変化するDNS運用とこれからの課題について
(DNS設計/運用者の目線から)

2024年1月19日

NTTコミュニケーションズ株式会社

自己紹介

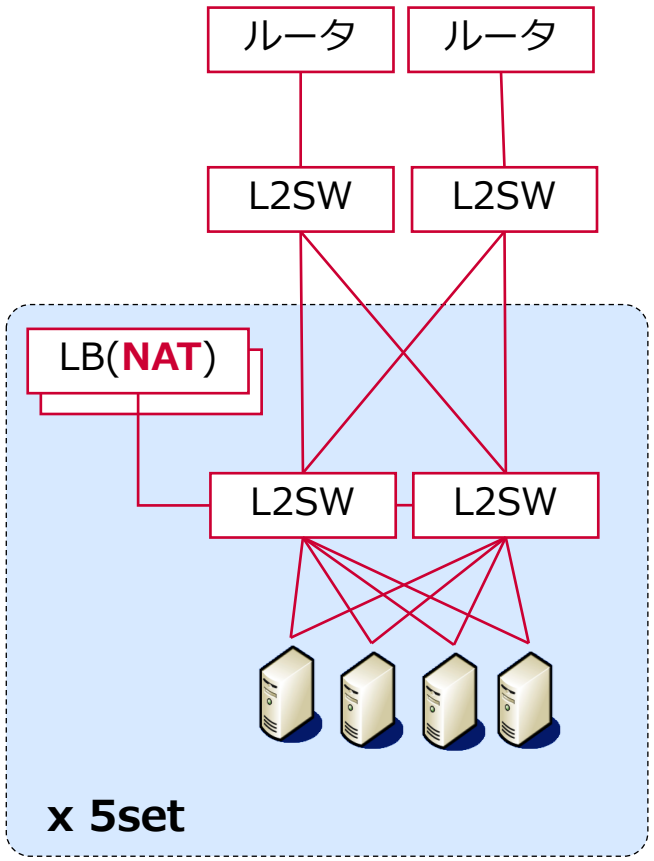
・名前：小坂　良太
・所属：NTTコミュニケーションズ株式会社
・業務：ISPとして提供しているフルサービスリゾルバ(キャッシュDNS)と
　　権威DNSの設計および開発　(回線サービス名：OCN)

・これまでに発表した内容
- DNSの可視化検討　(JANOG49 Meeting)
- DNSアプリの機能比較 on Rocky8編　(DNS Summer Day 2023)
- フルサービスリゾルバ利用状況　(Internet Week 2023)

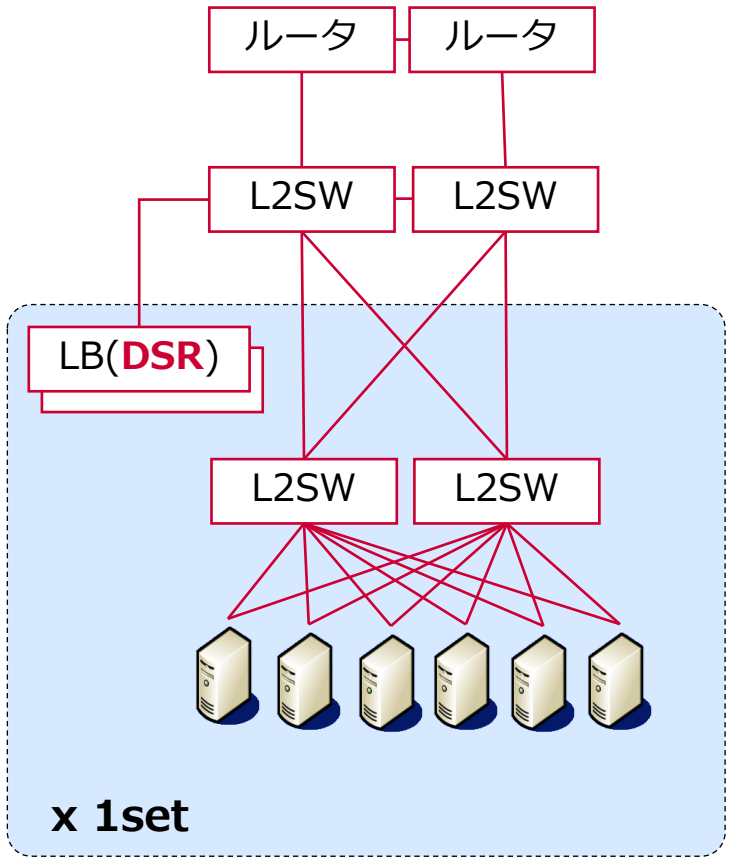
(※ 注意事項)
本日私の資料で紹介する構成および性能値は実際のものとは異なる場合がございます。

フルサービスリゾルバの構成遷移

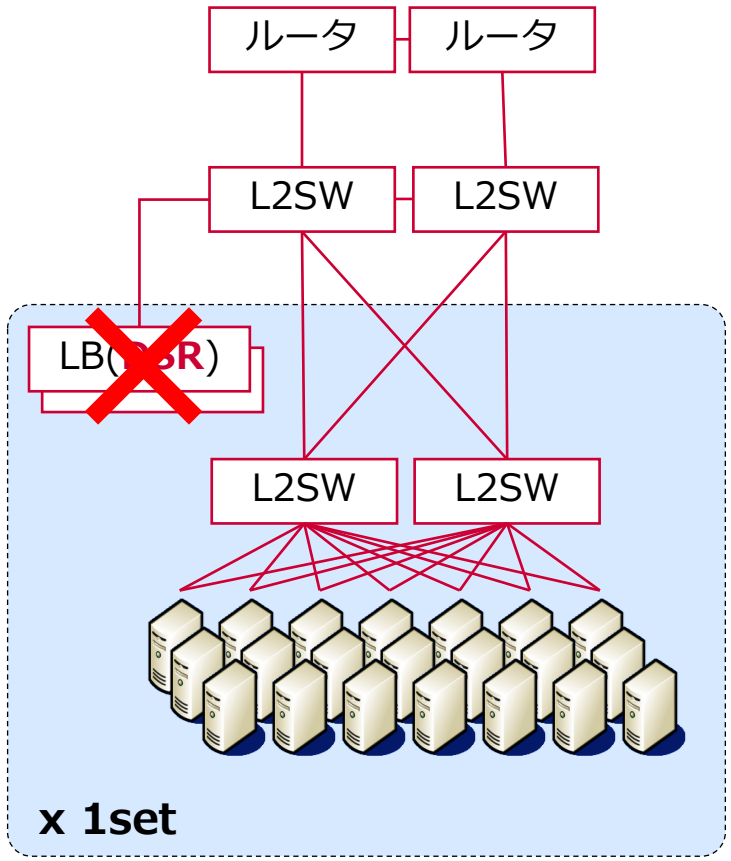
LBを用いたNAT構成
(～2019/12)



LBを用いたL3DSR構成
(2020/1～2020/12)



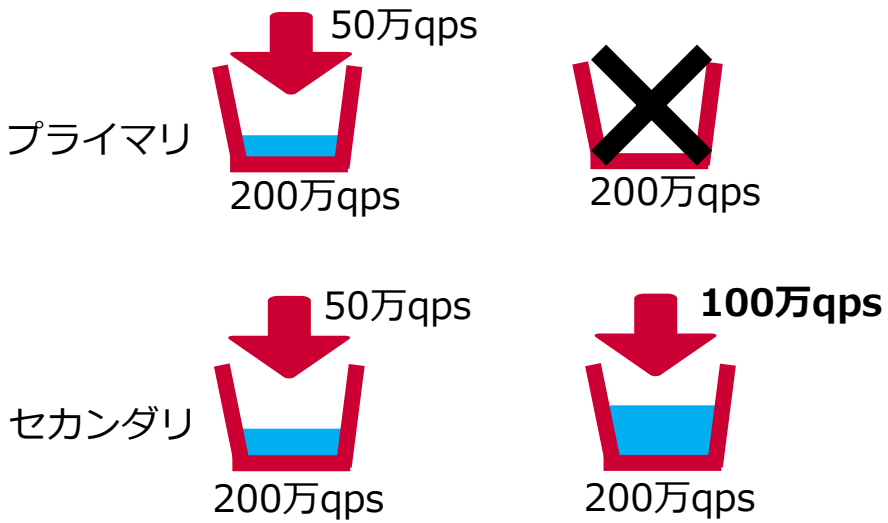
LBなし構成
(2021/1～)



DSR：Direct Server Return

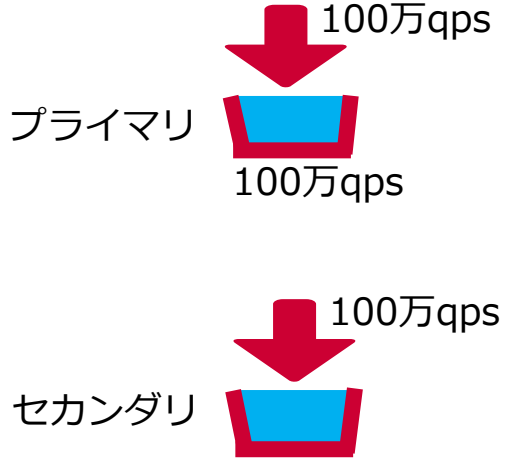
LBなし構成を導入した背景

DSR構成設計時の想定



NATからDSRへ変更することで
性能が大きく向上

現実



DSRを用いた次期基盤
開発中にトラヒックが2倍に

＋

商用のトラヒックパターン
ではLBの性能が想定と比べ1/2に

＝

1年以内に収容溢れの恐れあり
(一方が故障しても収容溢れ※)

対策案A 増設

更に2倍増えても
増設で耐えられるか？

対策案B 高性能LBの導入

商用に入れて
本当に性能が出るか？

対策案C LBなし構成の導入

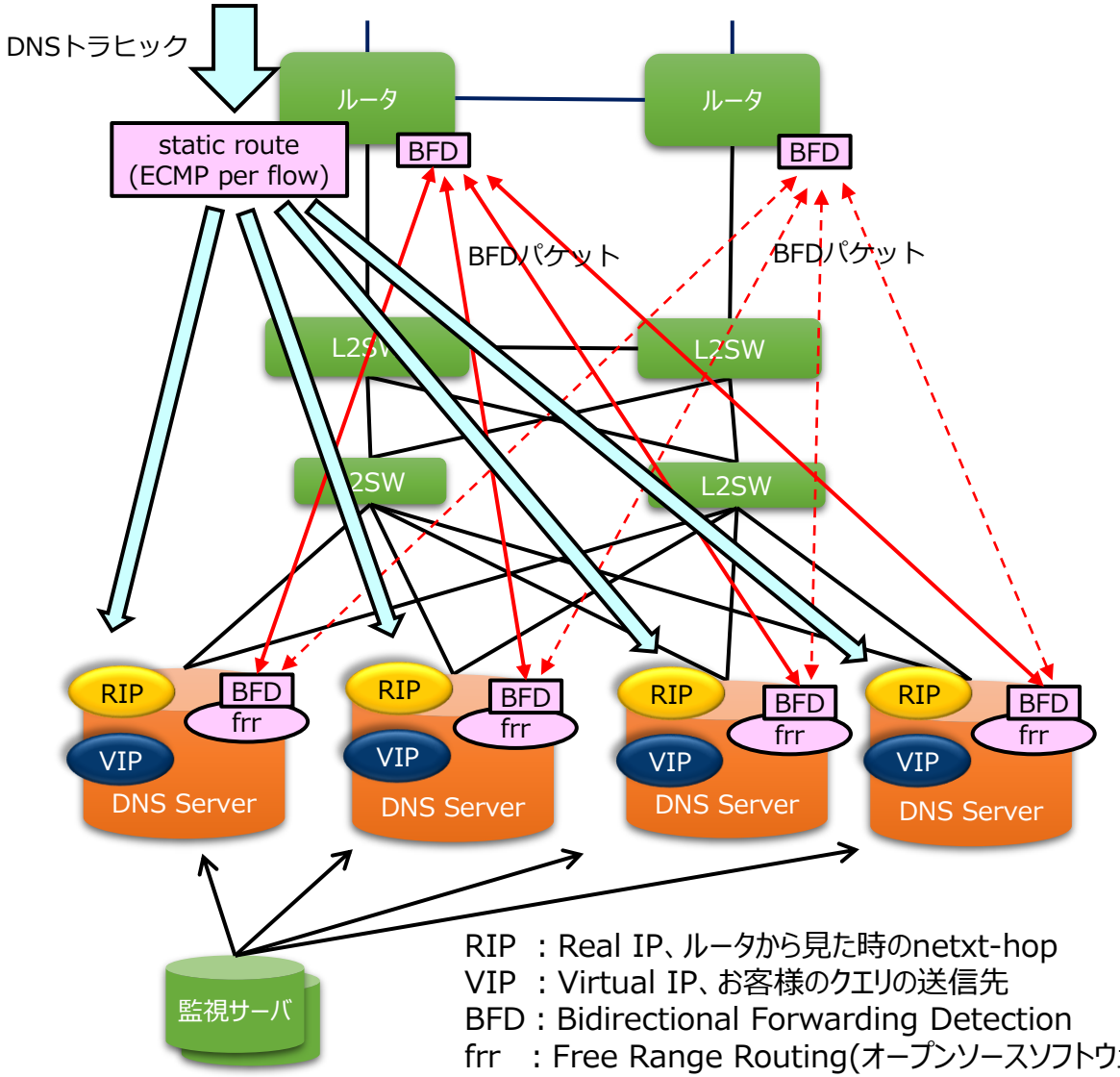
本当に動くか？

増設をしながら案Bと案Cを机上比較＆PoCによる簡易動作確認
LBなし構成のコアとなる部分の動作に問題がなかったため**案Cを採用**し設計着手

■凡例
- ：お客様のDNSトラヒック
- ：LBで処理中のDNSトラヒック
- ：LB(キャパシティ)

(※) 実際には収容溢れしないよう旧構成を保守延長し並行運用により対処

LBなし構成の概要



RIP ：Real IP、ルータから見た時のnetxt-hop
VIP ：Virtual IP、お客様のクエリの送信先
BFD：Bidirectional Forwarding Detection
frr ：Free Range Routing(オープンソースソフトウェア)

構成について

- ルータのECMPを有効にしVIPに対する**static routeの宛先を複数設定**することでトラヒックを負荷分散
  - ✓ BGP+BFDとstatic route+BFDを比較し導入ハードルの低いstatic routeを選択
  - ✓ 多くの機種で64台程度分散可能(デフォルト値に注意)
- DNSサーバのloインターフェースにVIPを設定
  - ✓ 元々はL3DSR導入のために設定したもの
- DNSアプリが応答不可の時に組外しできるようL7監視ツールを導入
  - ✓ DNSパケットを送信し応答がなければfrrプロセスを停止
  - ✓ **BFDを用いて動的な組み込み/組外しを実現**

安定性・性能について

- 約3年運用し大きなトラブルなし
  - 設計・検証期間は約6か月程度だったが手戻り等もなし
- 商用環境で100万qps以上処理
  - 10G基盤の場合、サーバ台数およびルータの性能に問題がなければ約500万qps処理可能(机上の計算。ボトルネックは帯域)
  - 複数のサーバにほぼ等分散(クエリの偏りはなし)

ハードウェアLBについて

元々抱えていた課題

LB増設にかかる人的コスト・時間的コスト

- LB増設に併せてL2スイッチやサーバも必要となる
- サーバがEoSになっていることが多く新規機種の選定、調達、試験に時間がかかる
- LBの調達も時間がかかる

トラヒックの平準化が困難

- ナップサック問題を解くようなもの
- 収容変更には工事を伴うので柔軟に入れ替えることが難しい

プライマリとセカンダリの分配状況と
将来の需要を考慮した収容変更

IPoE普及に伴い新たに生じた課題

増設(スケールアウト)では対応不可・スケールアップが必要

- IPoEでは1VIPにトラヒックが集中するため収容溢れの場合は上位グレードのLBへ買い替えが必要
  (DNSアドレスが複数あり分散収容できれば旧構成のように複数setにスケールアウト可能)

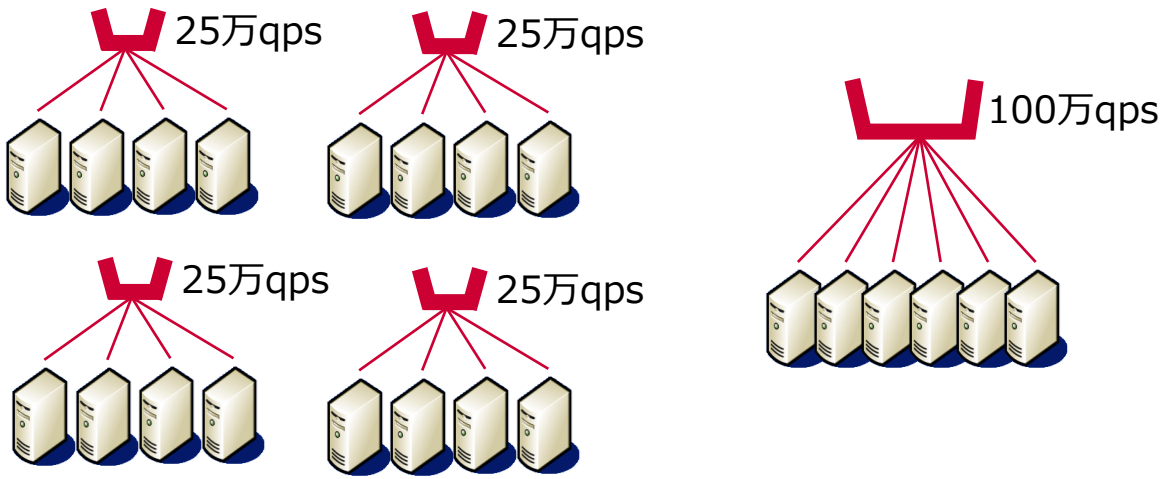
大きな土管を用意する必要があるがどの程度が「大きな」土管か？

- トラヒックが2年で2倍に増えている状況下では予測が困難
  - 線形に推移すると次は4年で2倍。指数的に推移すると4年で4倍
  - 200万qpsが400万～800万qpsに増えることを想定してLBを選定？

LBがなければ解決

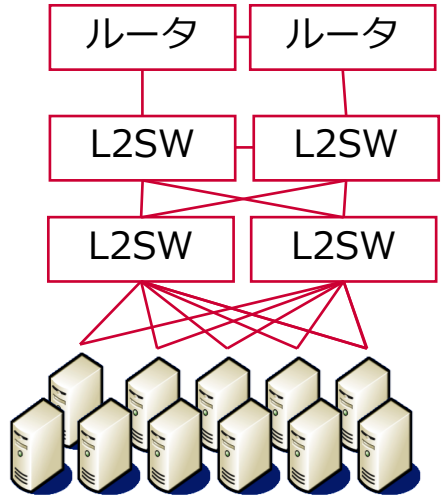
DNSサーバの管理コストについて

LBを用いたNAT構成



- **200万qps処理する基盤ではサーバは22台**
  - ✓ EoSによりサーバは3機種を運用

LBなし構成



- **200万qps処理する場合、サーバは11台**
  - ✓ EoS時に増設を行うことで1機種運用を継続

・分割損なし
・性能の高いDNSサーバ/アプリを使うことで高密度化
・サーバはLBと比べ安価なため将来を見越して多めに増設可能
(NAT構成時は25万qpsまたは100万qps単位での増設)

IPアドレスが変えられない問題について

背景と課題

参照用DNSとして払い出すIPアドレスの数が多い

- 元々はトラヒックを細かく分けて複数のLBに分散収容する設計ポリシー
- サービス毎、地域毎にIPを分けて払い出し (例：東北地域のADSL用DNS)
- プライマリとセカンダリを合わせると100IPを超える
- 設備更改やLB増設の際にそれらを移設する必要あり

参照用DNSのIP変更(集約)にはコストがかかる

- 何千台とあるネットワーク終端装置(NTE)の設定変更が必要

本来の用途以外の利用(手動設定)もありIP変更(削除)すると影響を受ける

- ADSL用DNSをフレッツ光のお客様が利用するようなケースあり

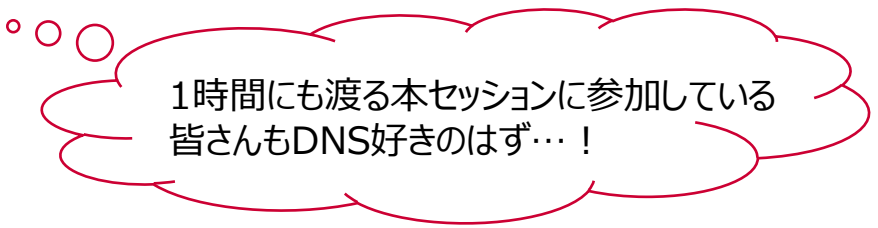
- **ISDNやADSLのサービス終了に伴い是正を目指す**
- **LBなし構成が安定運用できているおかげで以前と比べ稼働に余裕あり**

これからの課題：やるべきことが沢山問題

- EoS(Sale)、EoS(Service)、EoL(Life)対応の稼働を如何に削減するか
  - クラウド化できれば良いがIPが変更できないかつBYOIP(※)できないものをどうするか
  - 物(物理機器・構成パターン・IP)をどう減らせばよいか
    - とは言いつつ、アプリ冗長などコストを掛けるべき所にはコストを掛けたい

- 維持管理をするか新規開発をするか
  - サービスをクローズせずに長く運用しているため負の遺産は多い
    - 負の遺産を減らせば日々の負担を減るし、品質も向上する
    - ただし改善活動をどれだけ進めても新規機能の提供は稀
  - 負の遺産を解消しつつ、浮いた費用を使って今後も新規機能をどんどん提供していきたい

- 人手不足をどうするか
  - DNSが好きな方は多い印象
  - 積極的に採用していきたい



(※)BYOIP：Bring Your OwnIP