フルサービスリゾルバの
ロードバランサーなし構成について

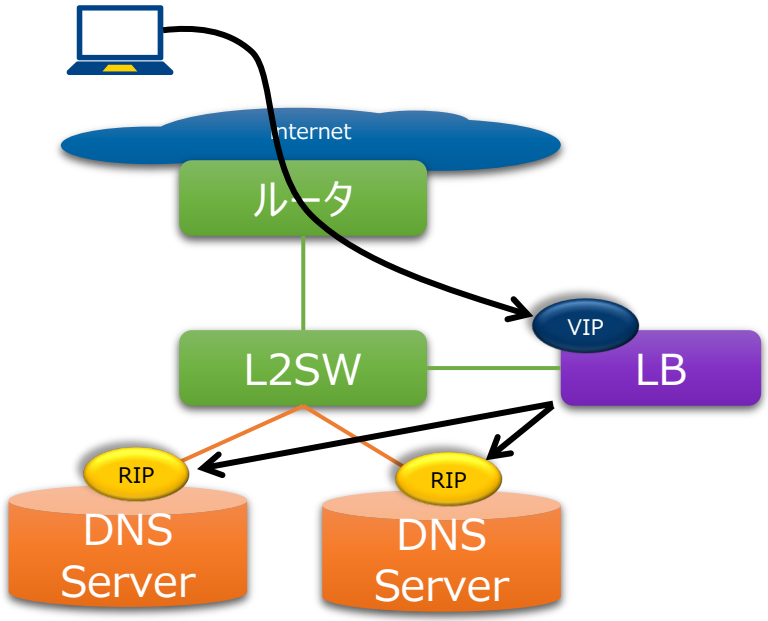
NTTコミュニケーションズ株式会社
小坂　良太

はじめに

本発表のターゲット	・DNSサーバを運用しておりロードバランサーを導入している方 ・本業はサーバで、NWについてはこれからスキルを伸ばそうとしている方
対象外	・設計や運用は未経験で、これからDNSシステムを設計・開発しようとしている方 ・NWに詳しく、anycast構成でシステムの設計/運用ができる方

本発表における用語について

用語	内容
キャッシュDNS	フルサービスリゾルバのこと。
DSR	Direct Server Returnのこと。 (今回紹介する新構成はL3DSRをベースにしています)
LB、ローバラ	ロードバランサーのこと。

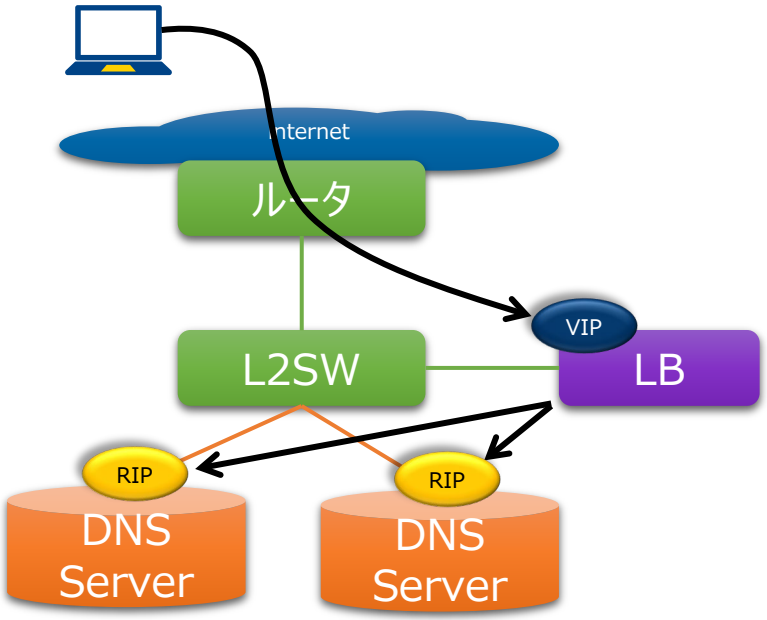


VIP
- お客様のDNSクエリを待ち受けるアドレス
- Virtual IP、VIP
- DNSのアドレス

RIP
- DNSサーバのIP
- Real IP、RIP、実IP
- 再帰問い合わせ用のIP

本発表では、
この呼び方を継続します

オンプレLBの課題



・オンデマンドにキャパシティを増やせない
  - オンプレなので当たり前
  - 調達期間もサーバに比べると長い

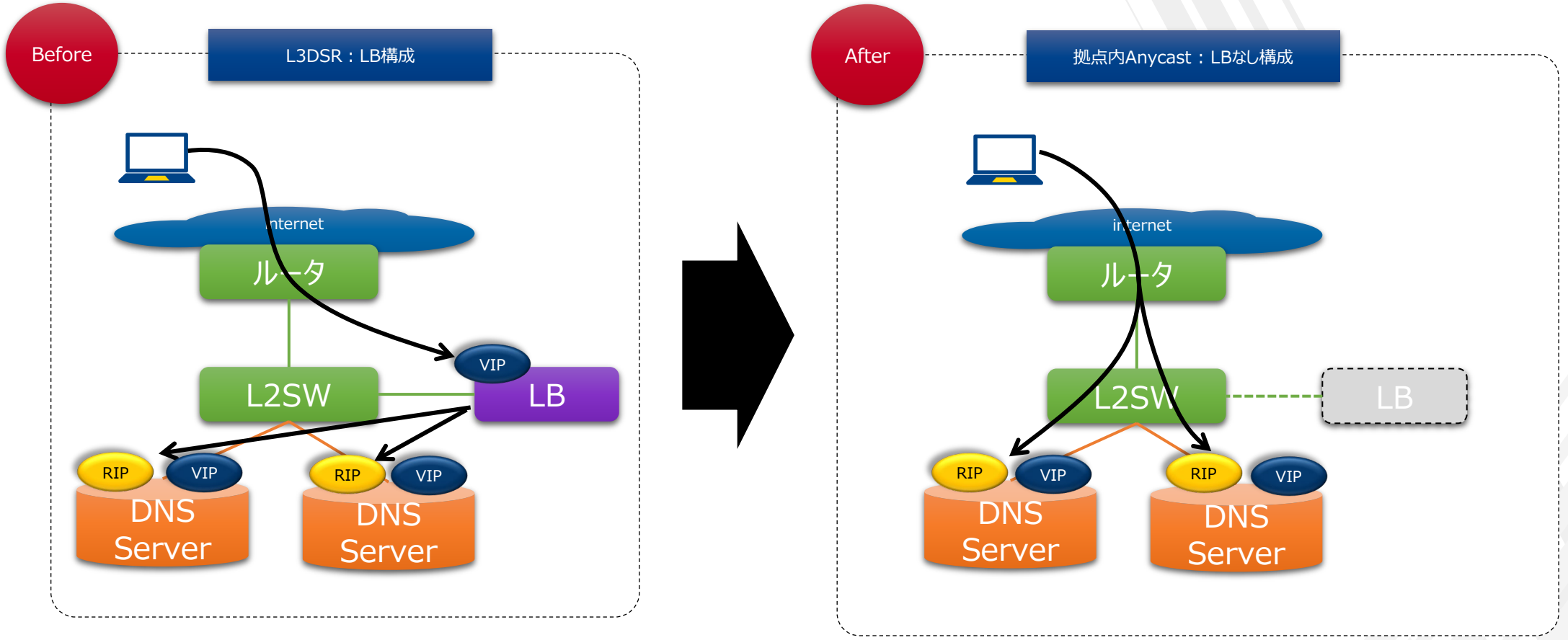
・スケールアウトができない場合がある
  - DNSのアドレスが複数あれば分散可能

・高コスト
  - 冗長や予備機を考慮すると余計にコスト高

・EoL/EoS対応

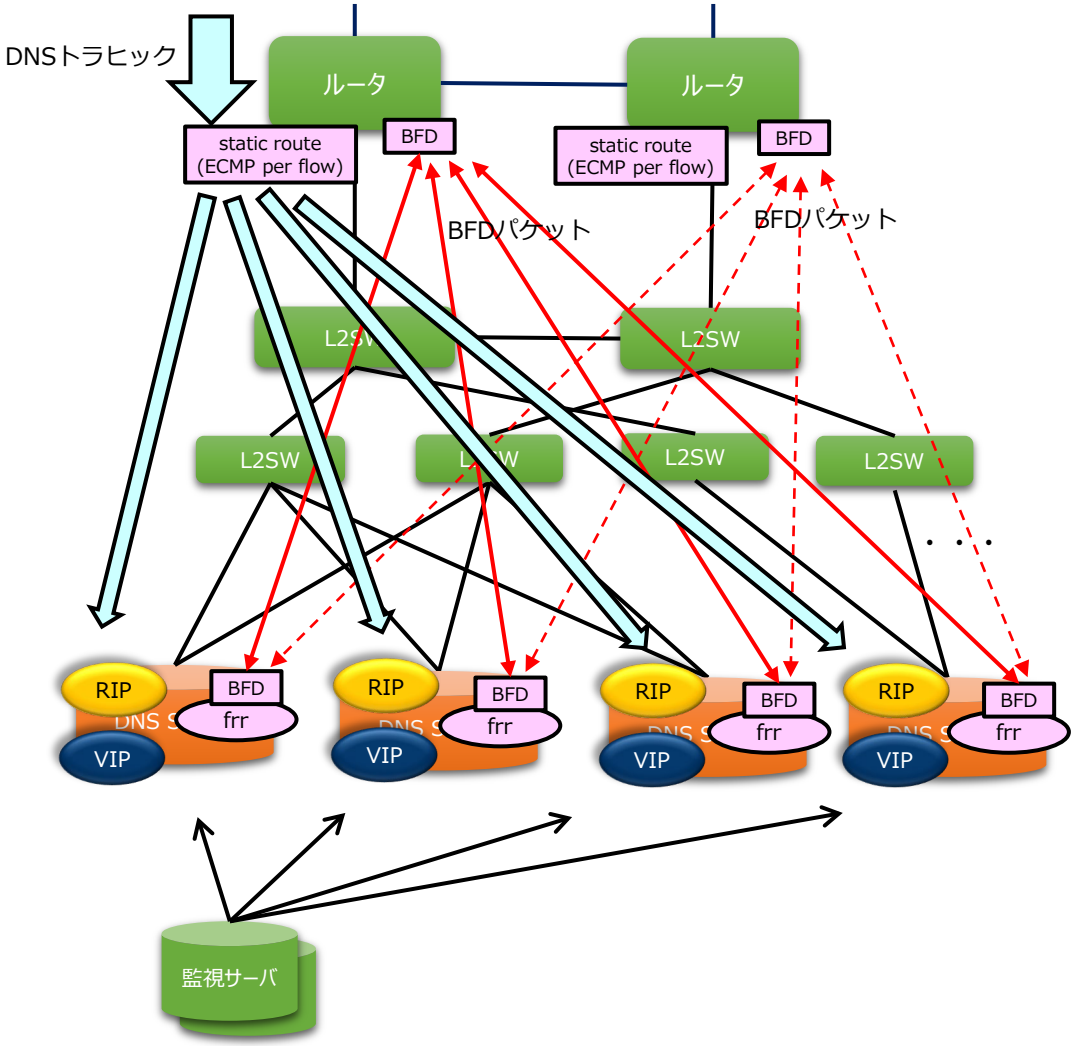
このような諸々のLB増設問題に対する1つのアプローチ
||
LBなし構成

LBなし構成の概要



・構成はL3DSR構成とほぼ同一
・ルータから見た時のnext-hopをLBからDNSサーバへ変えるだけ

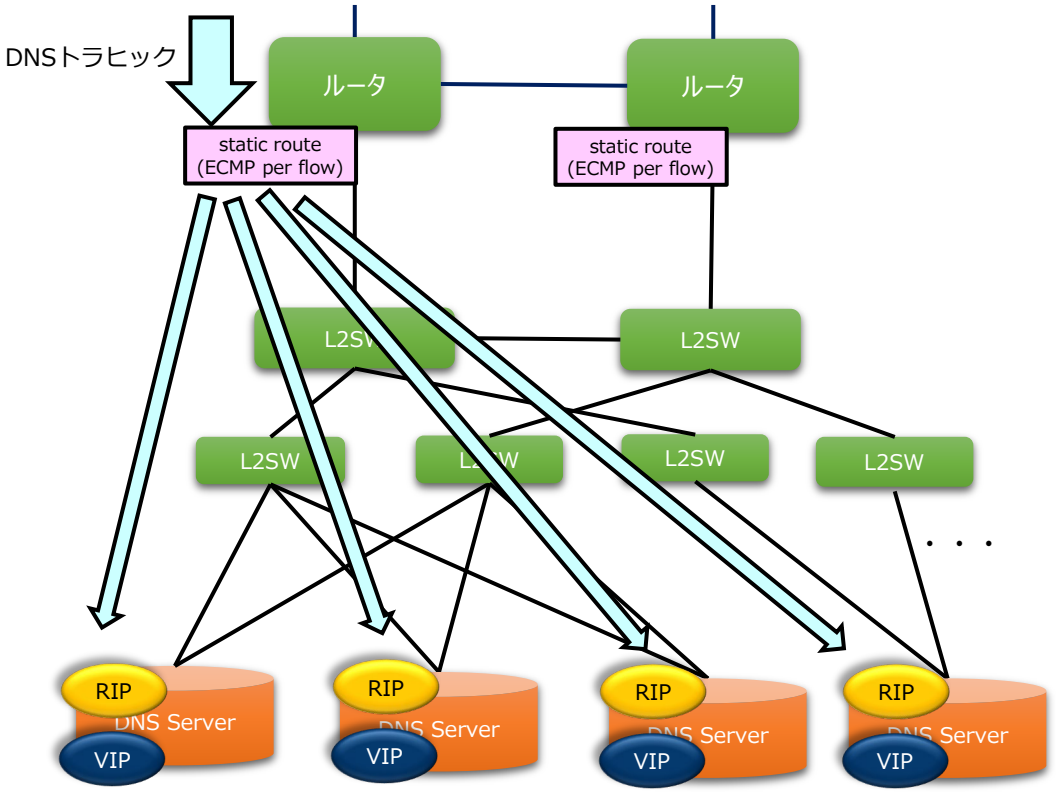
LBなし構成のアーキテクチャ



今回紹介するアーキテクチャ

- ECMP(+DSR)
- BFD
  - FRR
- ヘルスチェックツール

ECMP(+DSR)について



・ECMP=Equal Cost Multi Path

・等コストの経路(next-hop)が複数あるとトラヒックは
バランシングされる
⇒next-hopとしてサーバのIPを複数指定すれば良い
宛先IP(VIP)は/24でも/32でも指定可

・ルーティングプロトコルは何でも良い(BGP/OSPF/static)

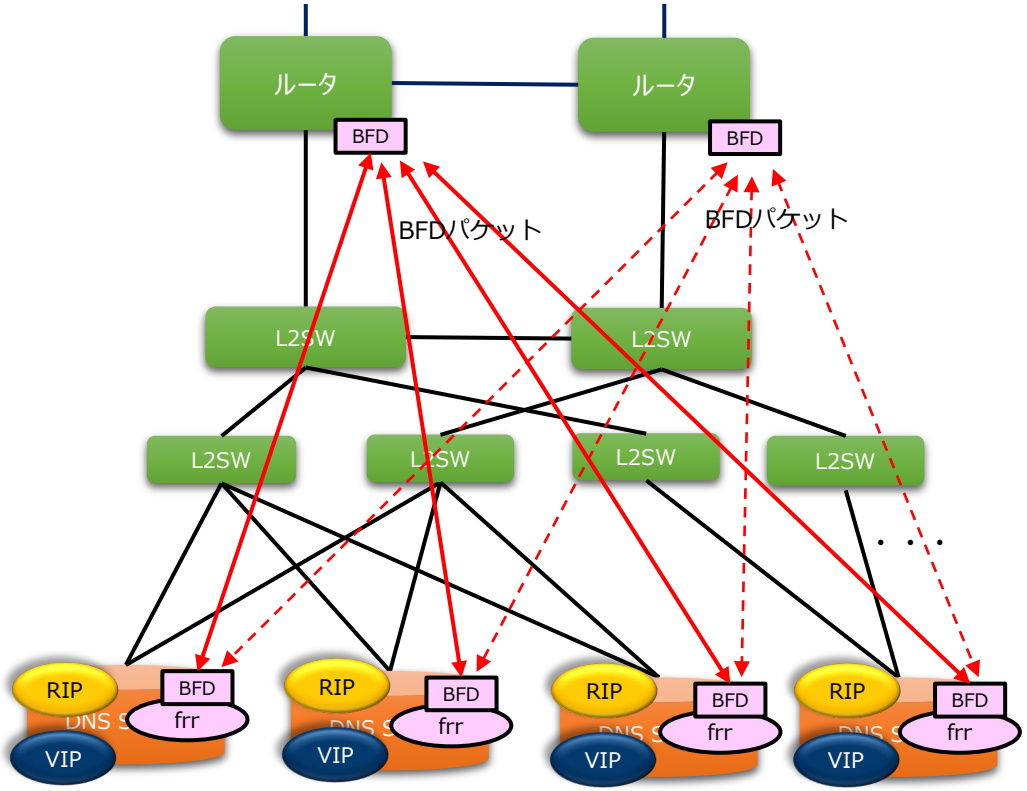
・チューニング箇所はバランシングポリシーのみ
- per packet
- per flow

} TCP通信を成立させるためper flowを選択

・DNSサーバからすると宛先IPがVIPなパケットを受信するので
loインターフェースにVIPの設定が必要(DSRと同様の設定)

課題：サーバ故障時に切り離し可能か、
またどの程度で切り離し/組み込みが可能か

BFDについて



・BFD=Bidirectional Forwarding Detection

・ルータ間で相互にpingを打ち合うようなプロトコル
(実際にはUDP 3784番ポートを用いる)

・BFDを用いるとmsオーダーで断検知が可能
また、L2スイッチを超えて断検知が可能
⇒つまり、サーバの切り離し・組み込みが高速に行える

・チューニング箇所は以下
- 断と見なす送信回数
- 最小受信間隔
- 最小送信間隔(※)

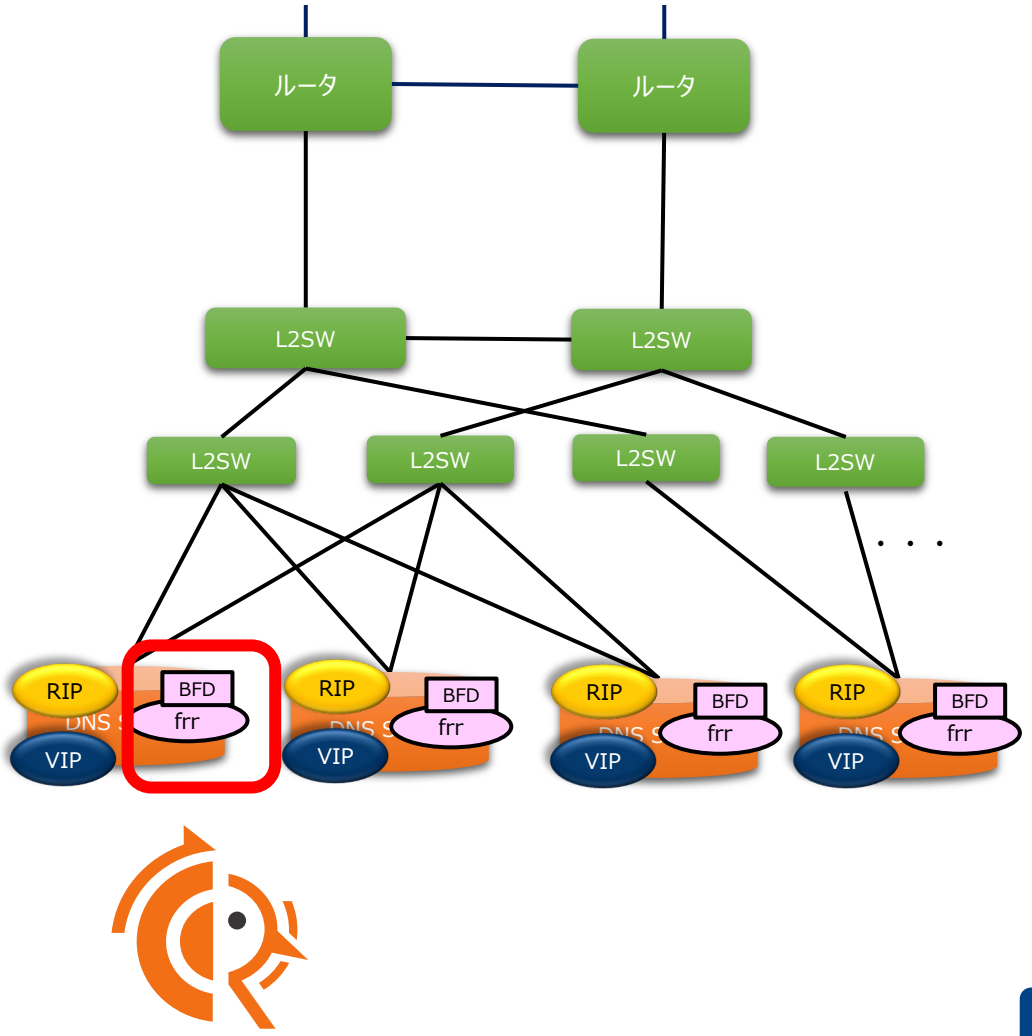
} 現在のLBのヘルスチェックポリシーを参考にチューニング

- echoモードの有効/無効 ：基本的に無効で良い
- Control Plane Independent ：Graceful Restartを考慮する場合に設定

(※)BFDセッションがdown時は送信間隔を調整する実装がされている場合があります.
この挙動により「意図した間隔でパケットがこない」「意図せずBFDがdownする」といったことが起きることがありました。

課題：サーバでBFDを動かすにはどうすればいいか？

FRRについて



- FRR=Free Range Routingというオープンソースソフトウェア
- Quaggaの代替ソフトウェアで一般的なルーティングおよびBFDをサーバで動作させることが可能
- BFDの軽微な不具合改修がされている7.4以降がオススメ(最新版は先月リリースされた7.5.1)
- コミュニティよりDebian PackagesとRPM Packageが提供されるためコンパイル不要(RHEL8ではredhatからrpmが提供される)
- 設定方法は以下
  zebra.conf：ホスト名の設定
  daemons：起動するデーモン(=BFD)の指定
  bfdd.conf：先程のパラメータを設定

残課題：リンクダウンおよびサーバ故障以外の障害(L7障害)でサーバを切り離すにはどうすればいいか？

本ページの
配布の予定はございません

ヘルスチェックツールについて(1/2)

ヘルスチェックツールについて(2/2)

本ページの
配布の予定はございません

NTT Communications
Go the Distance.

本ページの
配布の予定はございません

LBなし構成のキャパシティ

LBなし構成の課題・デメリット

・LBで担っていた機能が使えない
  - ヘルスチェック
  - rate-limit/流量制限
  - トラヒックの可視化(MIB)
  - 重みづけのバランシング
  - ICMPに対する代理応答

・何らかの代替手段が必要
  - 今回の発表ではヘルスチェックのみご説明させて頂きました

まとめ

・ECMPとBFDを組み合わせてLBなし構成を実現

・サーバでBFDを動かすためにFRRを採用

・LBなし構成のキャパシティは10G構成であれば500万qps程度収容可能
  - ただし、DNSの応答サイズに依存
  - また、クエリ数に応じたサーバ台数が必要

・LBで担っていた機能については代替手段の検討が必要

参考にさせて頂いた資料

- LINEのネットワークをゼロから再設計した話(JANOG43 Meeting)
  - https://www.janog.gr.jp/meeting/janog43/application/files/7915/4823/1858/janog43-line-kobayashi.pdf
- 自作ロードバランサ開発(JANOG40 Meeting)
  - https://www.janog.gr.jp/meeting/janog40/application/files/3415/0208/4443/janog40-sp6lb-kanemaru-03.pdf
- ロードバランサのアーキテクチャいろいろ(yanazuno様のHatena Blogより)
  - https://yunazuno.hatenablog.com/entry/2016/02/29/090001
- 頑張れIP anycast(JANOG34 Meeting)
  - https://www.janog.gr.jp/meeting/janog34/doc/janog34-acast-matsuzaki-1.pdf
- 高速切替手法の検討(JANOG19 Meeting)
  - https://www.janog.gr.jp/meeting/janog19/files/BFD_Suzuki.pdf

ご清聴ありがとうございました