零細組織でイケてるNWをつくる話

JANOG54 (in NARA) BoF at Room 205

2024年7月3日 15:55–16:20 (部屋枠 15:45–16:30)

一般社団法人生活情報基盤研究機構

登壇者：森祐佳，百瀬蓮

BoFの流れ

- 登壇者からの発表
  - 登壇者の自己紹介
  - AS63806の簡単な説明
  - iBGP+OSPFv3 を IPv6 link-local で動作させる手法の話
  - フレッツ光の上に Wireguard+BGP EVPN L2VPN を組んで暗号化された広域イーサネットを構築する話
  - Anycast で小規模組織でも安価に冗長化されたスケールアウト可能なWebサーバ群を走らせる話
  - 余談
- 質疑応答&フリートーク

発表者の自己紹介

- 一般社団法人生活情報基盤研究機構
  - 通称 Menhera.org
  - https://www.menhera.or.jp/
  - ざっくり言うと，生活に密着したITの研究開発で学術や文化，プライバシーなど公益に資することをやろうという非営利組織
  - AS63806を運用
- 森　祐佳
  - 代表理事，AS63806の運用担当者
- 百瀬　蓮
  - 技術系メンバー

自己紹介 (森)

- (一社)生活情報基盤研究機構 (Menhera.org) の設立者のひとり
- JANOG 初参加は前回の JANOG53 (FUKUOKA)
- AS63806 (MENHERA, Human-life Information Platforms Institute) の主運用担当者
- ネットワーク以外だと，哲学や医学などに興味があります
- 趣味は作曲など

X: @vericava

自己紹介 (森) つづき

やっていること：

- Asterisk (オープンソース VoIP サーバソフト) などで利用可能な時報音源アプリを作ったり (個人開発)
- IPv6 link-local を駆使して楽に小規模 AS を運用する方法について考えたり → 今日話します!
- など

自己紹介

百瀬　蓮

・僧侶@インド仏教

・リサーチャー@Skyland Ventures

・正構成員@一般社団法人生活情報基盤研究機構

主にzkp(ゼロ知識証明)を用いた情報基盤の社会実装に興味があります。

Nostrが好きなのでおすすめ記事貼ります(右が拙著)

最初にAS63806の簡単な紹介

小さな非営利のネットワークですが……

AS63806の成立の経緯

- (一社)生活情報基盤研究機構 (以下，Menhera.org) は「もともとITの専門家じゃない人たちが中心となって集まって，ITで公益的なイイことをしよう」という趣旨で始まった
  - 別に「メンヘラ」とかメンタルヘルスとはあまり関係ない
    - 誰でも入れます (ITやネットワークの専門家でも入れます)
- その目的に沿って，主に以下の2つの目的でASが誕生
  - ネットワークづくりや運用を実際に体験するという教育的目的や，研究用のネットワークとして
  - 非営利の事業 (OSS，実験的サービスなど) を行う基盤としてのネットワークとして
- Menhera.org はネットワークだけを専門的に行う組織ではない

AS63806に求められるもの

- 非営利の法人という性質と，外部から多額の寄附や補助金がもらえるわけではないので，メンバーの寄附で成立
  - そんなにお金がかかることはしていないので……
- リッチな人はあまりいないので，金銭的コストを下げる必要
  - このような実験的ネットワークの場合，自分たちでいろいろやれば，コストは下がる
- 運用するスタッフが充実していないので，人的コストを常時たくさんかけられるわけではない
- 同じく非営利の，HOMENOC様の実験ネットワークに参加し，接続を利用させていただいています
  - このあたりのAS運用のノウハウについてはHOMENOCさんのBoFでお話ししましょう……!

AS63806の構成

- ネットワークの分割
  - 実トラフィックが通るネットワークと，管理用のネットワーク
  - ASとしてのバックボーンと，そこにつながる各サイトのネットワーク
    - バックボーンにつながる各サイトネットワークにはプライベートASNを割り振り，eBGPでバックボーンに接続
    - 当然，ASの外にはJPNICから割り当てを受けた全体の IPv4 /24 と IPv6 /48 しか漏れない (細かい内部の割り当てと，プライベートASNはフィルタされる)
  - 各サイトのLANを結ぶイントラネット
- ネットワーク品質監視システム
  - バックボーンのは公開されています：
    - https://looking-glass.nc.menhera.org/

AS63806の技術の概要

- BGP full routes を食わせられるまともなルータがないので, Linux ソフトウェアルータを多く使用 (FRR)
- インターネットフルルートを持っていないルータやスイッチには中古市場で入手したハードウェアも使用
- 専用線を持っていないので，拠点が東日本に集中していることから，拠点間の接続にNTT東日本のNGN上の仮想化された接続を使用 → このあと技術的な面について詳述
- 仮想ルータや仮想サーバ，OpenvSwitch などを多く使用

iBGP+OSPFv3 を IPv6 link-local で動作させる手法の話

話題1/3

この設定で実現できること

- IPv4 と IPv6 の両方で iBGP ピアを張る必要がなくなるので管理が楽
- IPv4 と IPv6 の両方で OSPF ネイバーを張る必要が無……(ry
- IPv4 アドレスは各ルータ1個の割り当てでいいので，場合によってはかなりアドレスを節約できる

ヒントを得たもの

- 最近流行りの Clos アーキテクチャ，　fabric などの話
  - → バックボーンの iBGP にも活かせないか？？
- RFC5549 関連の過去の JANOG の発表

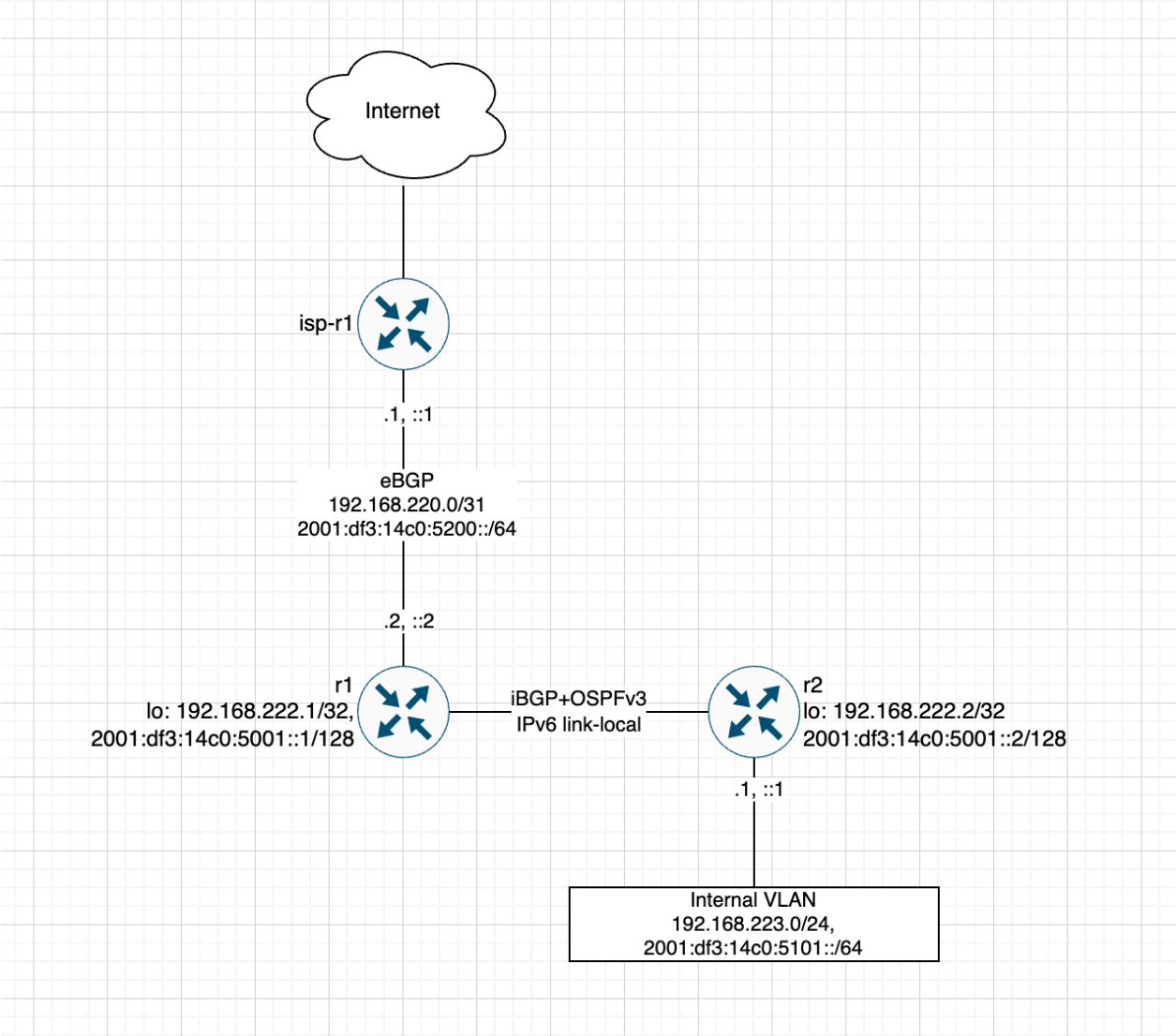
TL;DR

- ASのバックボーンで，IGP として OSPFv3 (IPv6) のみを走らせる
- その上で iBGP IPv6 ピアを張る
- RFC5549 を使えば IPv6 とその link-local nexthop を使って IPv4 の到達性も確保できる
- IPv6 でしか IGP/iBGP をやっていないので管理が楽
- IPv6 link-local 以外の実グローバル IP アドレスは IPv4/IPv6 それぞれループバックの1個づつのみ

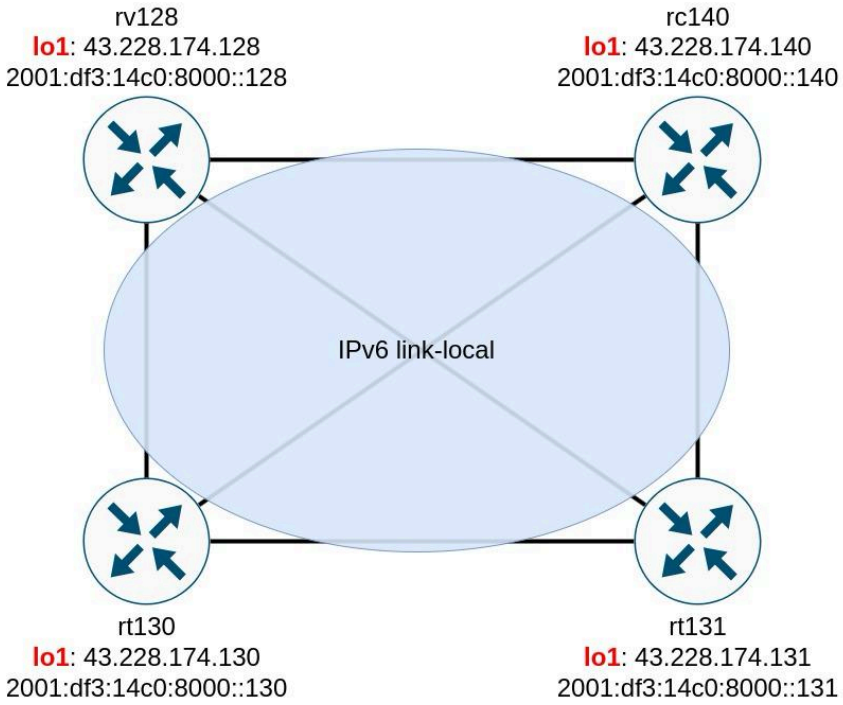
Qiita 記事（私の）

- https://qiita.com/metastable-void/items/08eb666e11b5502e0df8

概念図


Internet
isp-r1
.1, ::1
eBGP
192.168.220.0/31
2001:df3:14c0:5200::/64
.2, ::2
r1
lo: 192.168.222.1/32,
2001:df3:14c0:5001::1/128
iBGP+OSPFv3
IPv6 link-local
r2
lo: 192.168.222.2/32
2001:df3:14c0:5001::2/128
.1, ::1
Internal VLAN
192.168.223.0/24,
2001:df3:14c0:5101::/64

この仕組みで，AS63806 の実際のバックボーンで運用成功



フレッツ光の上にWireguard+BGP EVPN L2VPN を組んで暗号化された広域イーサネットを構築する話

話題2/3

前提

- 予算的に可能な組織は普通に通信会社の提供する広域イーサを使って構わない
  - → ここではそうでない場合，あるいは何らかの理由がある場合の話です
- フレッツ光などの閉域網につながった拠点が複数ある前提
  - インターネットでもできるが，品質が落ちる
- FRR をここでは使いました
- 暗号化はオプション，しかし FRRouting では IPv6 で直に EVPN L2VPN を張れませんでした……orz
  - 誰かやる気のある方がプルリクを送らないと実装されない？
- L2 接続性がほしい，あるいはそのほうが管理しやすい場面や場合に使用

Wireguard について

- 最近いろいろなものに利用されている L3VPN スタック
- 暗号化がいい感じにいろいろな環境でソフトウェアで高速につかえる
- ルーティングをポリシーとして静的に書きこんでおいて，それを MAC アドレスみたいにしてそれをもとに転送を行うので，動的ルーティングと非常に相性が悪い
- 2サイト間のピア2ピア接続ならば，すべてのIPアドレスが通るようにすることも可能だが，Wireguard の本領はひとつの UDP ポートだけをつかって，ひとつの仮想インタフェースで複数の接続先につなぐことだが，その場合任意のIPアドレスを通過させることができない

動機

- Wireguard で複数地点を結ぶと，Wireguard 上に設定されたアドレスでしか通信できなくなる → このアドレスを使用して，BGP EVPN をやればいいのでは
- せっかくなら EVPN L2VPN に

Set up 1/3: Wireguard

- まずフレッツ光等のアドレスに対して DDNS を設定しておく
- 公開鍵を生成
- このように，各拠点に1個のIPv4アドレス (IPv6 は FRRouring での BGP EVPN に対応していないのでいらない) を割りあてる
- 複数の拠点をひとつの接続で結んでしまって OK

```
[Interface]
PostUp = wg set %i private-key /etc/wireguard/%i.key
Address = 10.200.0.71/24
ListenPort = 50100
Table = off

[Peer]
PublicKey = ...
AllowedIPs = 10.200.0.11/32
Endpoint = ddns2.example.org:50100

[Peer]
PublicKey = ...
AllowedIPs = 10.200.0.12/32
Endpoint = ddns3.example.org:50100
```

Set up 2/3: L2/L3 VNI の作成

```
/etc/network/interfaces: (Debian/Ubuntu)

auto weth0
iface weth0 inet static
    pre-up /sbin/ip link add weth0 type bridge
    pre-up /sbin/ip link set weth0 addr 60:A4:34:43:e7:18
    pre-up /sbin/sysctl net.ipv6.conf.weth0.accept_dad=0
    post-down /sbin/ip link del weth0 type bridge
    address 10.200.1.71/24
    mtu 1370

iface weth0 inet6 static
    address 2001:df3:14c0:e001::71/64

auto vni101
iface vni101 inet manual
    pre-up /sbin/ip link add vni101 type vxlan dstport 4789 id 101 nolearning df unset
    pre-up /sbin/ip link set vni101 master weth0 addrgenmode none
    pre-up /sbin/ip link set vni101 type bridge_slave neigh_suppress on learning off
    post-down /sbin/ip link del vni101 type vxlan
    mtu 1370
```

Set up 3/3: FRRouting config

- 重要： bgp router-id はそのルータ(拠点)に割りあてた Wireguard IPv4 address
- wg100 は Wireguard インタフェース名

```
router bgp 4200000000
 bgp router-id 10.200.0.71
 no bgp default ipv4-unicast
 neighbor EVPN_VTEP peer-group
 neighbor EVPN_VTEP remote-as internal
 neighbor EVPN_VTEP update-source wg100
 neighbor EVPN_VTEP capability extended-nexthop
 neighbor 10.200.0.11 peer-group EVPN_VTEP
 neighbor 10.200.0.12 peer-group EVPN_VTEP
 neighbor 10.200.0.21 peer-group EVPN_VTEP
 neighbor 10.200.0.51 peer-group EVPN_VTEP
 !
 address-family l2vpn evpn
  neighbor EVPN_VTEP activate
  advertise-all-vni
  advertise-svi-ip
 exit-address-family
exit
!
```

Wireguard+BGP EVPN L2VPN (まとめ)

Menhera.org

BGP EVPN L2VPN

Wireguard (IPv4 L3VPN) → **暗号化**

フレッツ光などのアクセス線 (IPv6)

Anycast で小規模組織でも安価に冗長化されたスケールアウト可能なWebサーバ群を走らせる話

話題3/3

この話は少しだけです……

文脈

小規模組織等で単体の Web サーバを走らせるのはよくやる

→ 単体のサーバでやるのが現実的でなくなったら？

→ クラウド化するのがお金がかかってしまうような複雑・カスタム性が高いもの

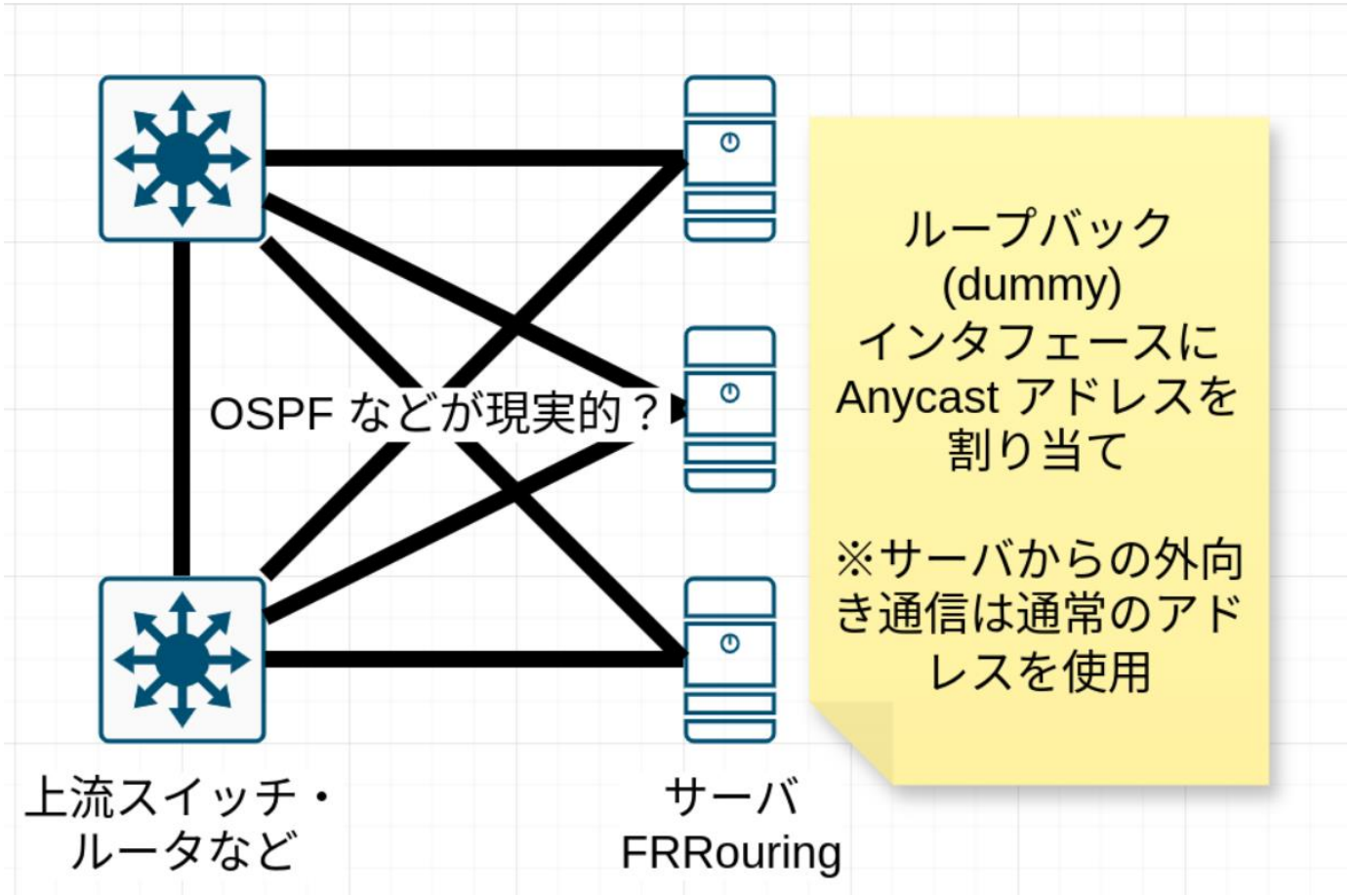
→ 複数のサーバをクラスタリングすることになる

- L3 で末端までやる方式が若干流行ったことがある気がする……
- FRR (FRRouting) のようなソフトウェアルータはサーバ機でよく利用される
- 安い汎用PCだけで少し重めの Web 的な負荷に耐えたい……

L3 で解決できること

- ロードバランサなどの上位レイヤで動くものが必要なことは確かに確実にある。
- しかし，単に複数のサーバが IP アドレスを共有するだけで構わない場合は多いし，そもそもアプリケーションの状態はそのようにスケールアウト可能であったほうがよい。
- IP anycast を使えば，共用の仮想IP (エニキャストアドレス) に対するリクエストを L3 で負荷分散可能
- DNS による負荷分散はうまくやるのが実はけっこう難しいので

割り当てるIPアドレスさえあれば，安い



負荷分散の階層

- 同じ拠点/DC内の複数サーバ間の分散 ← 今回はここも IP Anycast で
- 拠点間の分散 ← 近いところに接続
- (DNS 負荷分散)

余談・今後の課題

Open questions （ハード関連）

- インターネットルータ（フルルートを持つ）を仮想化するマシンはあまりネットワーキング以外の負荷と兼用しない方が良い？
  - ↑ 汎用仮想サーバと同じ物理マシンに収容したらそこのルータだけ遅い……？
- かといって Raspberry Pi 5 (8GB RAM) とかにフルルート食わせるのはかわいそう (Pi 4 の2倍の性能がある CPU らしいが……)
- i5 とかの省エネコンパクト自作PCに，グラフィックボードの代わりに個人向けPC的にはちょっと高いNIC (SFP+ x4 とか) を差したものが現状標準化している → どうなんでしょう？
  - ↑ メタル 10GbE の高密度な配線は無理があることがわかったため

現状の課題

- 頻繁にダウンする拠点がある
  - ハードウェア？ → 中古部品の自作PC，ラズパイ
  - 負荷かけすぎ？ → 性能が高くはない，いろいろ他の負荷が走ってる

→ 信頼性がまったくないとさすがにクリティカルじゃなくてもやってられないので，ちゃんとしましょう。

ありがとうございました

質疑応答 & フリートーク

- 非営利・営利問わず小さな組織のネットワークの主に技術面などについて
- 個人や非営利AS運用方面の話題に興味がある方はぜひ行かれてみてください↓
  - 関連 BoF：「アマチュアAS運用を議論するBoF」(一般社団法人 Home NOC Operators‘ Group 様，島根データセンター友の会 様)