNAT66/NAPTv6を行う

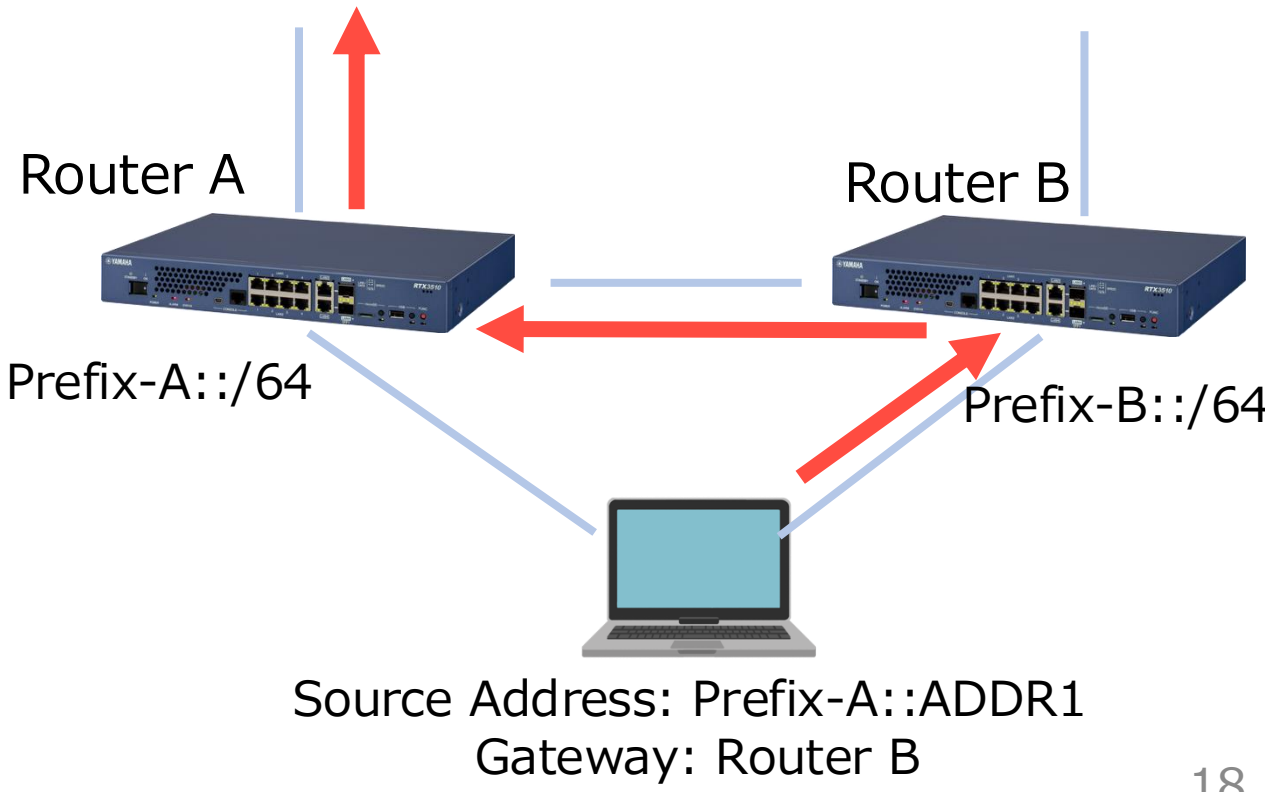
- 正しいPrefixへ変換され、上り下りの経路が固定される
- 特にステートレスなNAPTv6はセッション管理不要で有用
- 実装機器は高価な機種が多い

解決策③: PrefixごとにVLANを分割しクライアントに一意に割り当てる

- VLANを分割し別のゲートウェイを提供
- クライアントごとに割り当てるVLANを決定しローミングも可能
- ベンダー実装依存な部分

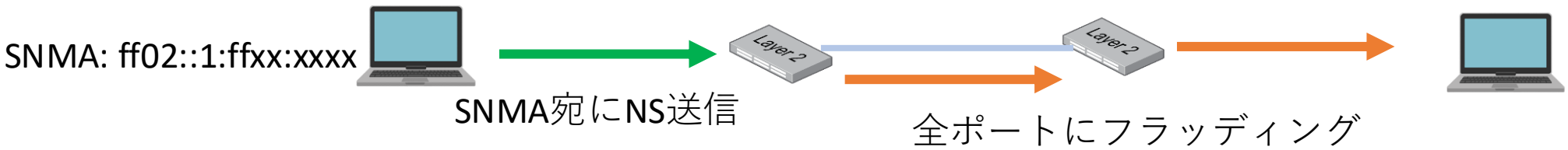
解決策④: ルーター間を接続して正しいGatewayを通るようにルーティング

- Router Aから割り当てられたPrefixをSource AddressとしたパケットがRouter Bに来た場合はRouter Aへ転送する
- **今回採用**



IPv6マルチキャスト設計

- 多数のクライアントを1つのサブネットで収容する場合、全クライアントの**Solicited-Nodeマルチキャストアドレスが全スイッチに転送**されL2MSテーブルに学習される
- エッジスイッチはL2MCテーブル上限が小さいものもあり、**MLD Snooping**を有効化した場合テーブルが溢れる危険性が高い
- MLD Snoopingを無効化することによってL2MCテーブルは作成されなくなるので溢れを考える必要性がなくなる
- SNMA宛の全NS/NAが全クライアントへフラッディングされるため、Airtimeが逼迫する可能性がある
  →Basic Rateを高めに設定して低レートを無効化 & IsolationでBUM抑制をする



DAD実行時の挙動

DNS冗長化手法

BGP Anycast

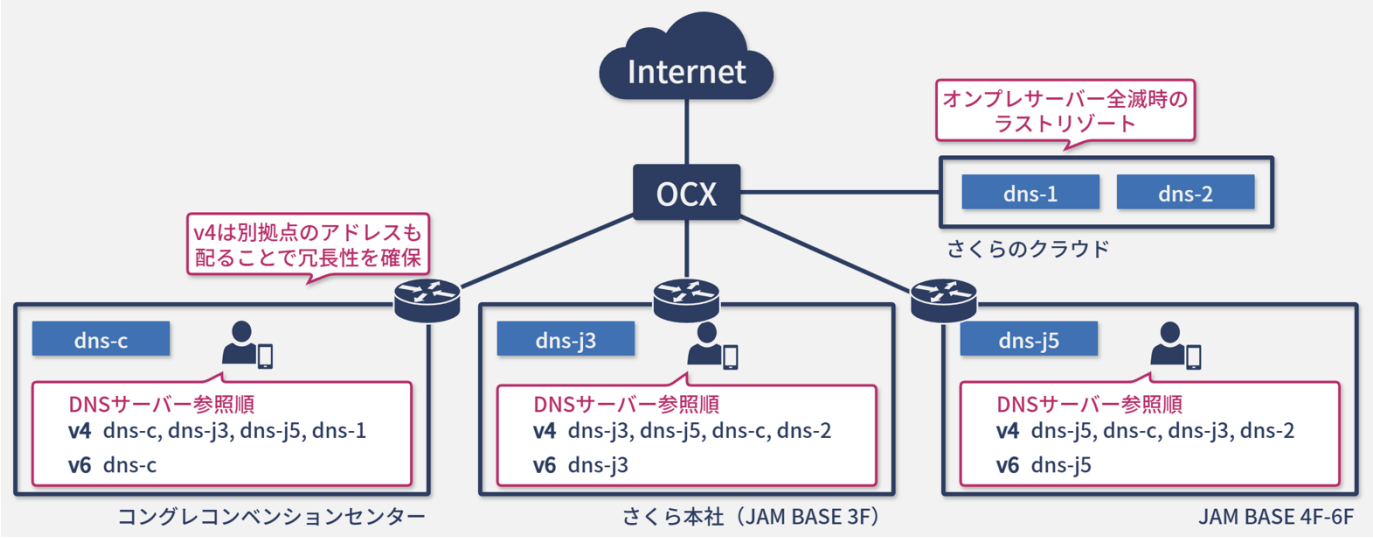
- 同一アドレスを複数拠点から広告
- イベントネットワークでは運用が困難

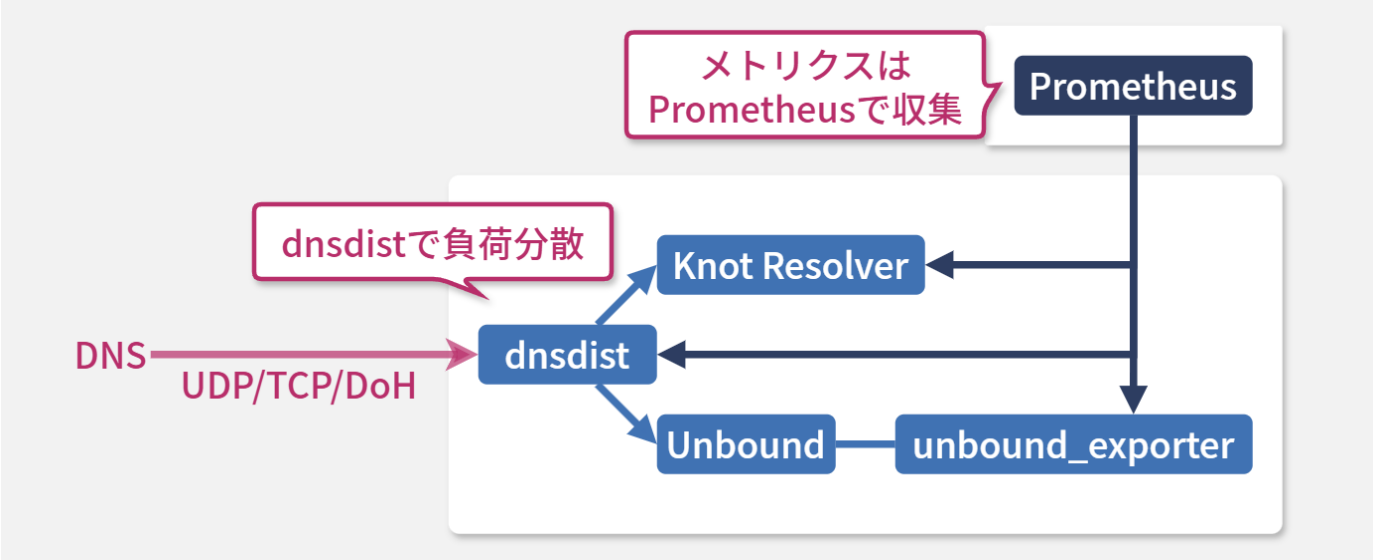
VRRP

- 複数のDNSサーバが仮想IPを共有
- フェイルオーバーの挙動が明確

複数DNS配布(今回採用)

- DHCPで複数のDNSサーバーを配布
- クライアントのフォールバック機構に依存





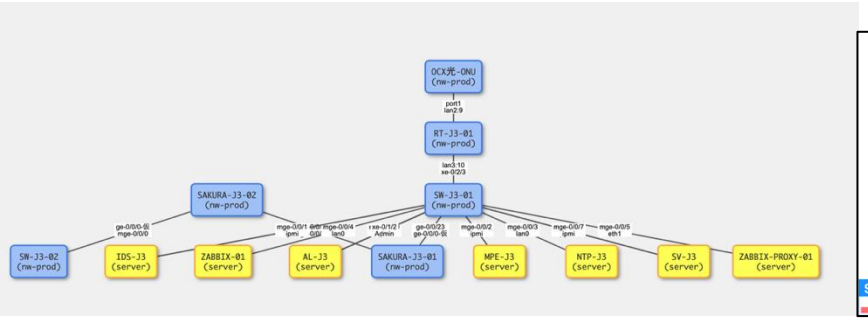
機器構成・コンフィグ把握の課題

- 特性上接続関係やコンフィグが頻繁に変更される
  →都度IPAMや構成図、配線図を手動で更新する必要性

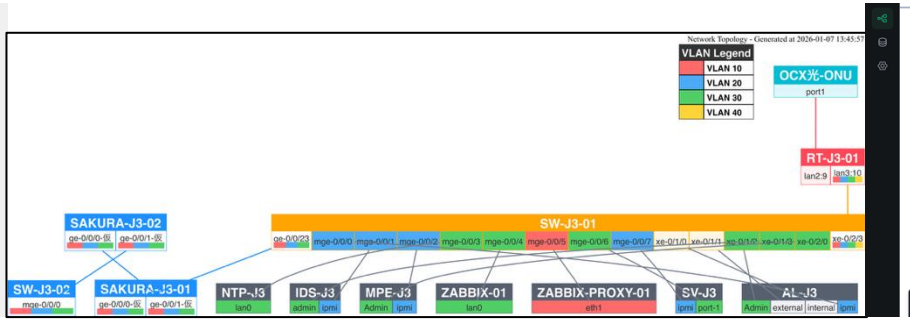
機器情報をNetBoxに集約し構成図は自動生成される仕組みの開発

- 既存ツールは”物理接続”に焦点が当てられていない印象
- インターフェイス同士が接続されていることを分かりやすく表現できない

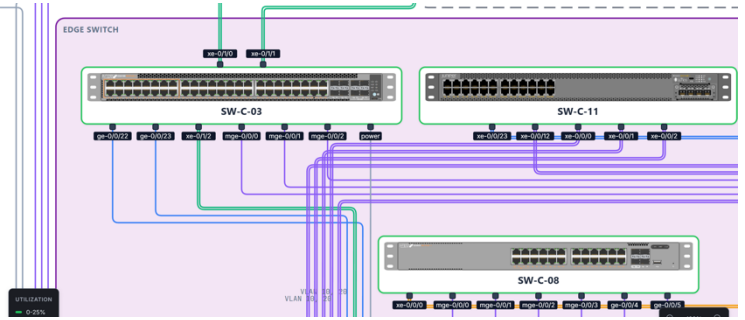
→イベントネットワークに最適な可視化ツールを作ってしまおう！



vis.js



Graphviz



自作ツール(shumoku)

機能

- 接続関係がわかりやすい構成図の生成
  - フロア単位の可視化に対応
  - NetBox更新時に構成図も自動更新
- 監視機能の統合
  - トラフィック量を構成図にリアルタイム表示
  - 機器のアラートの表示
    →トポロジーベースで障害ポイントを可視化できる
- 外部連携
  - トポロジー図からNetBox/Grafanaへワンクリック遷移
- OSSプロジェクトとして公開
  - https://github.com/konoe-akitoshi/shumoku