IRS25
AS内でのBGPコンバージェンスの短時間化

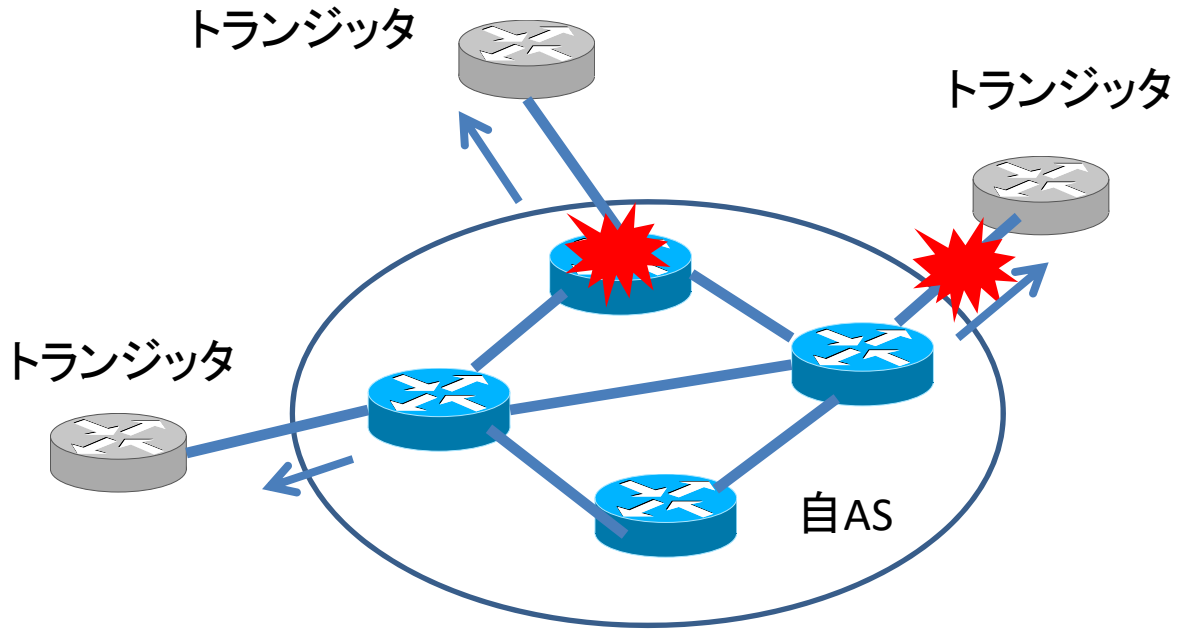
2016年09月20日

株式会社FORNEXT 篠宮

自己紹介

- 篠宮 俊輔
- 個人事業主 FORNEXT、株式会社FORNEXT
  2008年～
  自称ネットワークエンジニア
- ネットワークの技術面での支援を生業としています

本発表の話題



- 自ASから外へパケットを転送する方向
- 自ASのルータや、トランジッタとの間の線がダウンしたりで大規模なコンバージェンスを話題に

本発表の概要

- AS内でのBGPのルータダウン、リンクダウン時にコンバージェンスで時間がかかってしまう原因をざっと列挙
  - その中で、ネクストホップ属性がらみに注目
- ネクストホップ(NEXT_HOP)属性での経路の無効化の話をしたかった。
  - BFD使えば良いじゃん、fast external fail over使えば良いじゃん、というのもありますが。
- BGP PICのデモを(できたら)して終了

想定以上にコンバージェンスに時間がかかった?

- フルトランジットが1本切れたら、長い時間、到達できないネットワークがあったようだ
  - 出の広告も止まったからAS外が原因だったんじゃないか? うちは悪くない(希望的観測)、とうやむやにしてしまったり。
- ルータが1台落ちたら(以下略)
- AS内のリンクが1本落ちたら(以下略)

- 高い頻度で起こる事象ではないので、原因の追及が難しい
  - コンバージェンスに時間がかかったようです、で済ませてしまう?
- ネットワークが複雑なので、再現環境が作れない
- 実環境で試そうにも試しづらい
- 起きたときに観察しようもに、それも難しい

コンバージェンスに時間がかかる原因

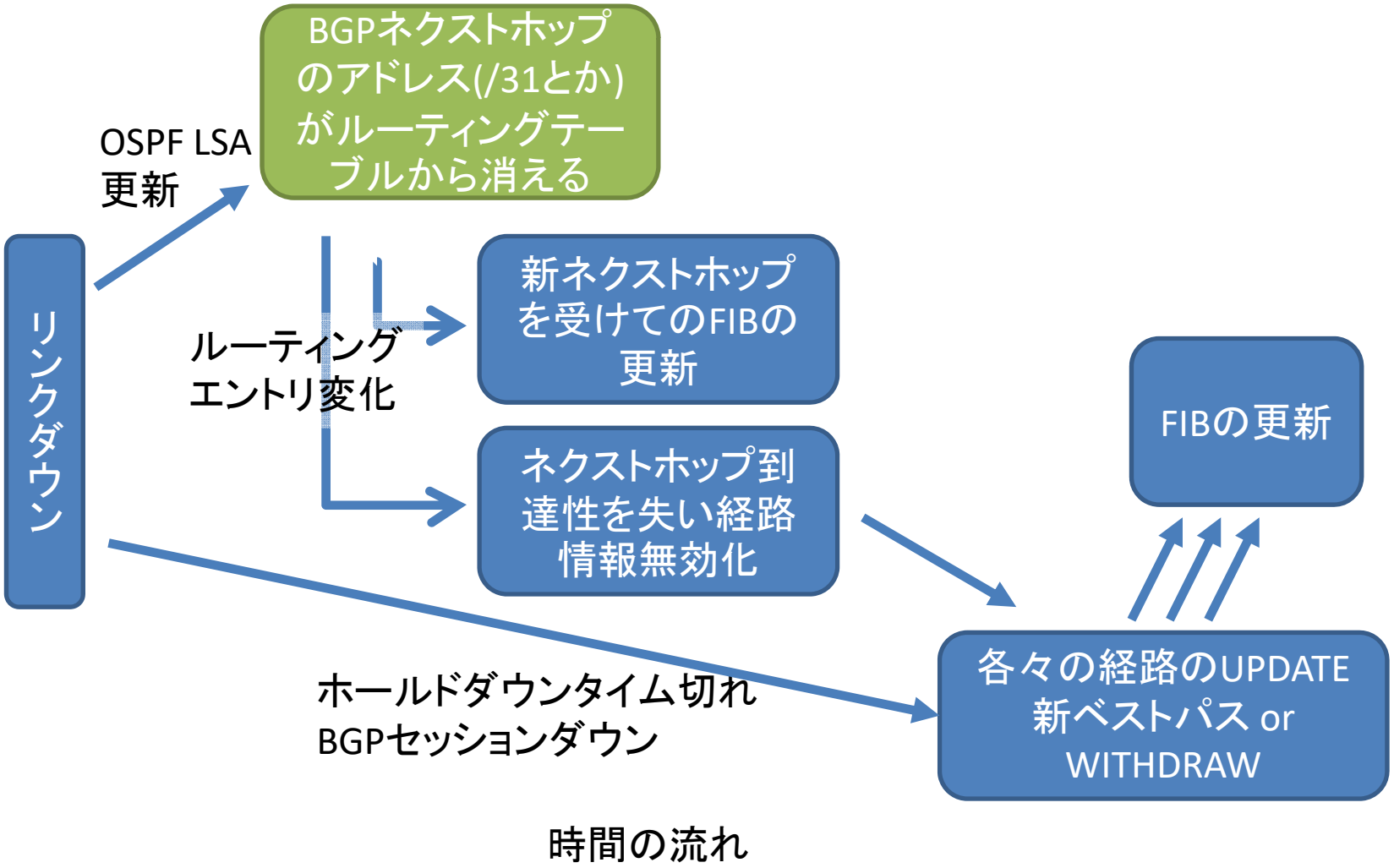
- 経路がたくさんある
  - 2016年9月現在で、フルルートは62万経路くらい
- 経路の処理には時間がかかる
  - 経路の受信、送信
  - ベストパスの決定
  - FIBへのインストール

  } 1000～1万経路/秒くらい?
- ダウン発生後、一発でベストパスが決まらなく、何度も経路の処理を繰り返す
- ダウン発生後の、BGPセッションダウン(そこからの経路取り消し)のタイムアウト待ち、WITHDRAW待ち。

コンバージェンス中に起きてしまう嫌なこと

- 転送しようとしたパケットの転送先がFIBにないのでパケット破棄
- 転送したパケットの転送先に受け取ってもらえず(転送先のルータ、リンクが落ちている)
- 隣のルータにパケットを転送したけど、隣のルータが送り返してきた(ピンポン)
- ホールドタイム切れまでの時間が長くてもどかしい
- なんか期待しない宛先にパケット投げた

対外とのリンクダウン等で起こること



経路情報のネクストホップ(NEXT_HOP属性)での経路情報の評価

- ベストパス選択アルゴリズムの最初や、「前提」にあるルール。

OS	記述
IOS	Why Routers Ignore Paths Paths for which the NEXT_HOP is inaccessible.
IronWare	1. Is the next hop accessible though an Interior Gateway Protocol (IGP) route? If not, ignore the route.
JUNOS	1. Verify that the next hop can be resolved.

- 「アクセスできる」、「アクセスできない」、「解決できる」って何?
  - 多くの実装で、ネクストホップを解決してみたらその学習元が、直接接続、スタティック、ダイナミックルーティング(BGPを含む)であると有効。
  - デフォルトルートで解決した場合は、有効にならない実装も
  - 学習元がBGPの場合は有効にならない実装も
- 有効でなければ、その経路情報は無効な情報として扱う。
- 経路情報が有効なのかを決める、BGPのメッセージを契機にしないトリガー。

ルータ、リンクが落ちた直後に経路のネクストホップが向く例

直後: ルータなりリンクが落ちてから、関係する経路がUPDATEされるまでの間

NEXT_HOP属性のアドレス	経路が向いた先	原因
自ASのアドレスブロック	null。 →全力でパケット捨てちゃった	自ASで使っているアドレスブロックをnullに向けるスタティック経路で解決してしまった
トランジッタAのアドレス	他のトランジッタB。 → 短時間だが、トランジッタBにたくさん流れた	トランジッタAのアドレスブロックのベストパスが、トランジッタB経由だった。
	IXでのピアなど。 →ピアから貰っていない経路宛てのパケット投げちゃった!	トランジッタAのアドレスブロックのベストパスが、ピア経由だった。
-	どこにも向かない(無効になる)	NEXT_HOPが有効では無くなり、経路情報が無効になる

有効なネクストホップの制限

OS	機能
IOS系	選択的 BGP ネクストホップ ルート フィルタリング route-mapやroute-policyで、有効なネクストホップの制限ができる。
IronWare	篠宮の知る限り、コンフィグではできない。 BGPで解決した場合は無効な経路となるので、それを上手く使う
JUNOS	policy-statementで、route resolution databaseへ書き込む経路を制限できる。 (副作用で、BGPなどメモリ消費が多い経路を制限すれば、空きメモリも増える)

- これにより、前述のリンクやルータダウン時に経路が一時的に変な方向を向くことを回避できる
- このネクストホップの制限を正しく使えば、ルータダウン、リンクダウンをイベントとして、BGP UPDATEに依らないイベント通知ができる。

数十万経路分BGP UPDATE vs IGPで1アドレス分伝搬

注意!

- 不用意に、JUNOSでshow route resolutionコマンドを叩かないでください。
- route resolution databaseの中身を表示するコマンドshow route resolutionを実行すると、ルーティングデーモンが落ちる不具合(PR1023682)のあるJUNOSがあります。(12.3R7、12.3R8、13.3R1～13.3R4など)

ネクストホップの判定で
正しく経路を無効にできると?

- 明後日の方向にパケットを投げる段階を減らせる
- nullにパケットを投げる段階を減らせる
- WITHDRAWによる経路取り消しを待たずに、ベストパスを送り始めることができる
- BGP PICが良く動く

コンフィグ例

IOS

```
# ネクストホップは、直接接続とOSPFの経路だけが有効

route-map rmap_bgp_nexthop permit 10
 match source-protocol connected
!
route-map rmap_bgp_nexthop permit 20
 match source-protocol ospf 1
!
router bgp xxx
 address-family ipv4
  bgp nexthop route-map rmap_bgp_nexthop
 exit-address-family
!
```

JUNOS

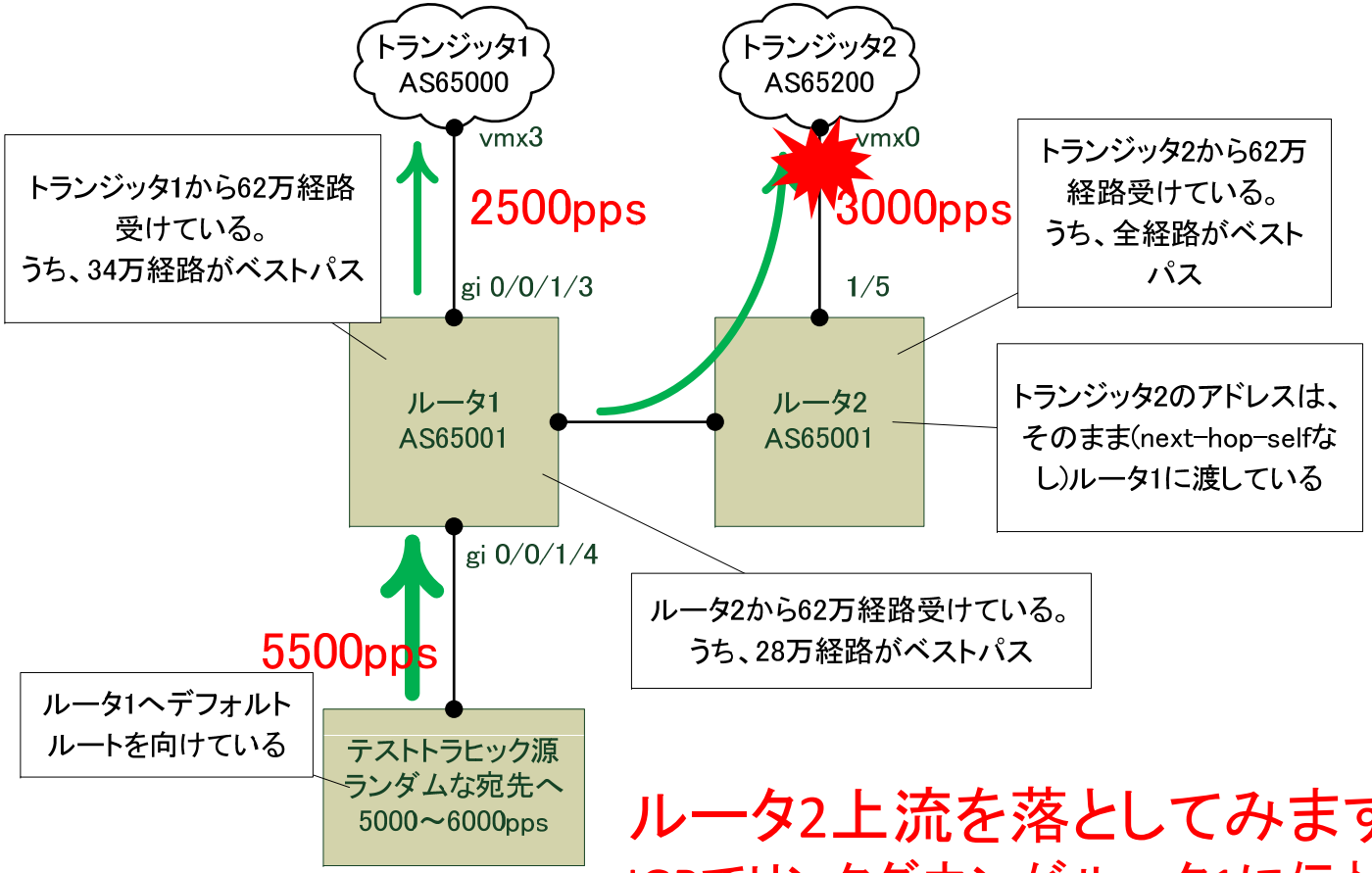
```
# ネクストホップの解決にスタティック、BGPは使わない

policy-options {
   policy-statement next-hop-resolution {
      term static {
         from protocol static;
         then reject;
      }
      term bgp{
         from protocol bgp;
         then reject;
      }
   }
}
routing-options {
   resolution {
      rib inet.0 {
         import next-hop-resolution;
      }
   }
}
```

BGP PIC

- Prefix Independent Convergence
- RIBから、ベストパス、第二案となる経路を選出。
- FIBには、ベストパスのネクストホップに加え、第二案となるネクストホップをインストールしておく
- ベストパスのネクストホップが無効になったら、問答無用で第二案での転送に切り替える
  - その後は普通にベストパスセレクション&FIBに反映

BGP PICのデモ

アクセリアさんの環境使わせて頂いております



ルータ2上流を落としてみます。
IGPでリンクダウンがルータ1に伝わり、
ルータ1がBGP PICで一気に切り替え。