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

自己紹介

吉川 知輝 (Tomoki Yoshikawa)

JANOG57 バックボーンチームリーダー

所属

- 京都大学大学院 情報学研究科 通信情報システムコース1年
- インターネット・セキュリティの研究

Home NOC Operators' Group 運営メンバー (AS59105)

- ピアリング
- バックボーンネットワークの運用と構築

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~20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
- 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

20分~70分：特定のテーマについての個別議論

- 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

- 2/7～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**
  - https://theme-submission.janog57.ishikari-dc.jp
- 事前に設定したテーマ+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
  - テーマに関連した議論をお願いします
- 複数の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JANOGerが入れ替わる
  - 議論に誰でも参加可能
- 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
  - https://drive.google.com/drive/folders/16mpSNXvjdY_eSTAhN4khVJrpBM1dNhKO?usp=sharing

リンク

議論テーマ募集サイト

https://theme-submission.janog57.ishikari-dc.jp

資料投稿用Google Drive

https://drive.google.com/drive/folders/16mpSNXvjdY_eSTAhN4khVJrpBM1dNhKO?usp=sharing

提案いただいた議論トピック

IPv6-Mostly Networkの提供について

- 今回は提供していません

IPv6を用意するかどうかについて

- 用意しています！詳細はこの後お話しします

IPv6にステートフルFWを入れるかどうか

- 入れています。苦労話は後ほど

ROV の implementation について

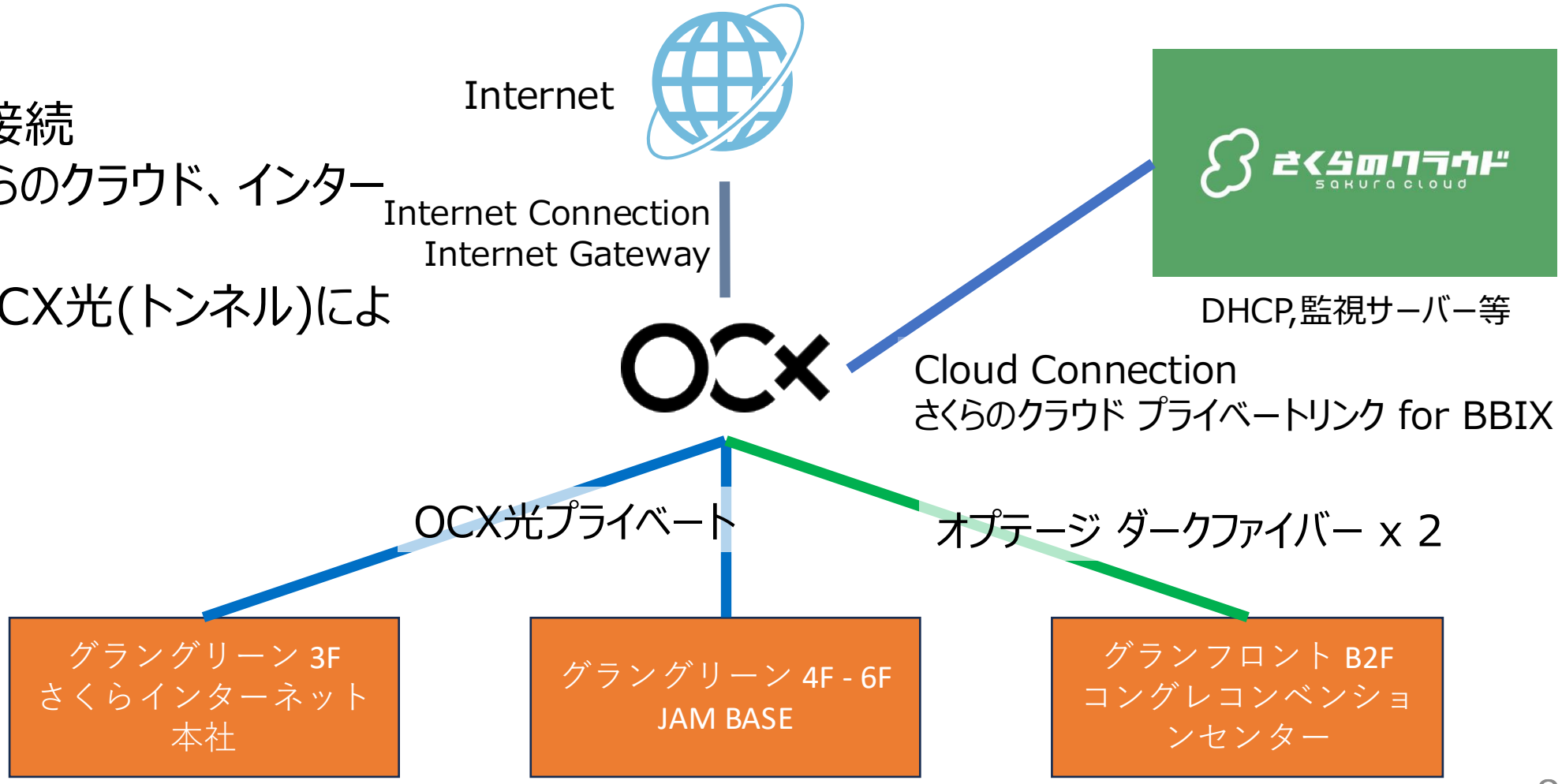
- 上流のSoftBank/BBIXが実施しているためROVが実装されています！

JANOG57 in OSAKA

会場ネットワーク構成紹介

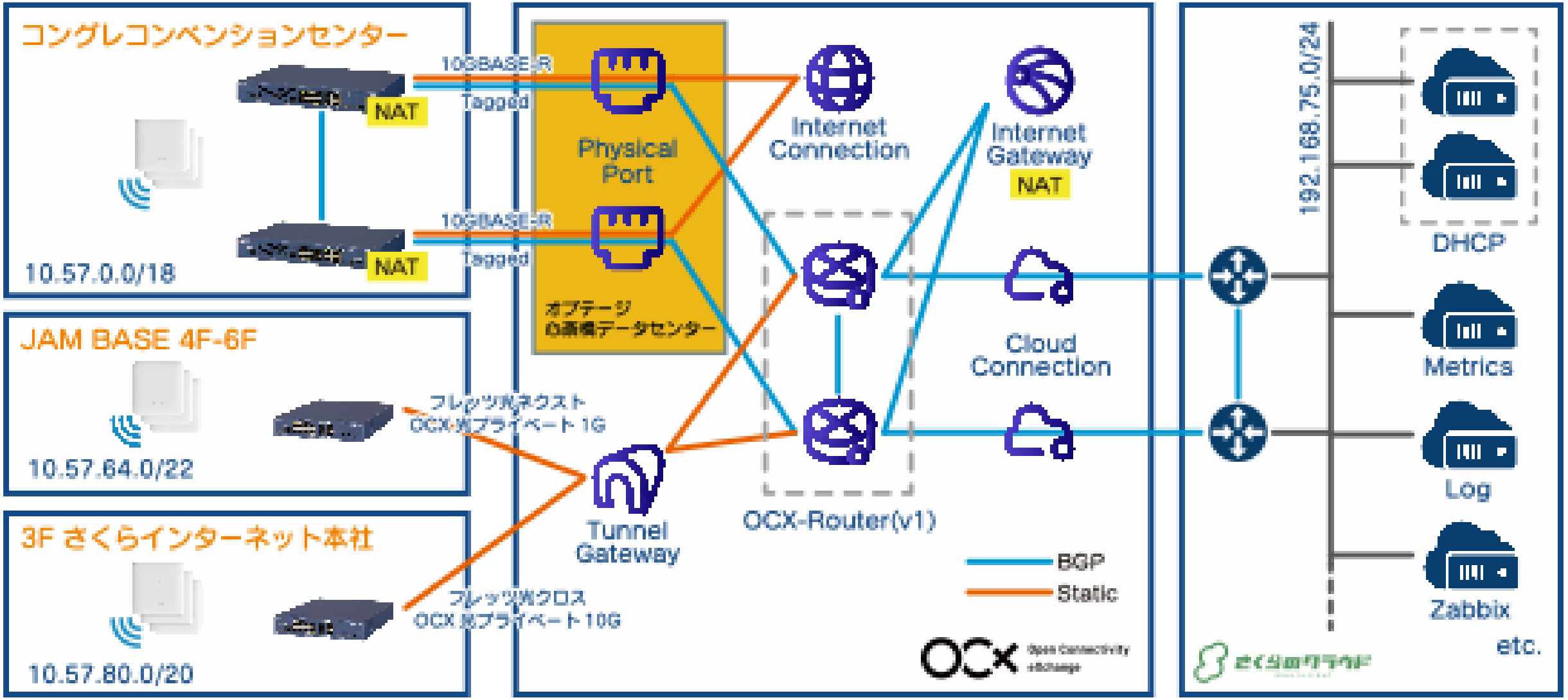
バックボーン

- 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



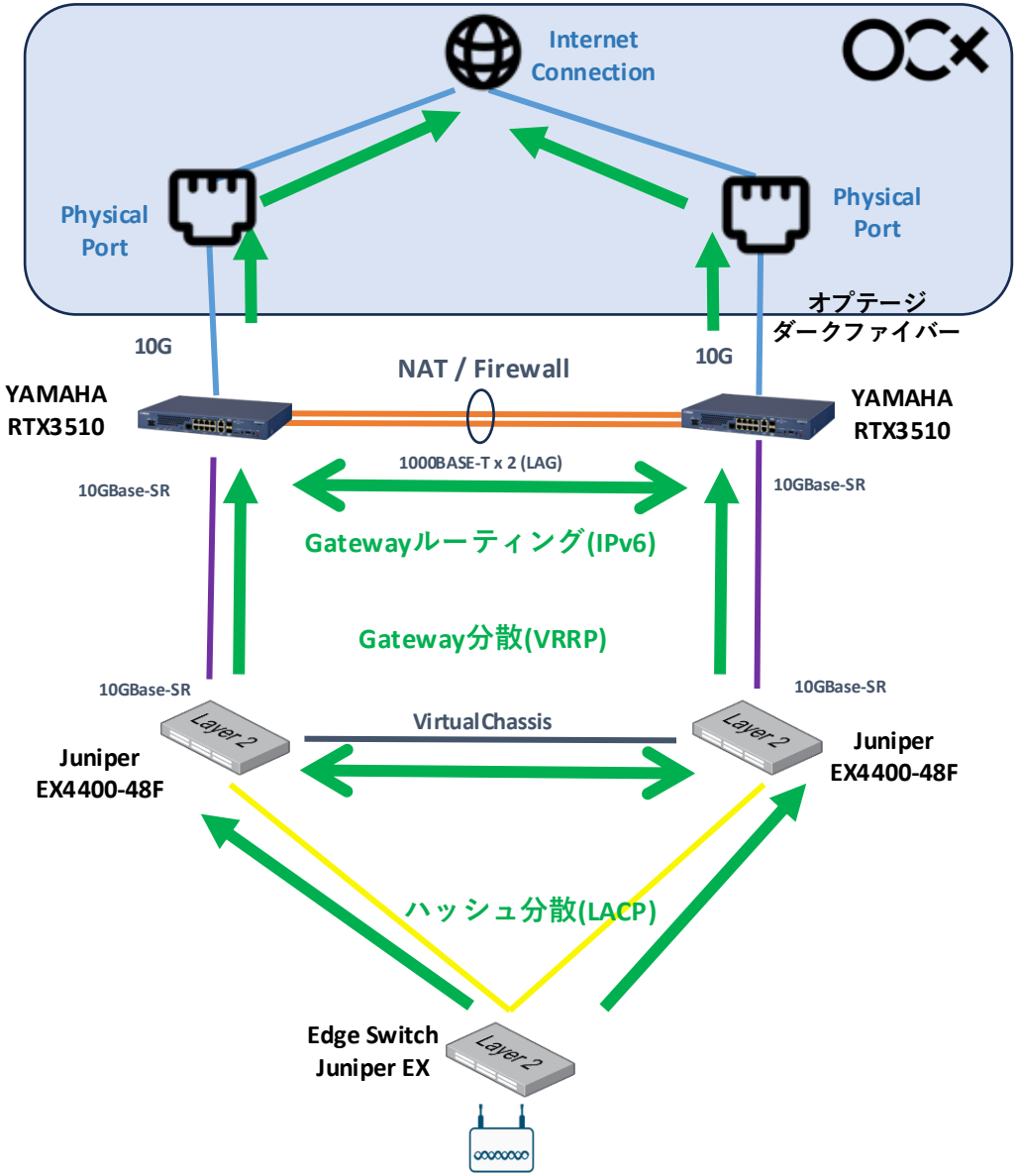
※IPv6 はフレッツ光クロス IPoE (BBIX)

バックボーン概要図


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

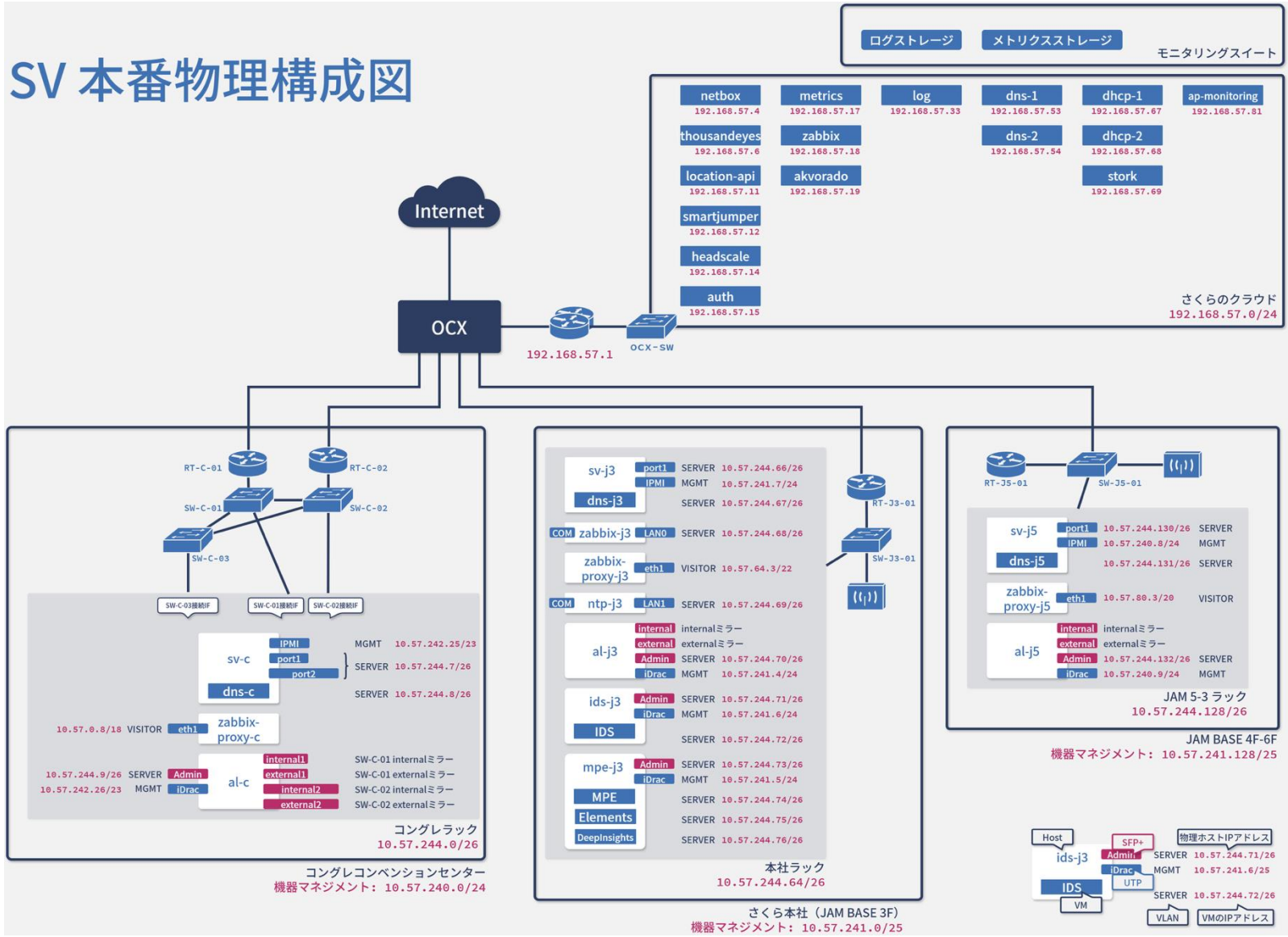
コングレ ネットワーク構成

- 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



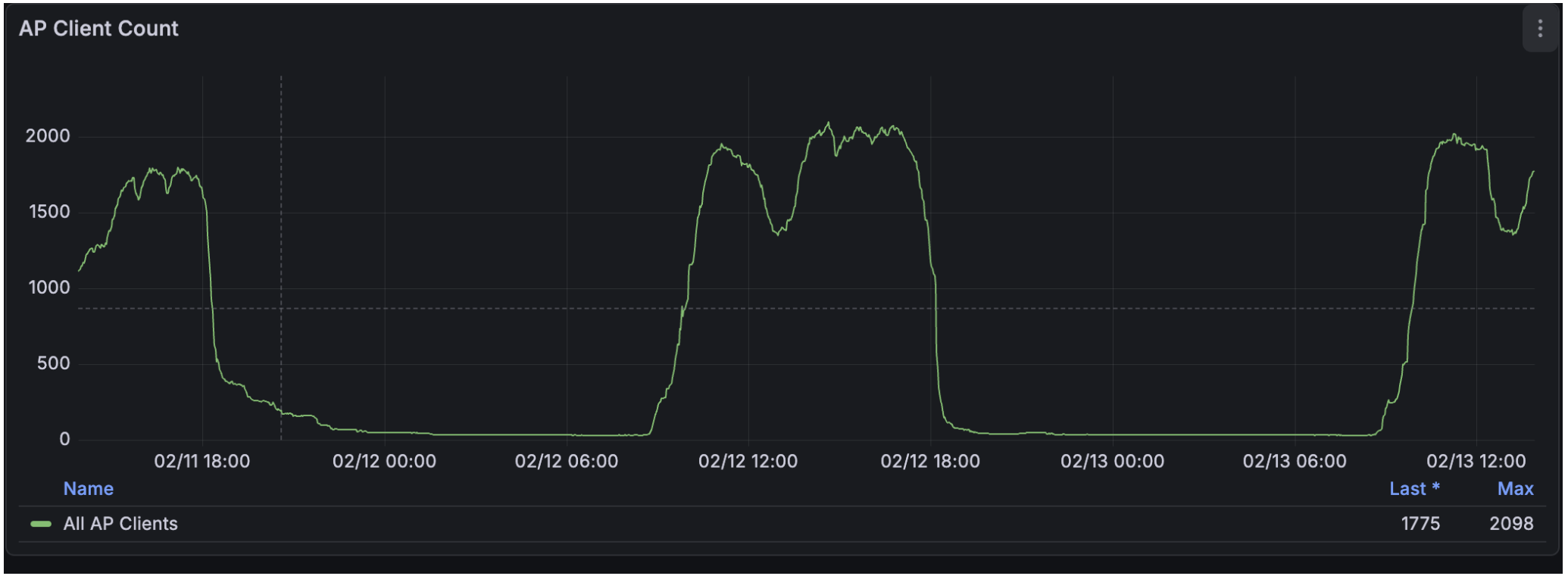
サーバー構成

- 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RTT順に参照
  - dnsdistによる負荷分散
  - Knot Resolver/Unbound
  - DoHに対応
- 監視
  - Prometheus
  - Grafana
  - Zabbix
  - Mist System



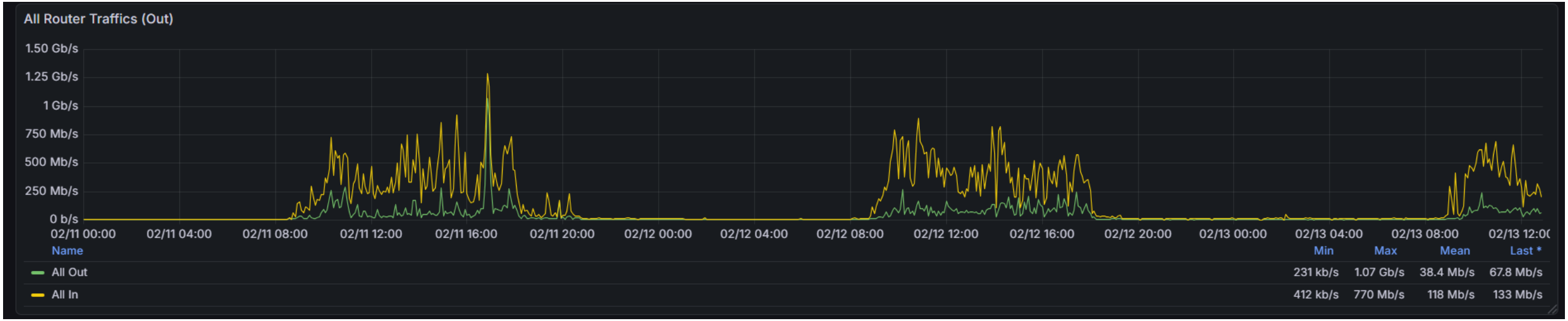
モニタリング結果

- クライアント数
  - 最大2098クライアント



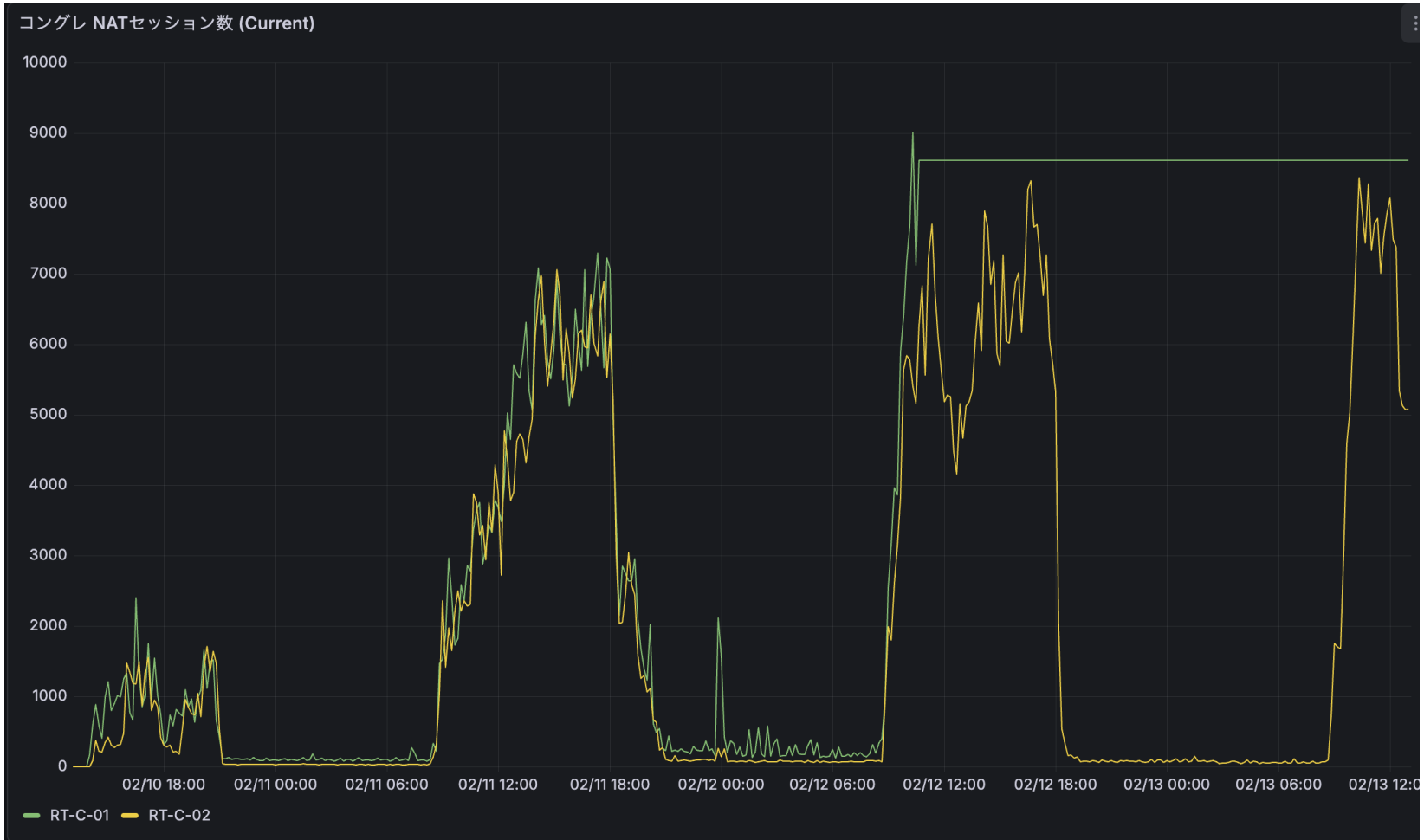
モニタリング結果

- トラフィック
  - 3会場合計で最大1.8Gbps



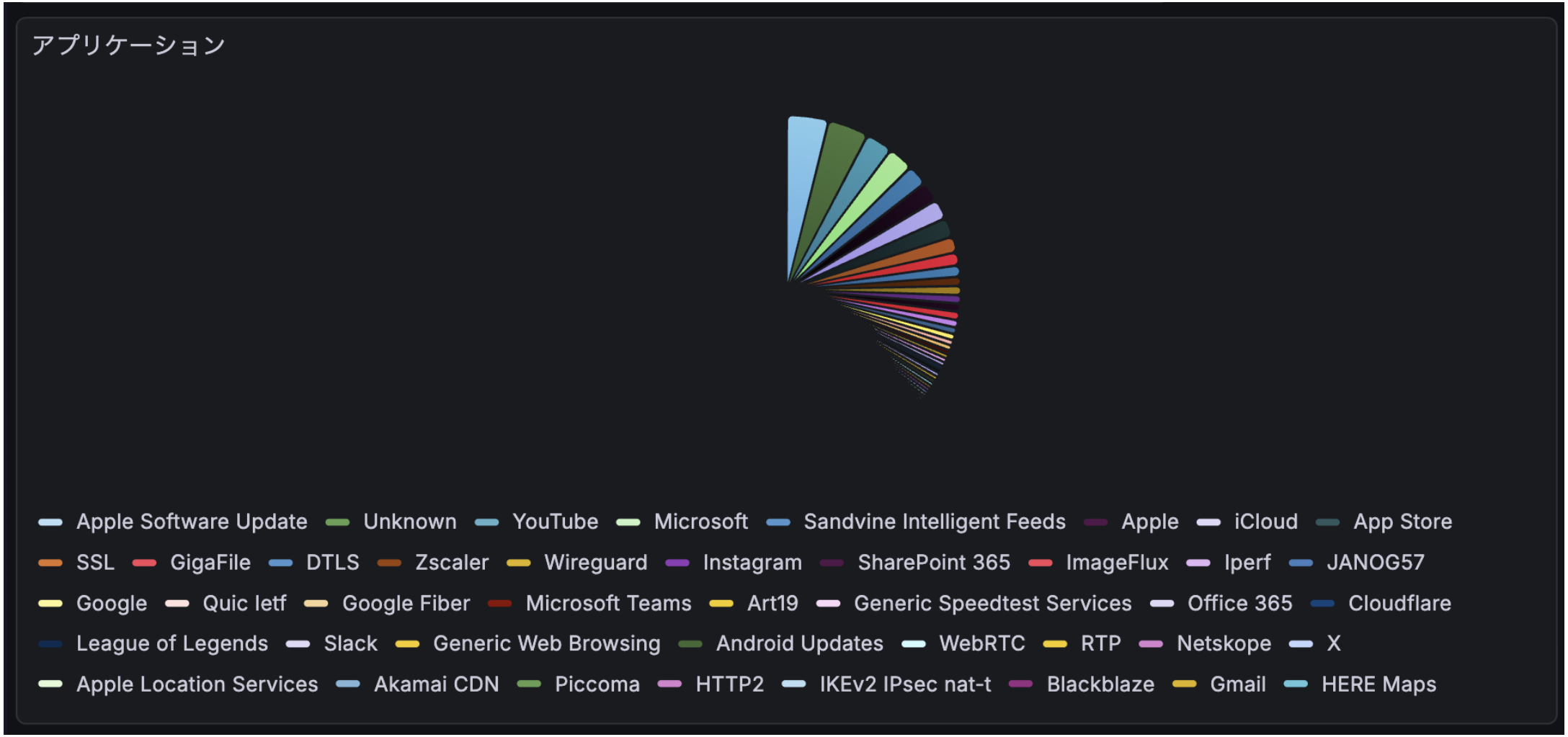
モニタリング結果

- NATセッション数
  - コングレ 最大17000セッション



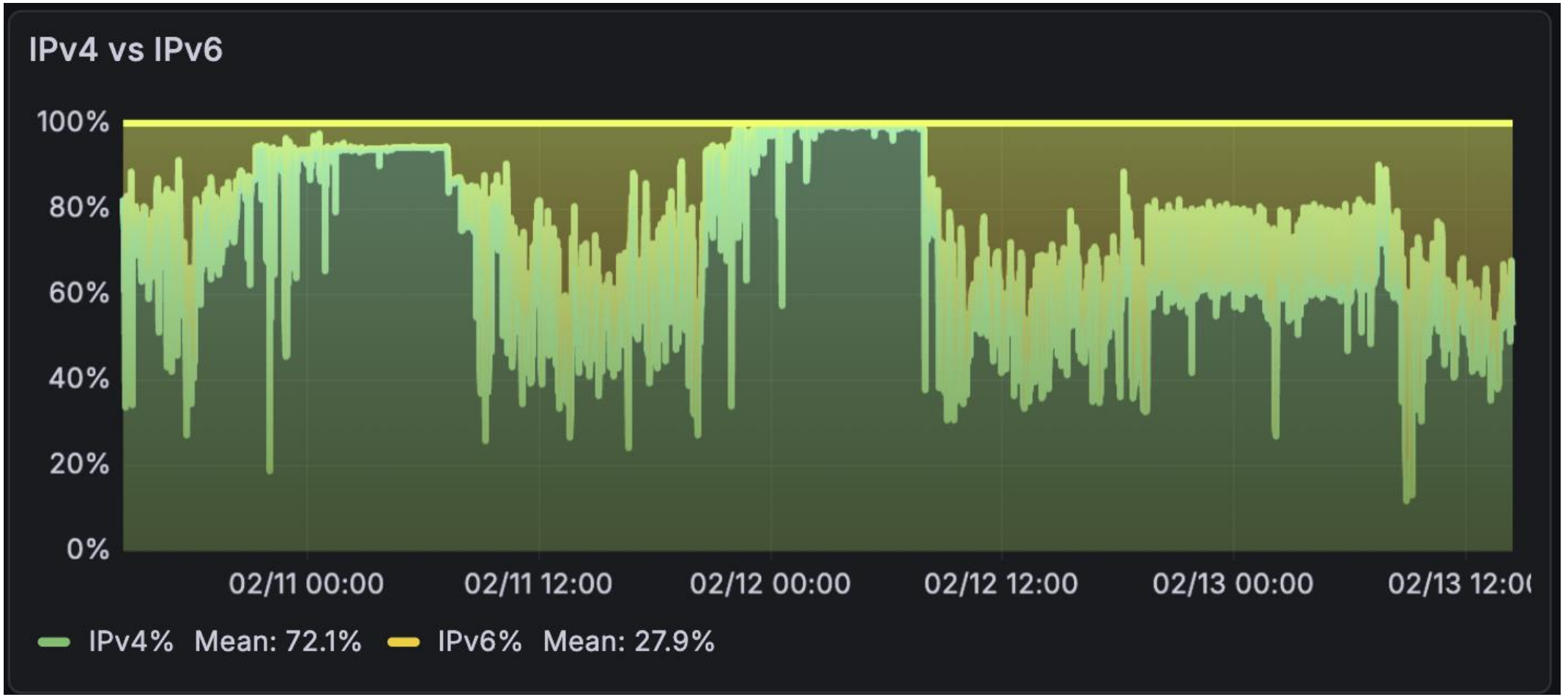
モニタリング結果

- 使用されたアプリケーション



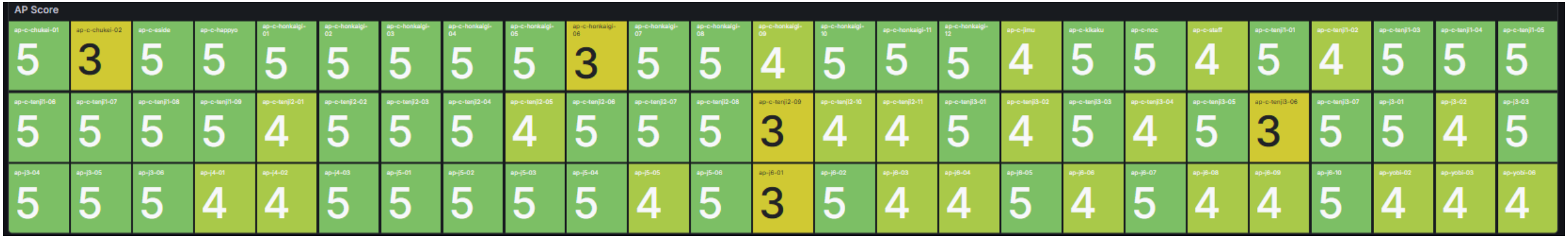
モニタリング結果

- IPv4 VS IPv6
  - 大まかに6:4程度



モニタリング結果

- Deep Packet Inspectionによるモニタリング
  - AppLogic Networksさんにご提供いただいているActiveLogicによる観測結果
  - 各APでの接続状況を1-5で評価



モニタリング結果

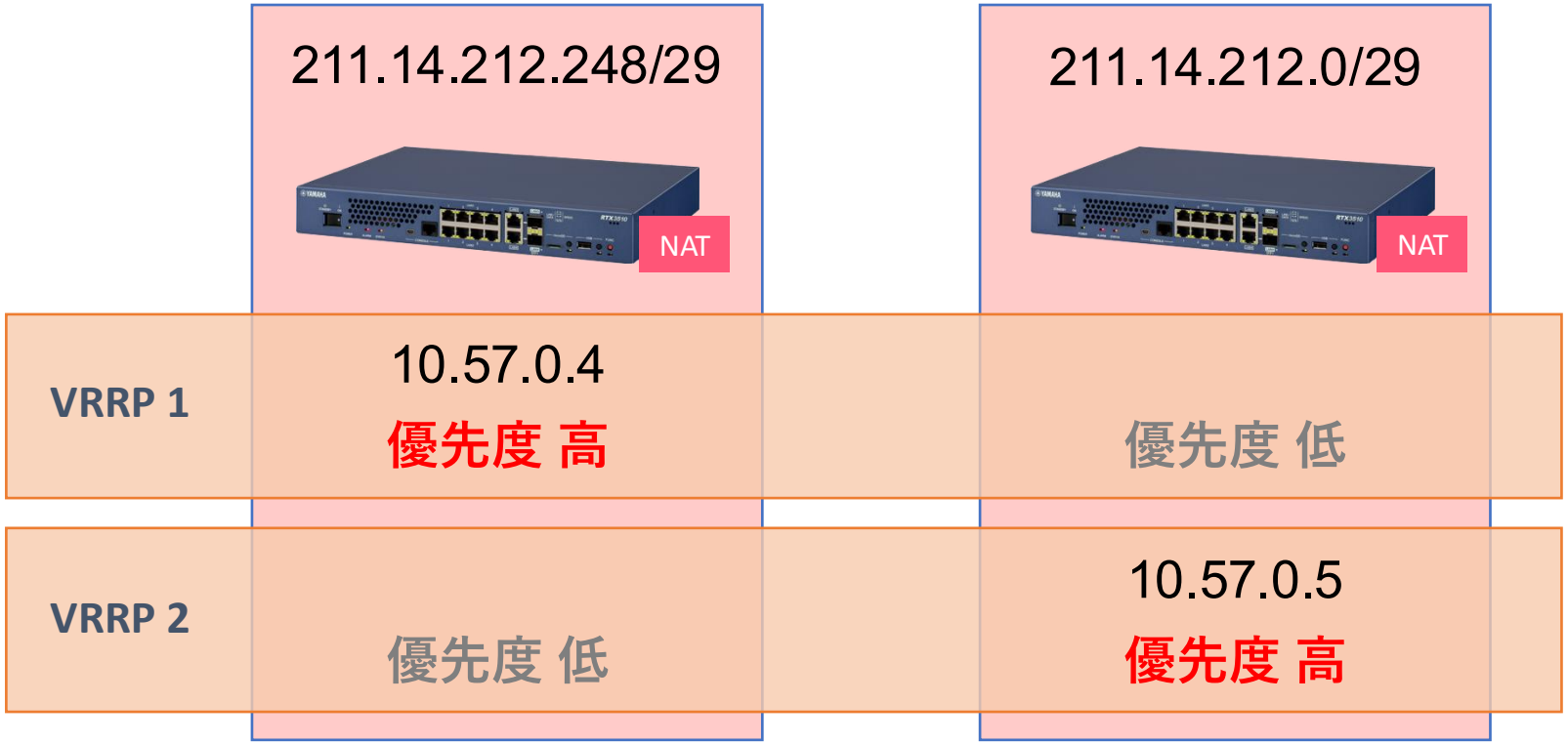
- ケーブル
  - LANケーブル
    - 距離　2,393m
    - 本数　119本
  - 光ファイバ
    - 距離　2,290m
    - 12本

議論トピック

議論トピック

1. IPv4の冗長化とトラフィック分散
2. IPv6の冗長化とトラフィック分散
3. IPv6マルチキャスト設計
4. DNS冗長化手法
5. 構成管理ツール
6. WindowsにおけるDHCPv6の挙動
7. 帯域制限の実施
8. eduroam/OpenRoaming
9. バーストDNSクエリ
10. ケーブル敷設計画

IPv4の冗長化とトラフィック分散



- VRRPにより冗長化したゲートウェイを2つ用意
- 2台のDHCPサーバで異なるゲートウェイのプールを用意し提案する
- 冗長化とトラフィック分散を達成
- 上流の死活をVRRPステートに反映させることに注意

IPv4の冗長化とトラフィック分散

＋：2台のDHCPサーバーを独立させることができる

＋：gateway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

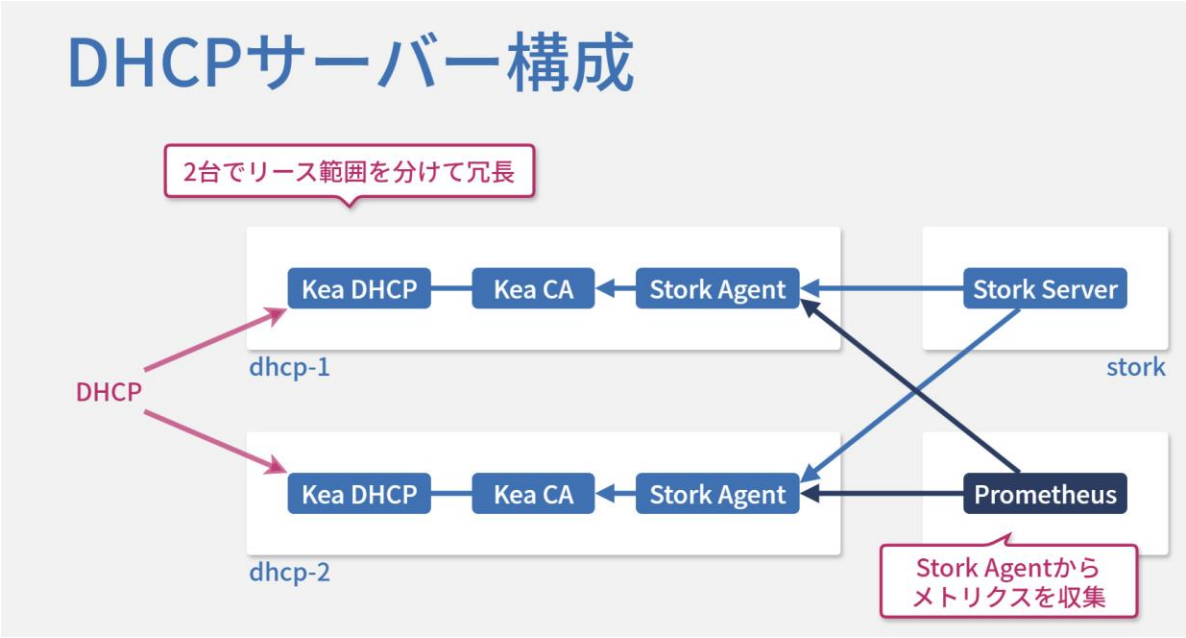
他に可能な方法

2台のDHCPで別のgatewayアドレスを通知

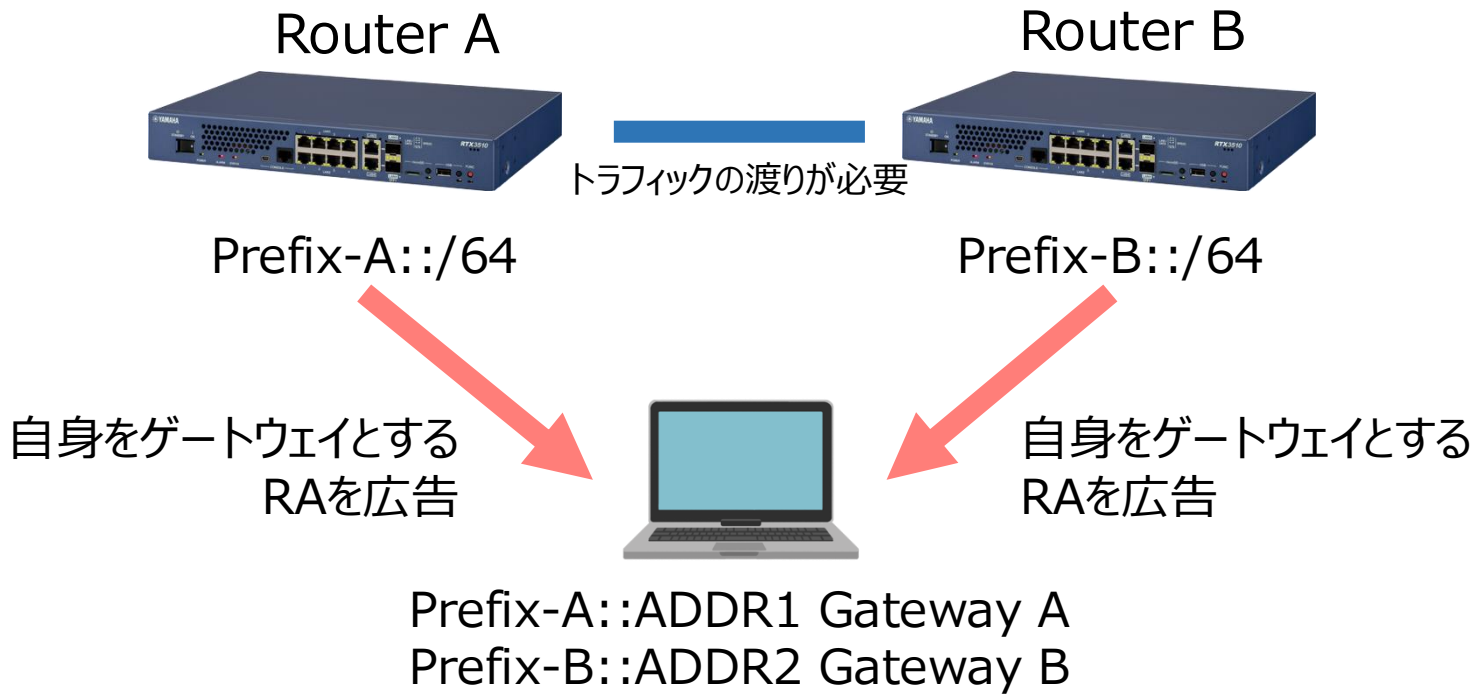
- 検証時に偏りが大きかった(7:3)

Kea HAのLoadBalancingモード

- 2台を独立させたかった



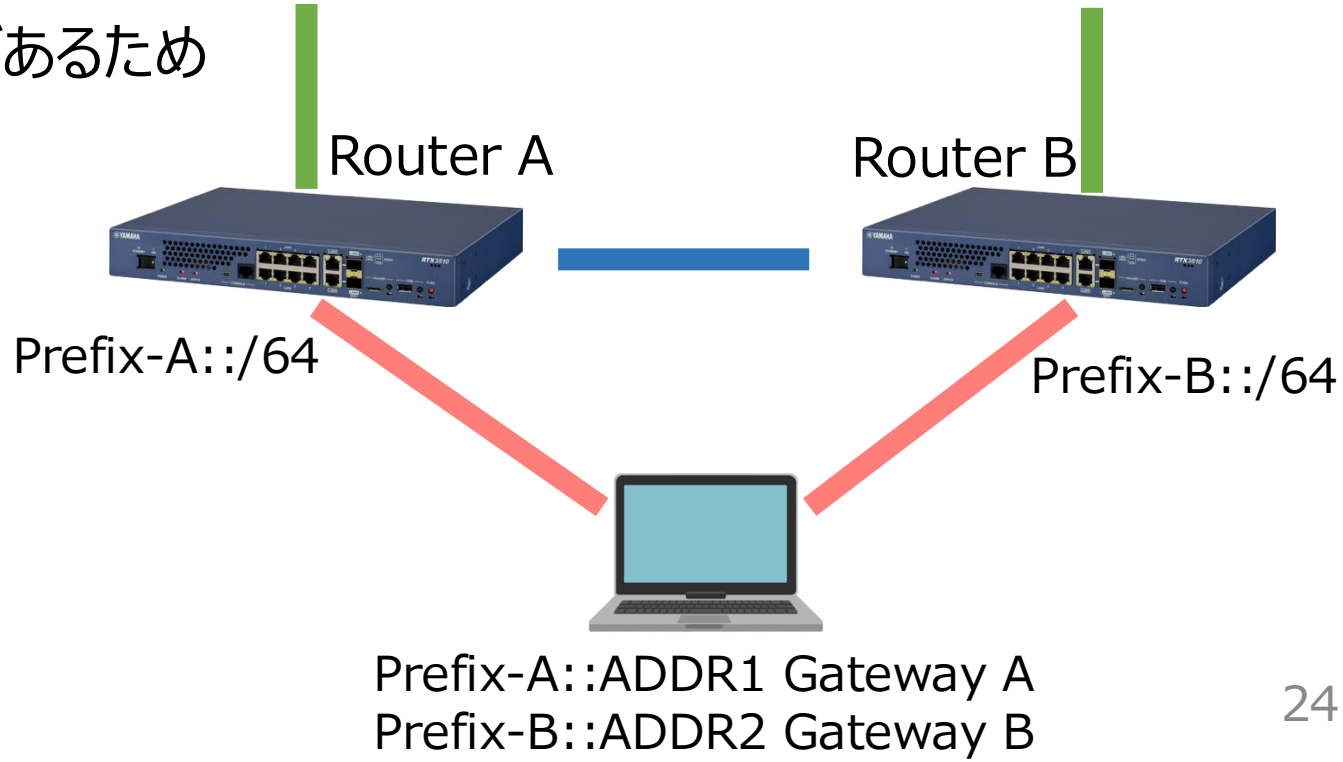
IPv6の冗長化とトラフィック分散



- 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

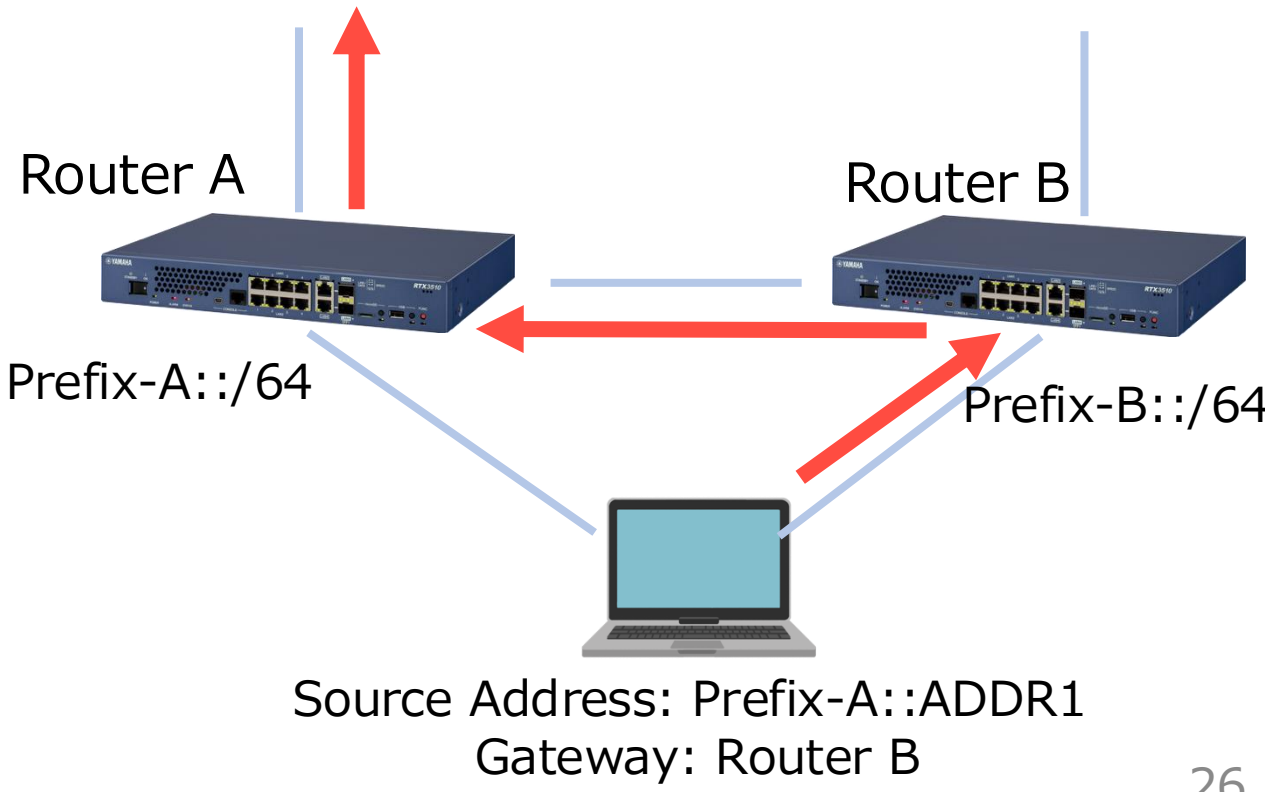
- 正しいPrefixへ変換され、上り下りの経路が固定される
- 特にステートレスなNAPTv6はセッション管理不要で有用
- 実装機器は高価な機種が多い

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

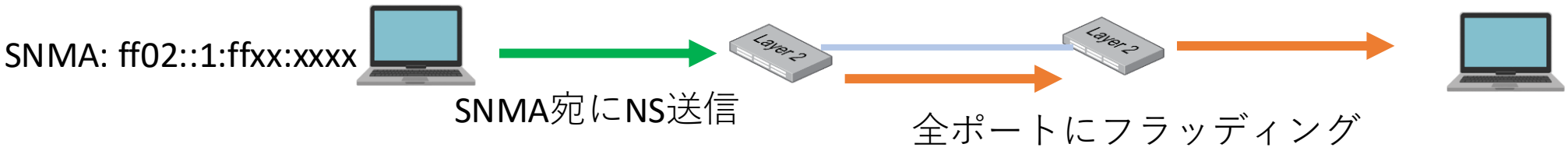
解決策④: ルーター間を接続して正しいGatewayを通るようにルーティング

- 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

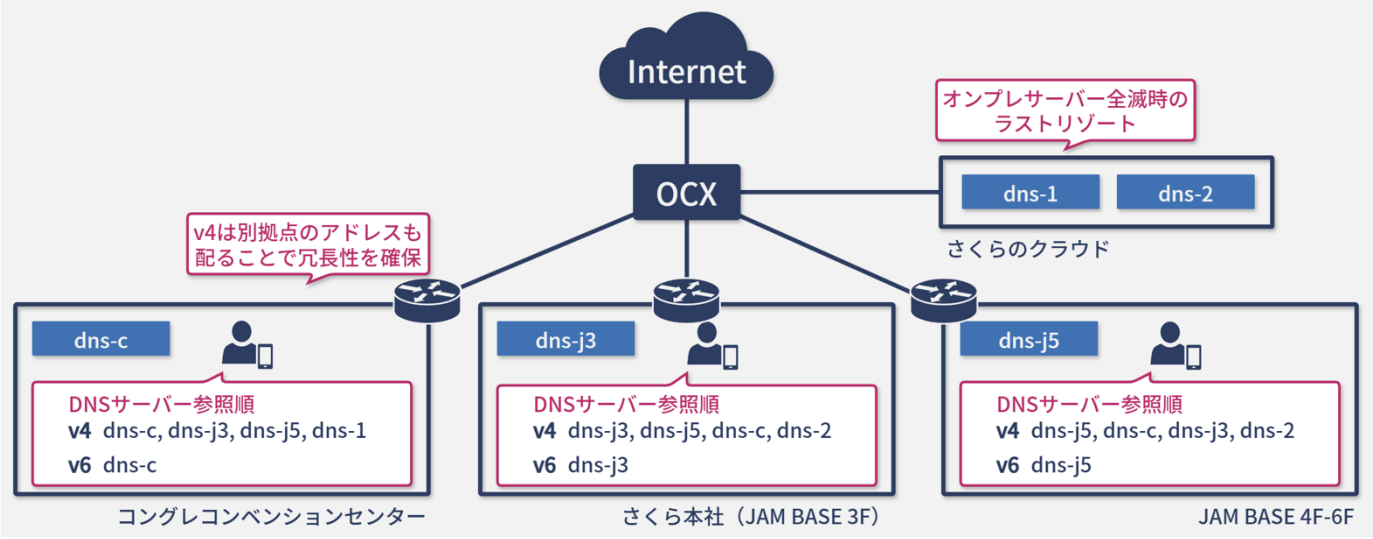
- 同一アドレスを複数拠点から広告
- イベントネットワークでは運用が困難

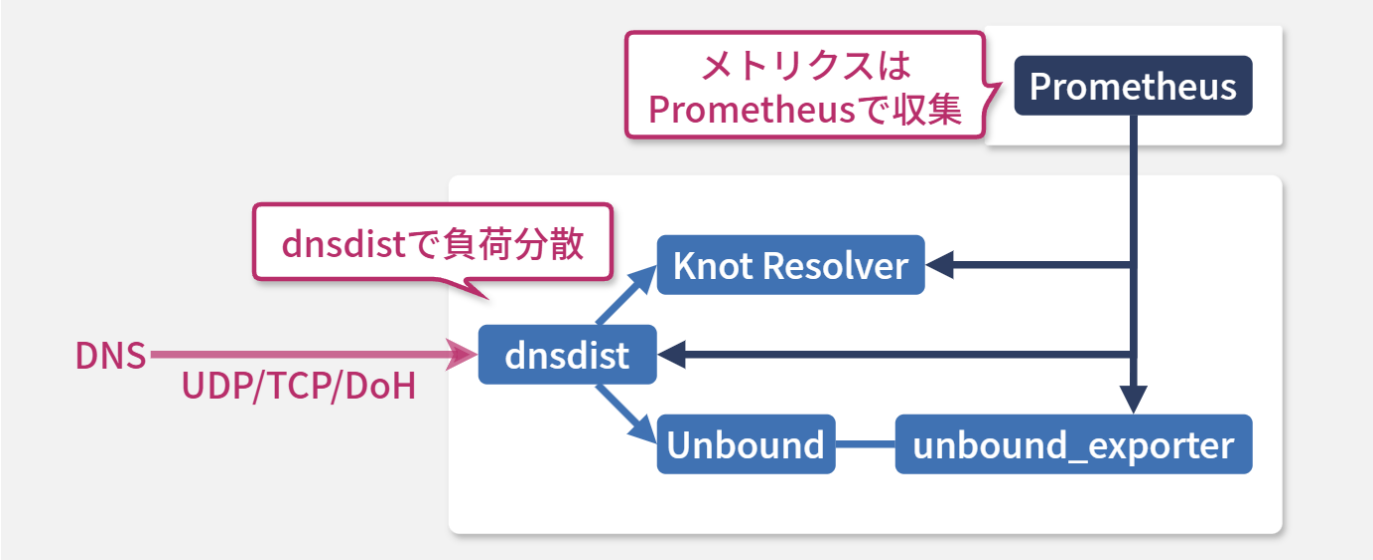
VRRP

- 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

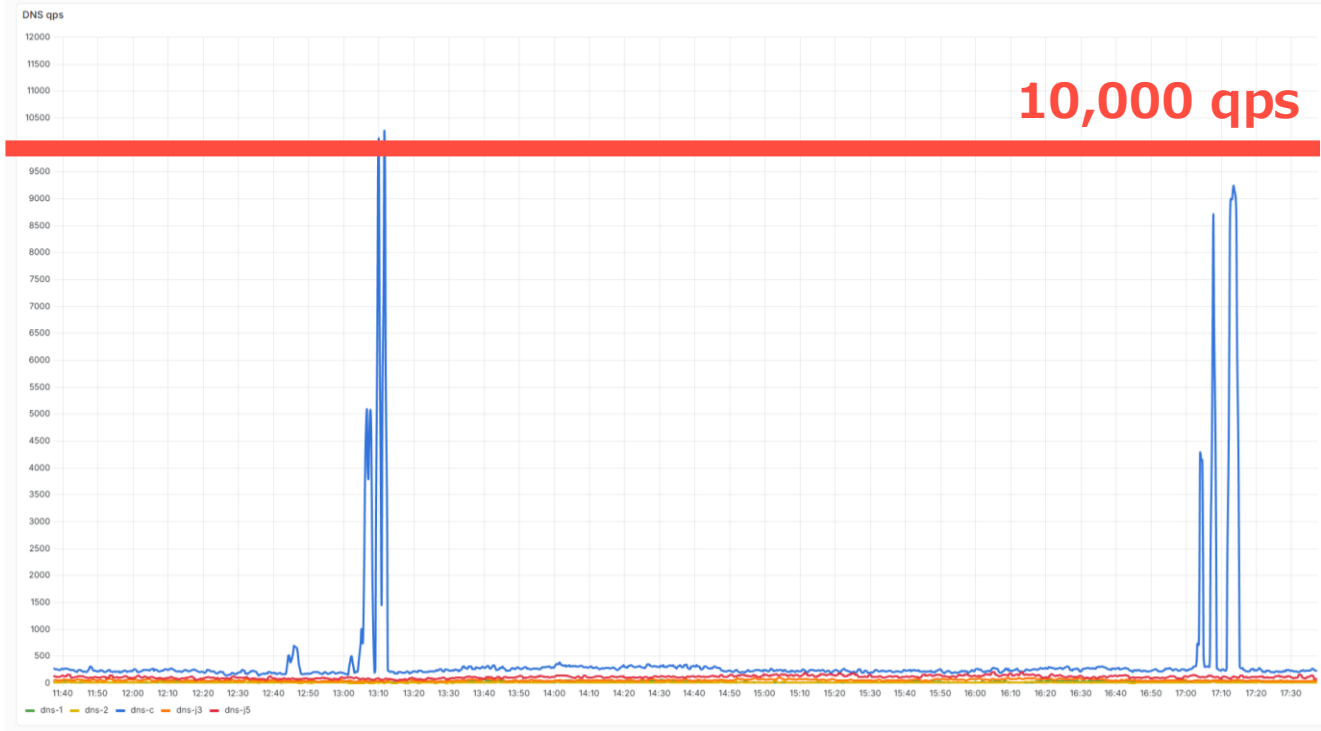




バーストDNSクエリ

観測事象(コングレ)

- DNSサーバーに対するクエリ数が突発的に急増するスパイクを**7回**観測
- 最大 **10,000 qps (queries per second)** 超過
- 通常はpeak 300 qps 程度
- **キャッシュヒット率はほぼ100%**
- 4回目はキャッシュヒット率が低下



監視項目

- qps
- キャッシュヒット率
- サーバードロップ数

バーストDNSクエリ

プライバシーポリシーに基づきDNSクエリログは保存していないため具体的なクエリ内容の特定は不可能

トラフィック特性からの推測

- 異様に高いキャッシュヒット率→通常のランダムなドメインではなく同一/少数のドメインに対する大量のクエリ？
- 単一のクライアントによる**大量のクエリ送信**であったと推定

結果

- ホットステージ期間の検証で**完全にランダムなドメイン**のクエリ送信では最低でも**3,000qps**まで耐えることを確認していた
- ランダムなドメインへの大量のクエリ以外では平常時と変わらないドロップ率
- **全体的な障害にはならなかった**

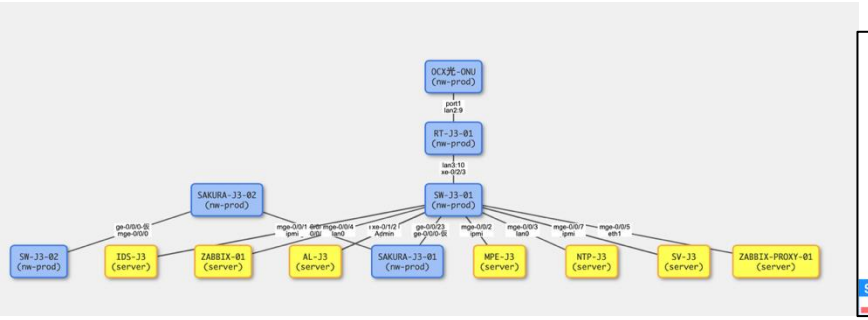
機器構成・コンフィグ把握の課題

- 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

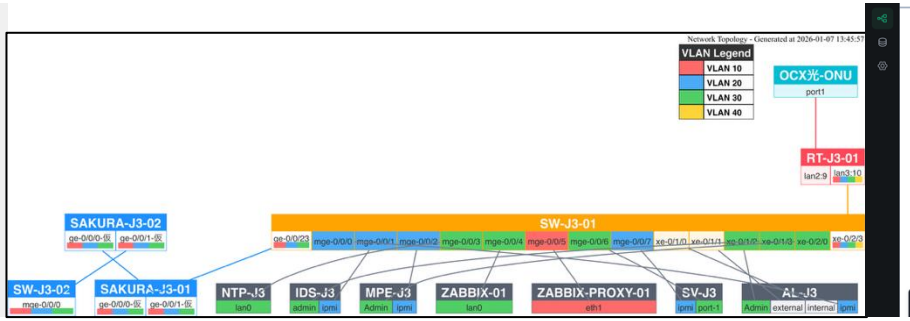
機器情報をNetBoxに集約し構成図は自動生成される仕組みの開発

- 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

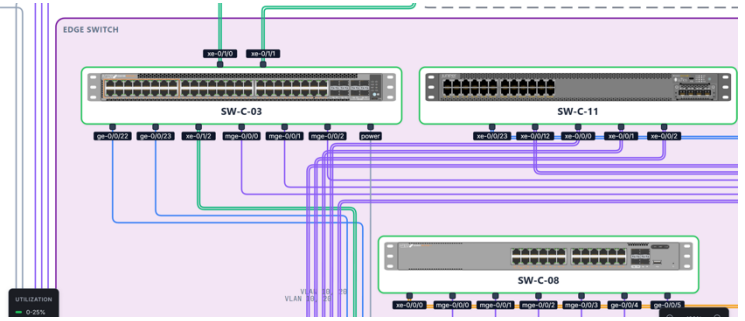
→イベントネットワークに最適な可視化ツールを作ってしまおう！



vis.js



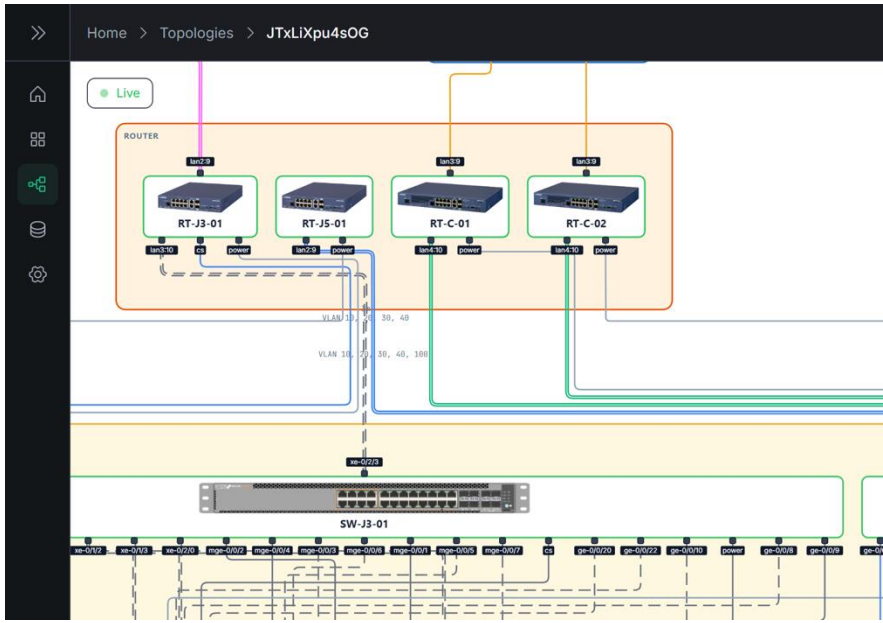
Graphviz



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
  - CI/CDでNetBoxが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



WindowsにおけるDHCPv6の挙動

現象

- WindowsクライアントにのみDNSサーバーとしてルーターのアドレスが追加されていた

想定動作

- IPv6アドレス配布：SLAAC (Stateless Address Autoconfiguration)
- DNSサーバー通知：RA (Router Advertisement)のRDNSS Option
- DHCPv6：**不使用** (Stateless/Stateful共に運用しない)

調査結果

- ルーターから送出されるRAのパケットをキャプチャ
- M-Flag (Managed Address Configuration): 0 (Off)
- O-Flag (Other Configuration): 0 (Off)
- RDNSS Option: 正常なDNSサーバーアドレスのみ

WindowsにおけるDHCPv6の挙動

原因

- 1つのルーターに誤って**DHCPv6 サーバーの設定が入っていた**
- パケットをgrepしていたため気が付くのが遅れた

Windowsだけに追加された理由の仮説

- ルーターはO-Flagを追加することでアドレス以外の情報をDHCPv6で取得可能なことをクライアントに伝えることができる
- O=0 の状態はクライアントがDHCPv6パケットを送出すること自体をプロトコルとして禁止するものではない
- WindowsだけはO=0でも**DHCPv6 Information-requestを出してDNSサーバーを取得しに行くのでは？**

帯域制限の実施

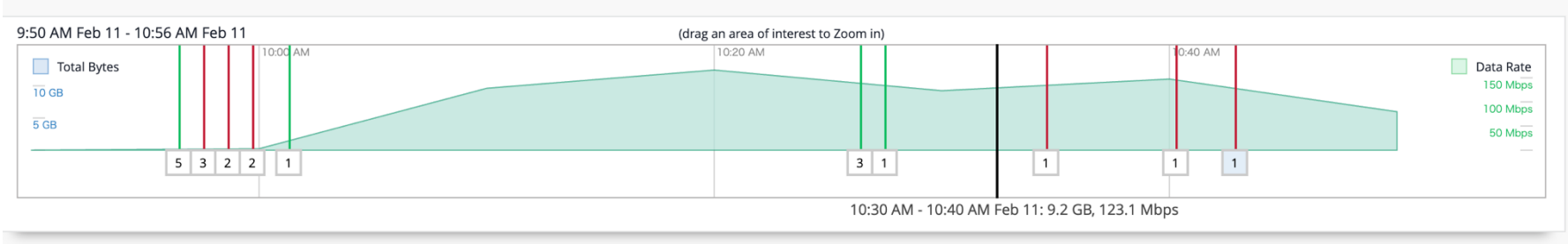
JAMBASE4-6Fは会場LANコンセントをお借りしている都合上帯域が**1G**

帯域圧迫

- Day1午前、利用者が300名程度の段階でバーストトラフィックにより帯域使用率が **80%** を超過しアラートが発報
- 特定の単一ホストによる **140Mbps / 累積51GB** の帯域占有を観測

推測される原因

- 少数クライアントによる大量の帯域消費



帯域制限の実施

- 特定ユーザーによる帯域占有を防ぎ公平性を確保するため、クライアントあたりの帯域制限を実施
- **下り 30Mbps / 上り 20Mbps** (Day1 昼に適用)

結果

- その後は接続性は改善した
- 有効性を確認した上で他拠点でも帯域制限を実施

eduroam/OpenRoaming

- 今回eduroam/OpenRoamingを提供しています
- 輝日さんにプロバイダとして認証基盤への接続を提供いただいています
- RadSec(RADIUS over TLS)でプロバイダRADIUSサーバに接続
- **JANOG57向けにOpenRoamingプロファイルを提供**
  - プロファイル配布数：**1183**

トラブル

- Day1で接続性が著しく低下
- 一部のみ接続可、大多数が接続不可
- RadSec再送により帯域を消費

Radiusセッション上限の緩和

- プロバイダで1 Global IPあたりの同時セッション数が **32** に制限されていた
  - **RADIUS(UDP)ではなくRadSec(TCP)を使用するからこそのトラブル**
- 1証明書あたり **10,000セッション** へ緩和
  - TCP Timeoutの時間を短縮
- 接続性が大幅に改善
- Access-Accept が正常に返却されるようになり再送パケットが収束

実際に使用されたセッション数

- **Peak 50セッション程度**
- パケットの観察の結果クライアントの認証が完了したタイミングでセッションを切るので少量で済んでいるのでは？

ケーブル敷設計画

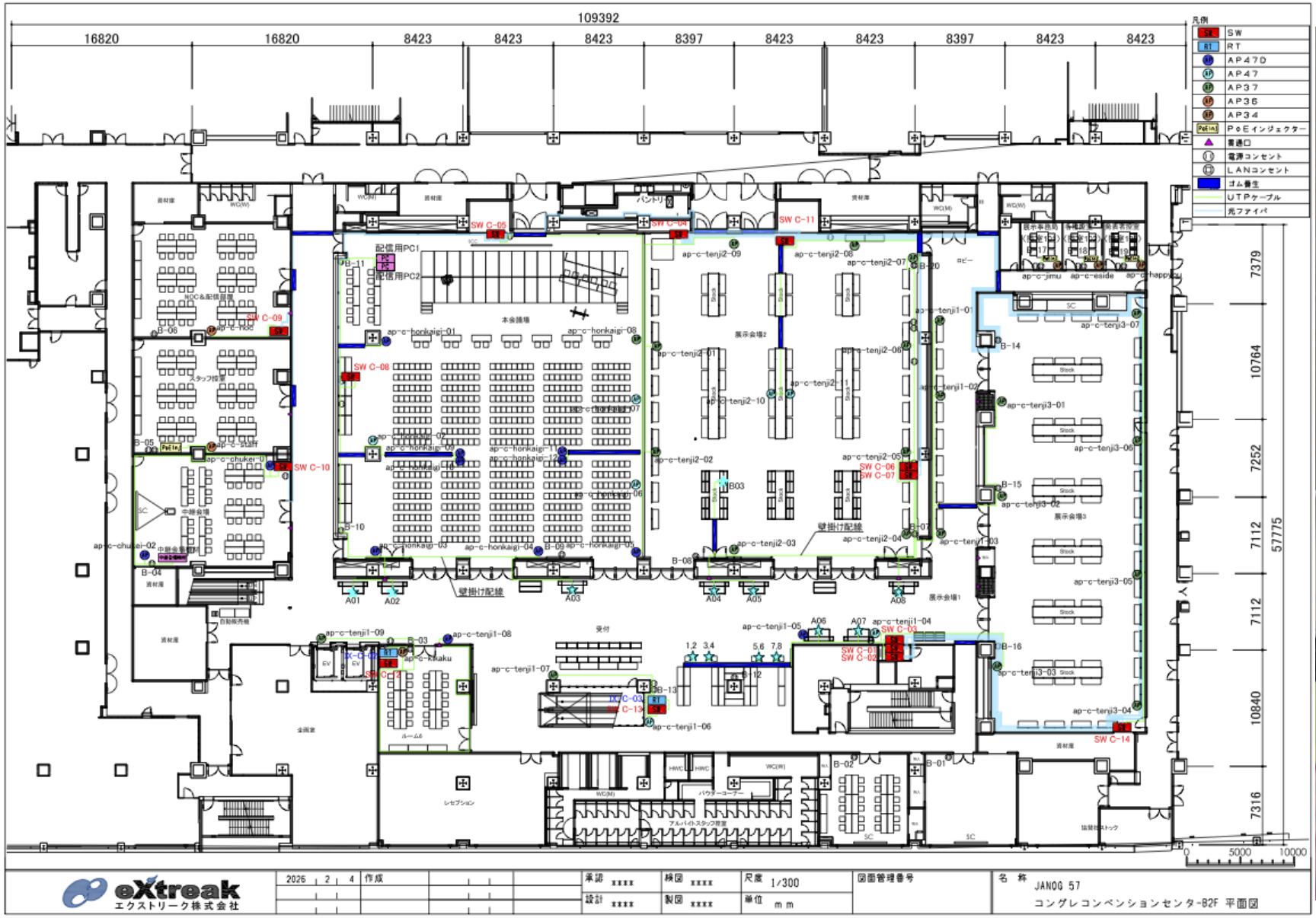
会場図面

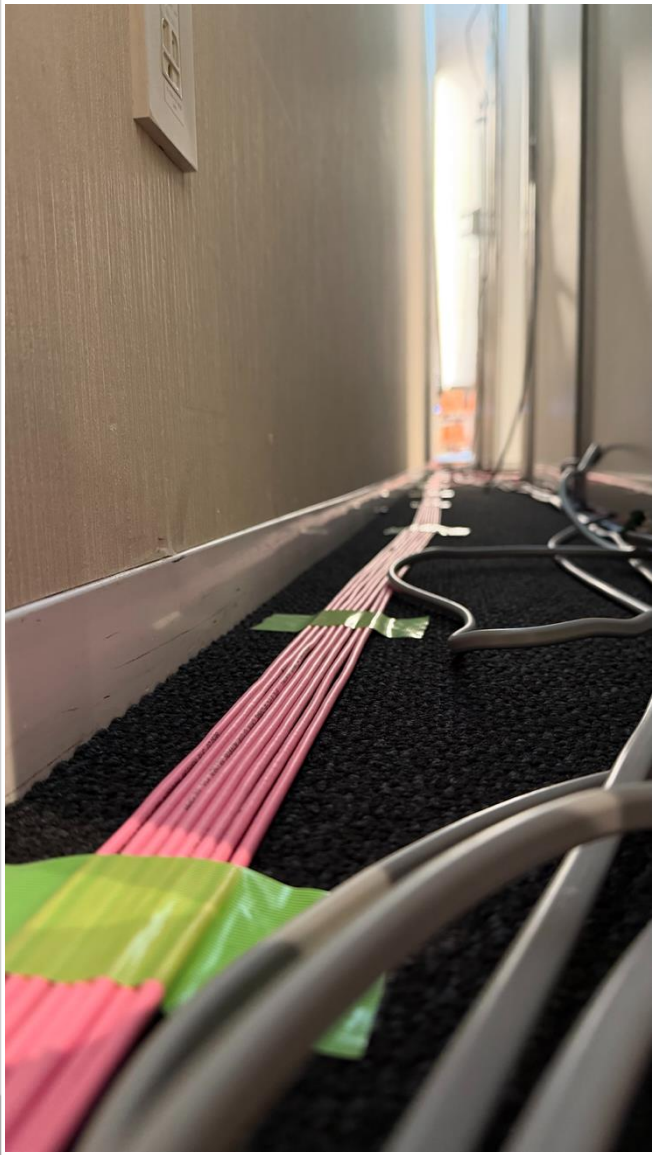
- 会場側で提供される図面に後付けでブース机や椅子配置を追加していくと徐々にスケールが荒くなる
- NOC向けに図面が届く頃にはスケールが荒くなり正確な配線計画が困難

独自図面作成(eXtreakさん)

- 下見時の**現地実測に基づく独自の配線図面**をゼロベースで作成
- レーザー距離計、メジャー(、目視)
- 独自図面をもとに配線計画を作成
  - NOCメンバー複数による同時編集(NOC独自)
  - どこを養生固定するかまで決められるとルールが統一されて管理がしやすくなる

作成した独自図面


109392
16820
16820
8423
8423
8423
8397
8423
8423
8397
8423
8423
凡例
SW
RT
AP47D
AP47
AP37
AP36
AP34
PoEインジェクター
電源コンセント
LANコンセント
UTPケーブル
光ファイバ
配信用PC1
配信用PC2
本会議場
展示会場1
展示会場2
展示会場3
壁掛け配線
7379
10764
7252
7112
7112
10840
7316
57775
0
5000
10000
eXtreak
エクストリーク株式会社
2026 2 4 作成
承認 ****
設計 ****
検図 ****
製図 ****
尺度 1/300
単位 mm
図面管理番号
名称 JANOG 57
コングレコンベンションセンター-B2F 平面図



投稿いただいたトピック