IRS28
iBGPでのBGPネクストホップの話
(next-hop selfするしない話)

2018-03-09

株式会社FORNEXT 篠宮

事後資料

自己紹介

- 篠宮 俊輔 (しのみや しゅんすけ)
- 株式会社FORNEXT
  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
  - ネットワーク設計、構築
  - AS運用(の技術的な部分)

発表内容

- 今回も、IRS25に続きネクストホップの話です
  - 同じIXに二つ目のインタフェース接続する作業をして思ったことがあるので共有、意見交換
  - 発表者(篠宮)はnext-hop selfは基本的にしない派です
- 今回も、IRS25同様、BGP PIC edgeが出てきます

next-hop selfするしない - する理由。

- ネクストホップ(対外との接続のリンクに付けたアドレス)が、プライベートアドレス、リンクローカルアドレス
- IGP(広義)に、他ASのプレフィックスを乗せたくないから
  AS内経路は皆BGPに載せるネットワークだとよりある?
- IGPの経路をなるべく少なくしたい
  対外接続の本数だけ、経路が増える
- ネクストホップが意図しない方向に向く事への対処
  - ネクストホップが向く先が他の方向に誘導されてしまうような経路の注入事故、事件
    - /30でトランジッタ、ピアと接続しているけど、/30、/31、/32の経路をBGPで受けた
    - IXのセグメント(/24や/23)より細かい経路をBGPで受けた。/23になったIXがあるので、より注意。

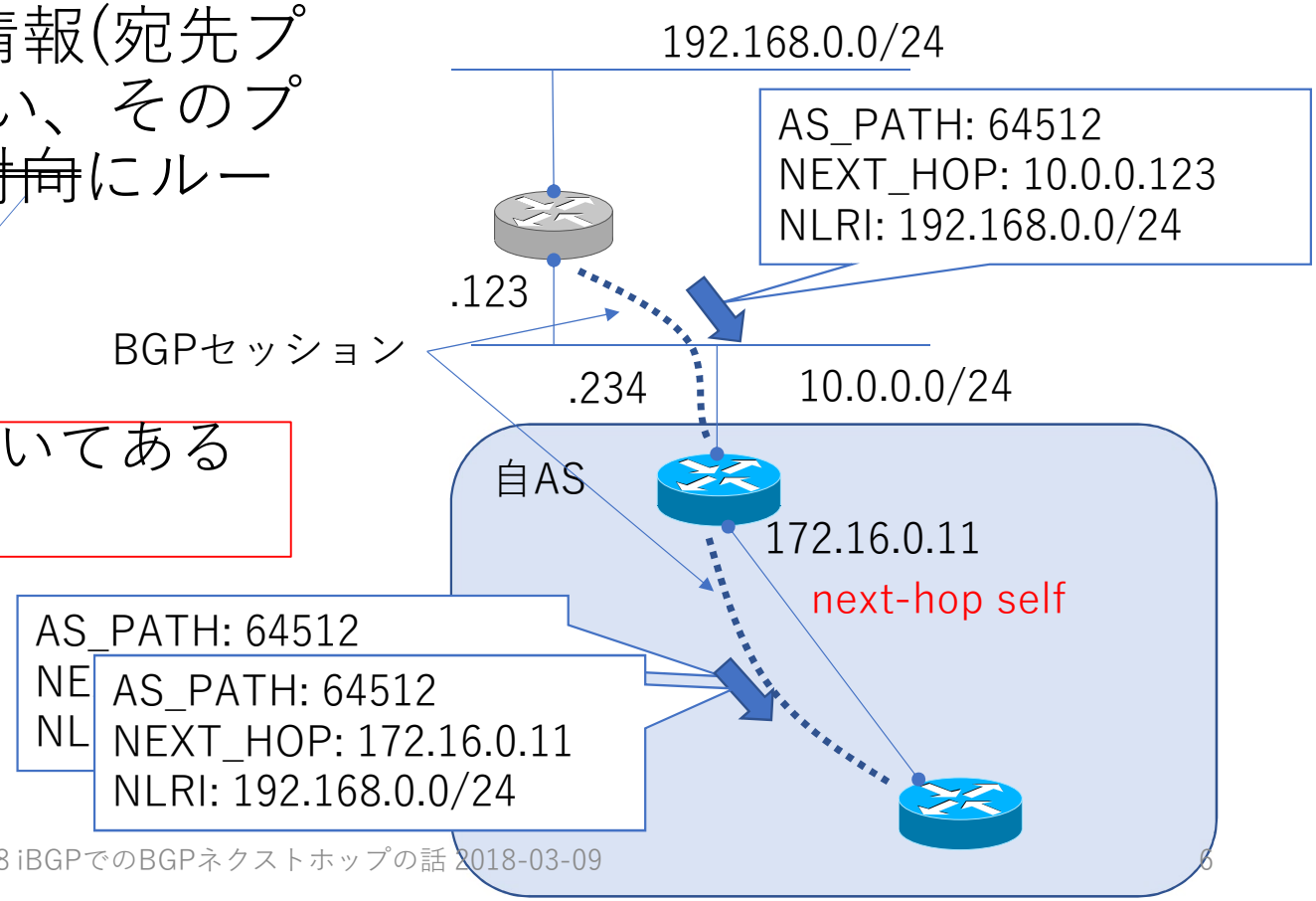
next-hop selfするしない - しない理由。

- 対外接続のリンクダウンをIGPで伝えたい
  - このリンクのアドレスへの到達性は、IGPでAS内に伝える
- ベストパスを決める際に使えるパラメータはなるべく維持したい(減らしたくない)
- BGP PIC edgeが活きる場面を増やしたい
  - next-hop selfは、ネクストホップを集約、隠してしまうのでPIC Edgeが発動できる場面が減る

ネクストホップとnext-hop selfのおさらい

- BGPで対向から経路情報(宛先プレフィックス)をもらい、そのプレフィックス宛ては~~対向~~にルーティング

- NEXT_HOP属性に書いてあるアドレスへ



BGP PIC edgeおさらい

- Prefix Independent Convergence
- RIBから、あらかじめベストパス、バックアップパスとなる経路を選出
  - バックアップは、異なるネクストホップとなるように選出
- FIBには、ベストパスのネクストホップに加え、バックアップパス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、問答無用でバックアップパスに転送に切り替え、パケット転送は継続。
  - その裏で、普通にベストパスセレクション(RIB更新)→FIBに反映

BGP PIC補足
(事後資料で追記)

- BGP PICには、BGP PIC core、BGP PIC edgeの2種類が有る
- ID(draft-ietf-rtgwg-bgp-pic-07)から引用

> 6.1. BGP-PIC core
>
> This section describes the adjustments to the forwarding chain when a core link or node fails but the BGP next-hop remains reachable.

> 6.2. BGP-PIC edge
>
> This section describes the adjustments to the forwarding chains as a result of edge node or edge link failure.

- 本発表で話題としているのは、BGP PIC edge。エッジノード、エッジリンクの故障時に働く方

BGP PIC edge: JUNOSでの経路表示例

192.168.0.0/24

ベストパス

```
@[BGP/170] 00:01:22, MED 250, localpref 1000
   AS path: 64512 I
validation-state: unverified, > to 10.0.0.123 via xe-0/0/0.0
```

@ = Routing Use Only
# = Forwarding Use Only

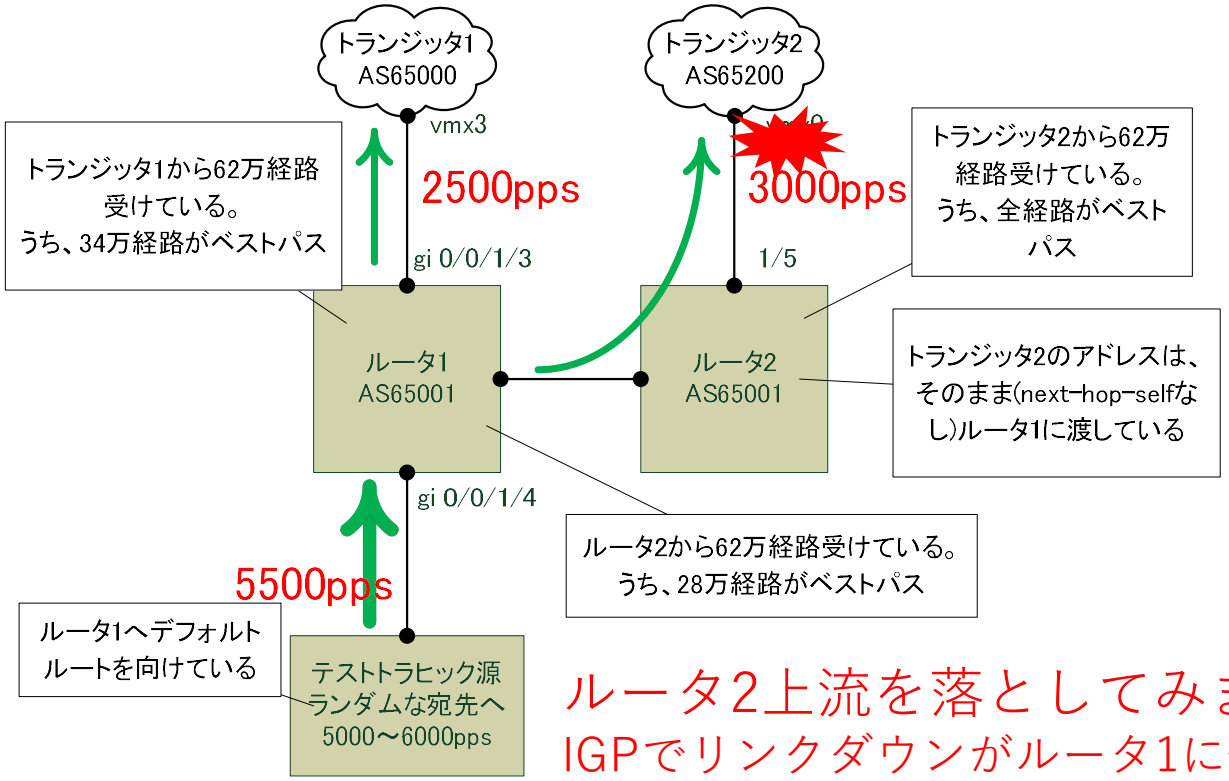
```
[BGP/170] 00:02:14, MED 250, localpref 1000, from 隣のiBGPのルータ
   AS path: 64512 I
validation-state: unverified, > to 10.0.0.123 via xe-0/0/0.0
```

バックアップ

```
[BGP/170] 1w0d 04:45:42, MED 350, localpref 1000
   AS path: 65500 64512 I
validation-state: unverified, > to 10.255.0.1 via xe-0/0/1.0
```

```
#[Multipath/255] 00:01:22, metric 250
  > to 10.0.0.123 via xe-0/0/0.0
    to 10.255.0.1 via xe-0/0/1.0
```

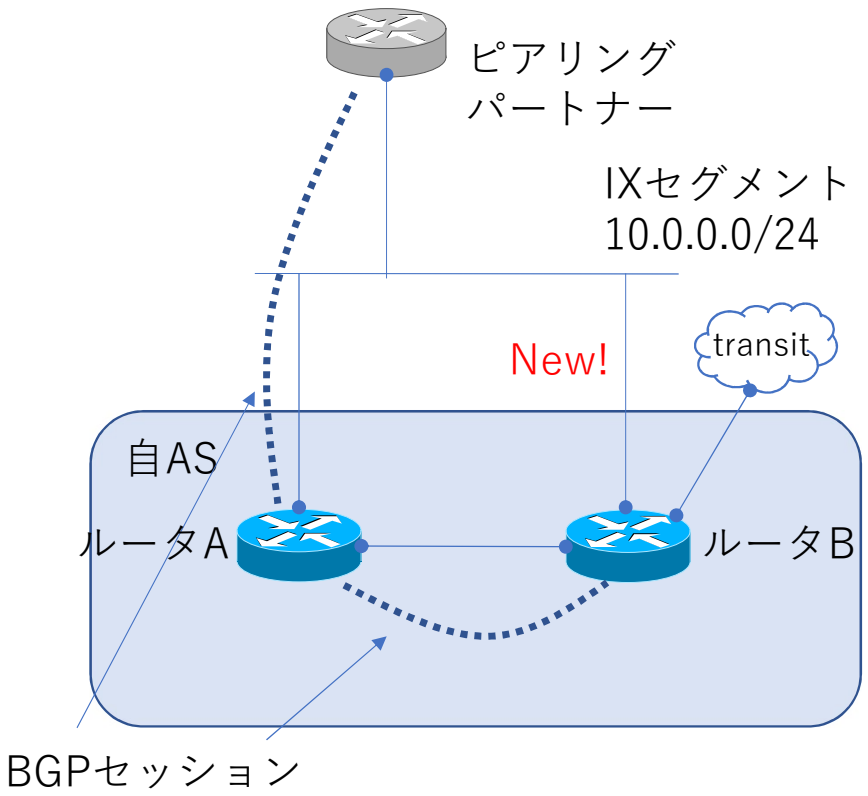
BGP PIC edge(のデモの時使った図)



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 edgeで一気に切り替え。

ここから本題

- 基本的にはnext-hop selfしない設計をしている人(発表者)が、IXに二つ目のアドレス、インタフェースで接続した際に、いろいろ思うことがありました。

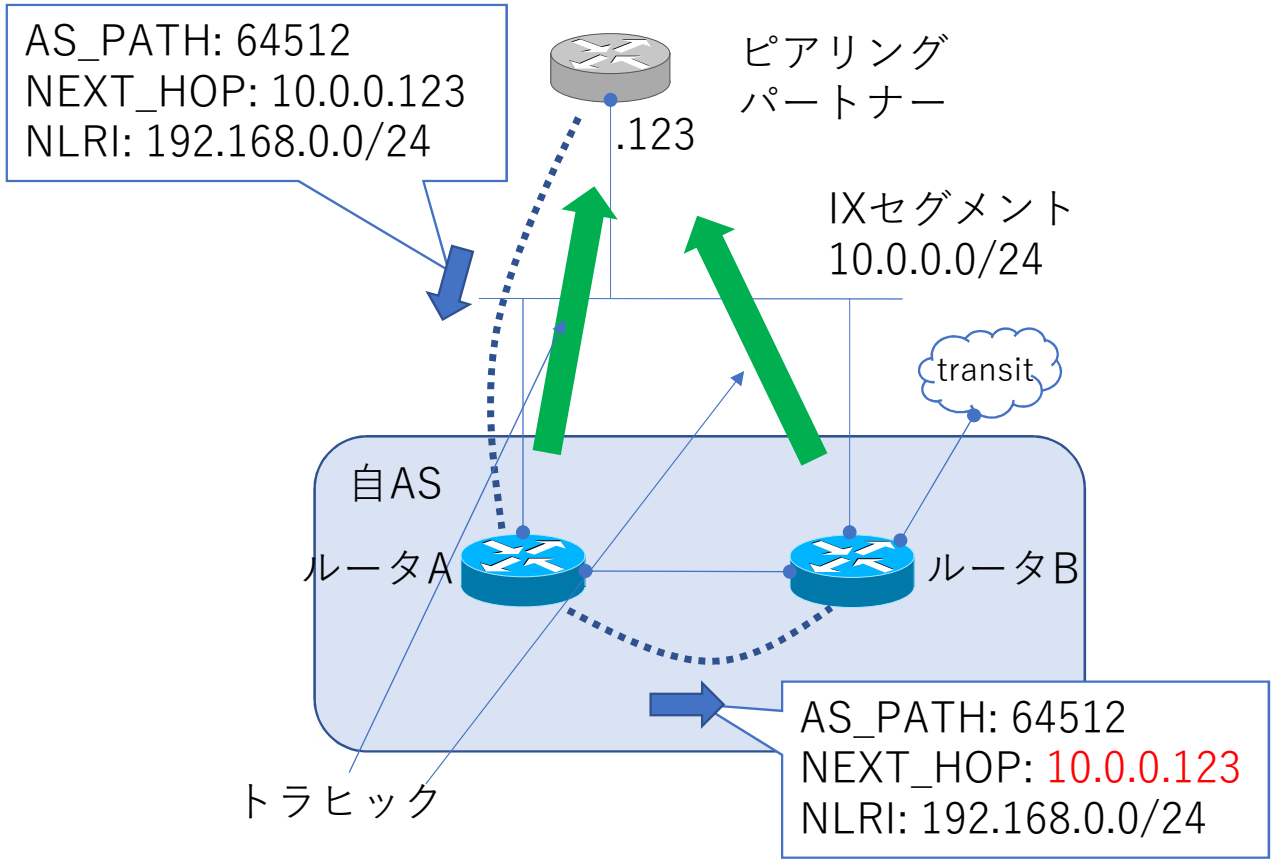


next-hopしないと

罠?1	ピアリングしていないルータから、トラヒックを送る
罠2	1本目、2本目どちらでもピアリングしているピアとは、1本落としてもトラヒックが流れ続ける
罠3	IXへの開通作業時のIX事業者の確認用ネットワークに接続するとパケットロス
罠4	IX内でユニキャストフラッディング (川上さんの資料参照)

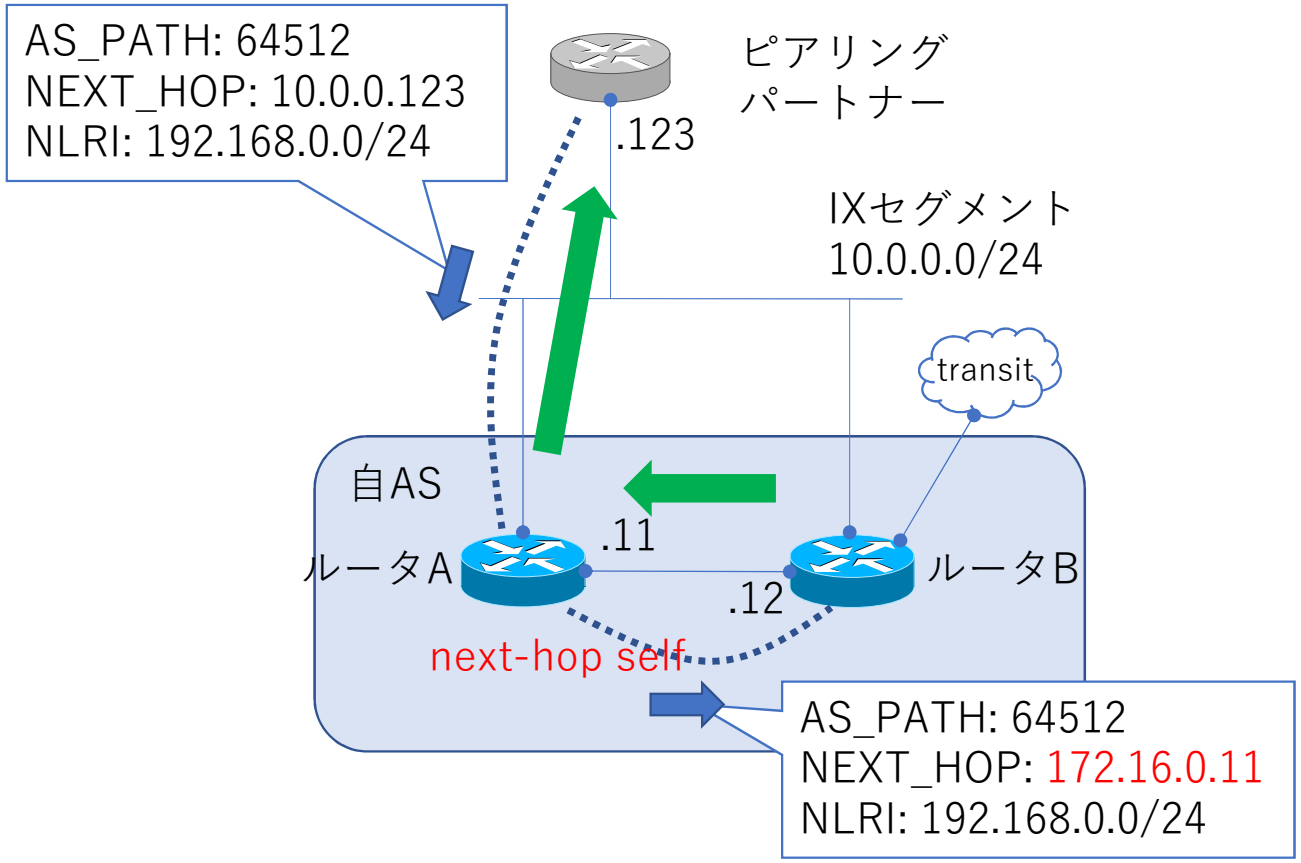
罠?1: ピアリングしていないルータから、トラヒックを送る(1)

- ルータBは、ルータAから受け取った192.168.0.0/24の経路情報に従ってパケット転送。
- ルータBは、ピアリングしていないのに。
- 必ずしも、悪いことではない。ピアリングパートナーとの間で調整つけばOK。



罠?1: ピアリングしていないルータから、トラヒックを送る(2)

- next-hop selfすると、ルータBはルータA経由でトラヒック送信。



突然ですが、質問です

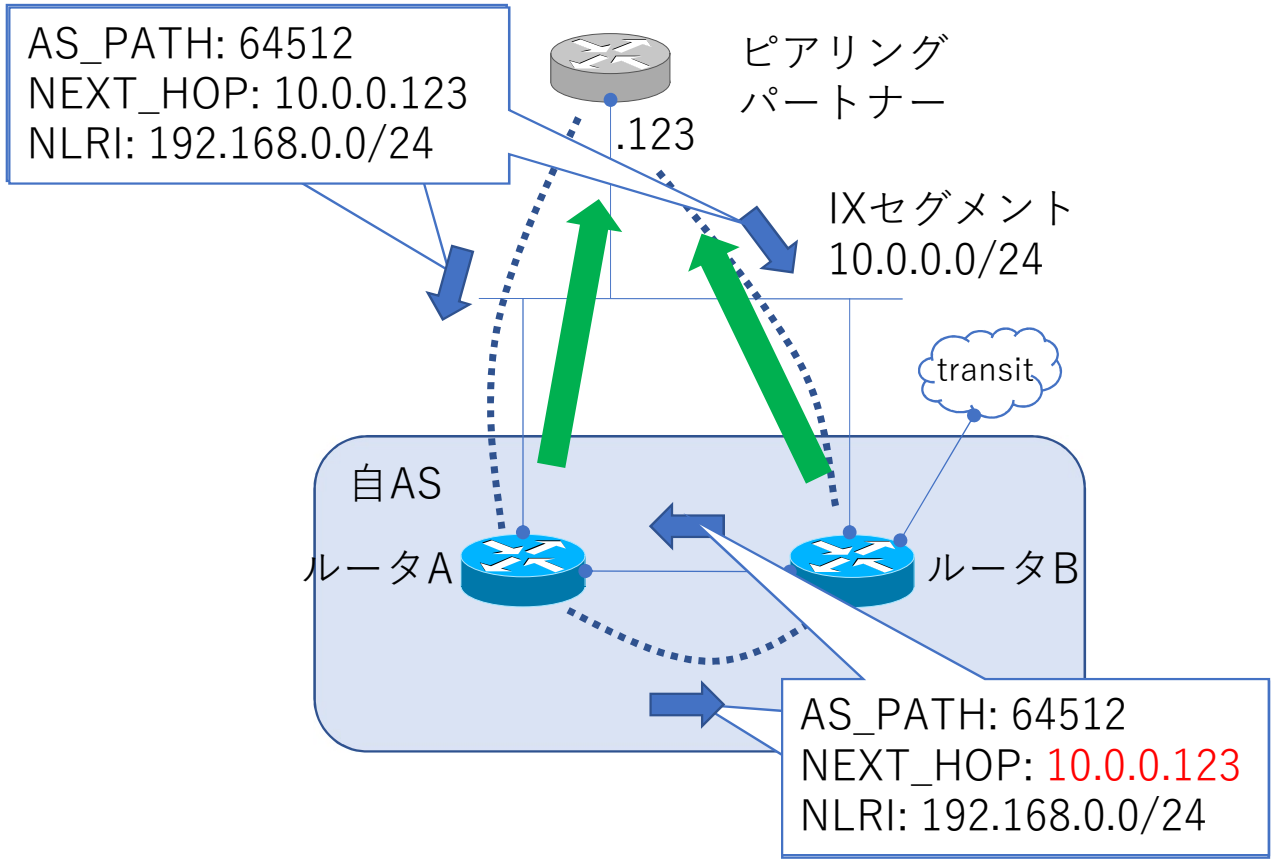
- 「ピアリングパートナー」の立場で、先ほどのようにBGPセッションを張っていないルータ(B)からトラヒックが送られてきたらどう思います? どうします?

どれでしょう?
ルータAとはピアリングしてるのだから、OK。
気にしていなかったけど、OK。
流れるのでしょうがないので、OK。
流されたらしょうがないけど、だめ。
だめ。
だめ。パケットフィルタ、レートリミットしている。

参加者からは、明確に「だめ」という声はありませんでしたが、上記トラヒックが送られてくるのは気持ち悪い、という声が多かったです。

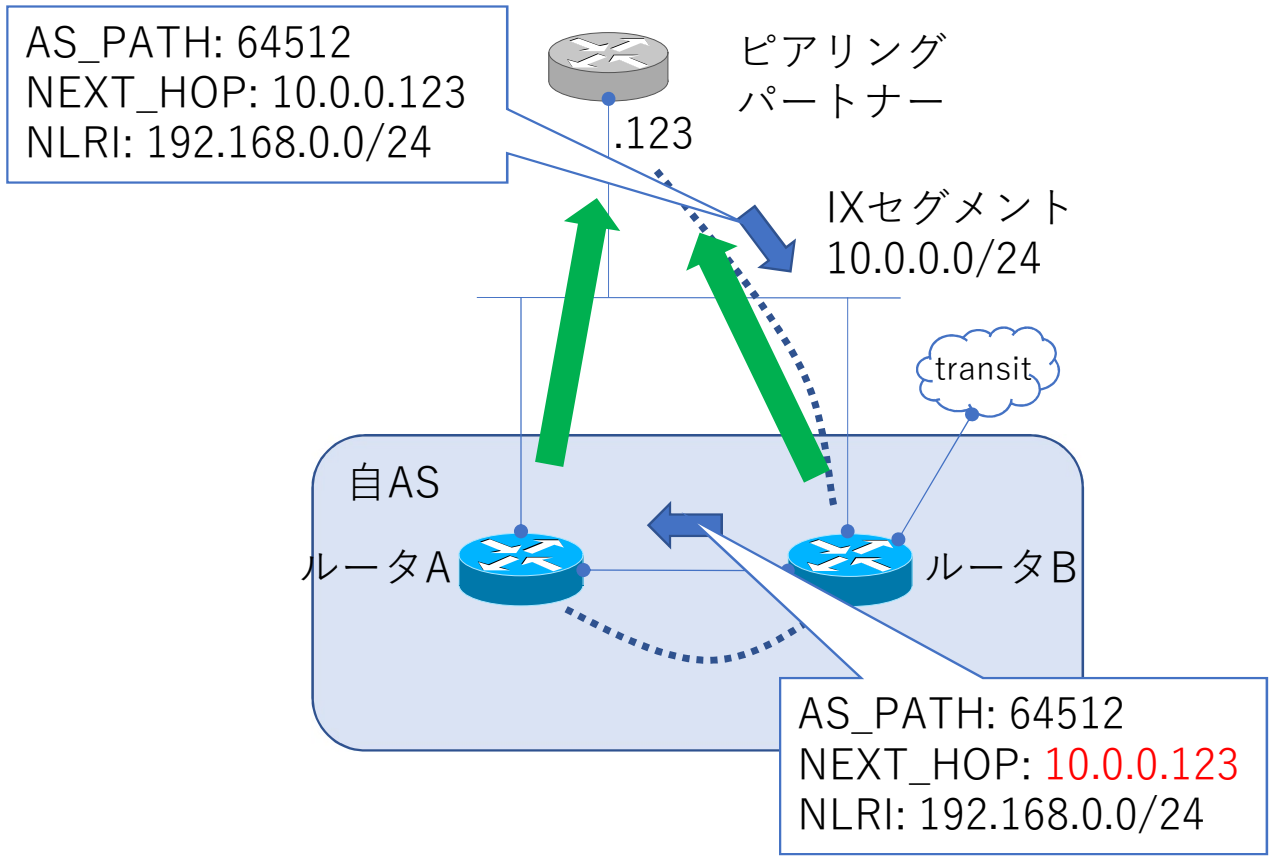
罠2: 1本目、2本目どちらでもピアリングしているピアとは、1本落としてもトラヒックが流れ続ける(1)

- 今度は、ルータA、Bどちらもピアリングしています。
- メンテナンスのため、ルータAのIX接続を切りたいな、トラヒック迂回させようかな、と思ったとする



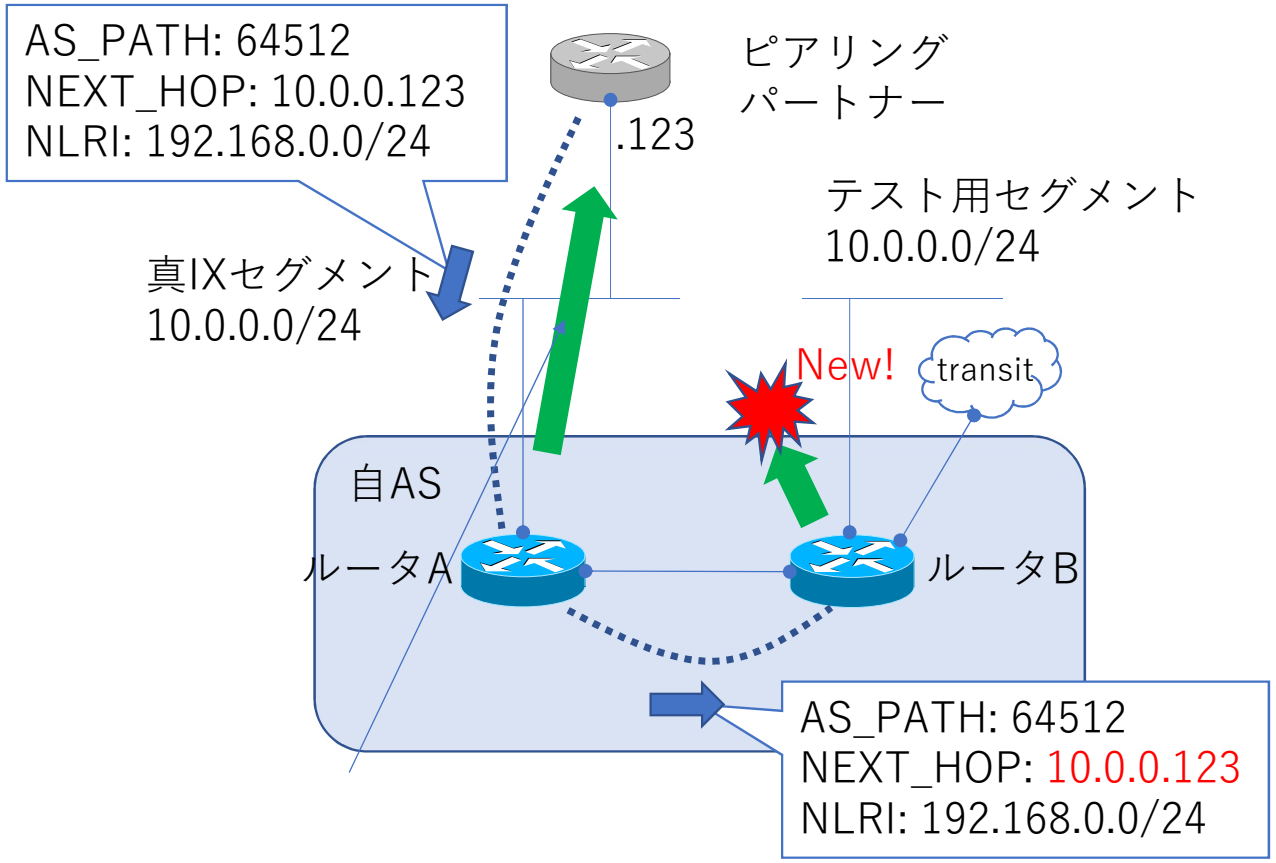
罠2: 1本目、2本目どちらでもピアリングしているピアとは、1本落としてもトラヒックが流れ続ける(2)

- 今度は、ルータA、Bどちらもピアリングしています。
- メンテナンスのため、ルータAのIX接続を切りたいな、トラヒック迂回させようかな、と思ったとする
- ルータAのBGPセッション落としても、相変わらずトラヒック送信。



罠3: IXへの開通作業時のIX事業者の確認用ネットワークに接続するとパケットロス

- IXに接続も受けるとき、利用者がアドレスの付け間違い、変なパケット送ってないか確認のため、テスト用セグメントに接続させるIX事業者さんいますね。



IXとnext-hop self

- 確かに、IXへの接続はnext-hop selfした方が良いかも
- でも、BGP PIC edgeが・・・
  - IX接続で扱う経路数はトランジットより少ないから、まぁ、良いかな。
  - トランジットはnext-hop selfしない(BGP PIC edgeのため)
  - IXから受けた経路は、next-hop selfする(罠回避のため)
- そのためには、iBGPネイバとの設定で行うnext-hop selfではなく、出の経路ポリシで選択的にnext-hop selfが良いかな?
  - 罠?1で取り上げた、BGPセッション張っていないルータからトラヒックを送るのをOKとするピアリングパートナーもいらっしゃいますし。

next-hop selfとBGP PIC edge

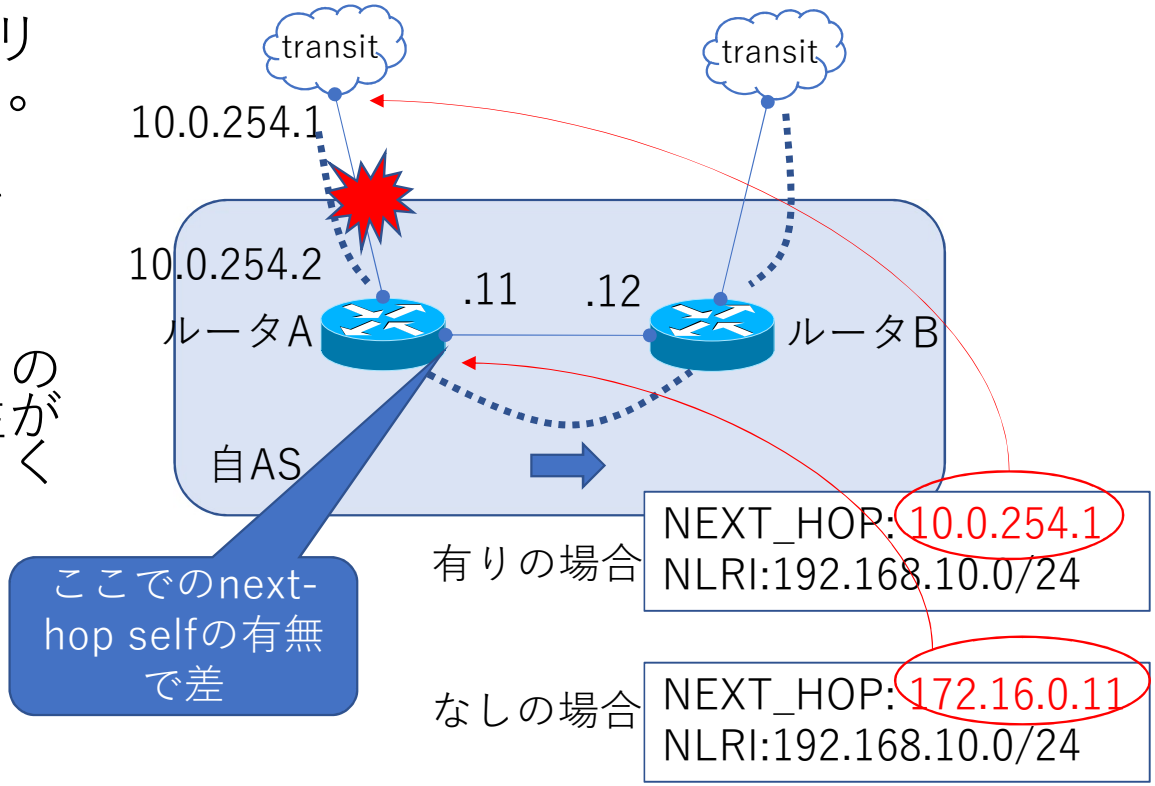
- BGP PIC edgeは、ネクストホップへの到達性がなくなったら発動
- バックアップパスとしてベストパスのとは異なるネクストホップのパスを選ぶ

- next-hop selfでネクストホップを変える(集約)してしまうと、発動できる場面が減る
- バックアップパスがいまいちになってしまう

相性悪そうですよね。

next-hop selfにより、BGP PIC edge発動の場面が減る

- ルータBの視点で。
- ルータAのトランジットとの間のリンクがダウンした場合を例とする。
- ルータAがルータBに渡す経路にnext-hop selfせすると、発動できない。
- iBGPはループバックアドレスで張ったりするので、その場合、このループバックアドレスへの到達性がなくなるのはルータが落ちたときくらい。



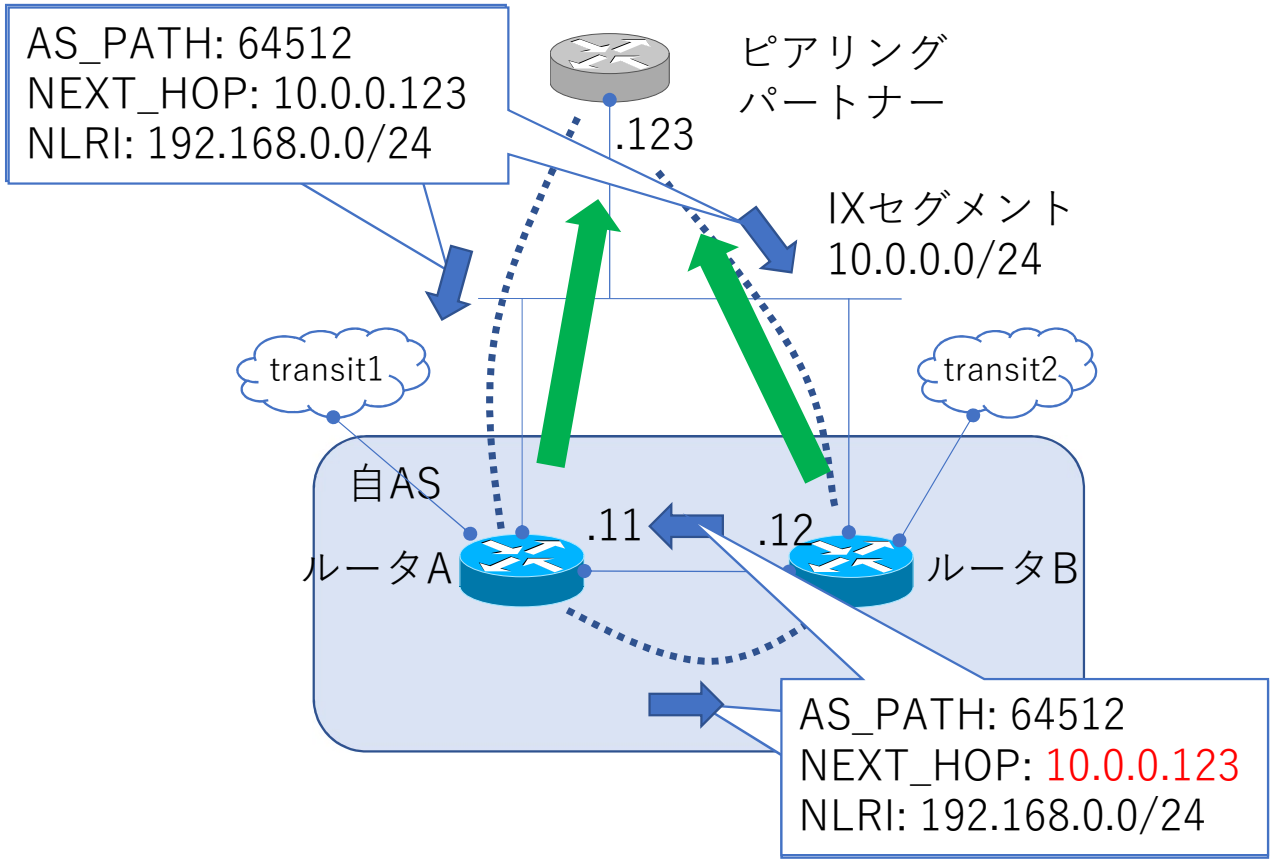
例:next-hop selfでBGP PIC edgeのバックアップパスがいまいちに: next-hop selfしない場合

```
192.168.0.0/24
        @from 10.0.0.123
ベストパス  AS path: 64512 I
        > to 10.0.0.123 via xe-0/0/0.0

        from 172.16.0.11
        AS path: 64512 I
        > to 10.0.0.123 via xe-0/0/0.0

        from 10.254.2.1 (transit2)
        AS path: 65502 64512 I
バックアップ  > to 10.254.2.1 via xe-0/0/1.0

        #[Multipath/255]
        > to 10.0.0.123 via xe-0/0/0.0
        to 10.254.2.1 via xe-0/0/1.0
```



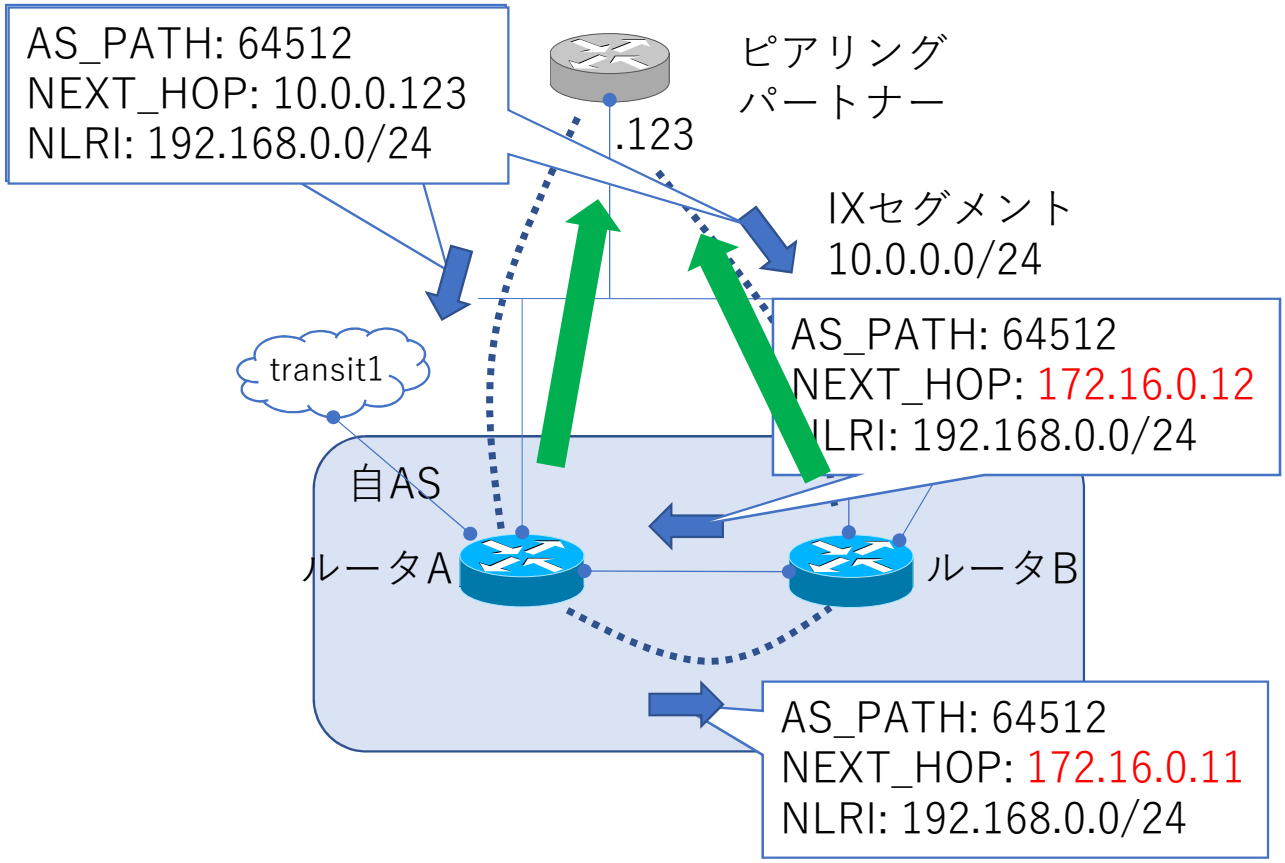
例: next-hop selfでBGP PIC edgeのバックアップパスがいまいちに: next-hop selfする場合

```
192.168.0.0/24
        @from 10.0.0.123
ベストパス  AS path: 64512 I
        > to 10.0.0.123 via xe-0/0/0.0

        from 172.16.0.11
バックアップ AS path: 64512 I
実はルータA  > to 172.16.0.11 via ae0.0
でのIX経由

        from 10.254.2.1 (transit2)
        AS path: 65502 64512 I
        > to 10.254.0.2 via xe-0/0/1.0

        #[Multipath/255]
        > to 10.0.0.123 via xe-0/0/0.0
        to 172.16.0.11 via ae0.0
```



例: next-hop selfでBGP PIC edgeのバックアップパスがいまいちに: next-hop selfする場合

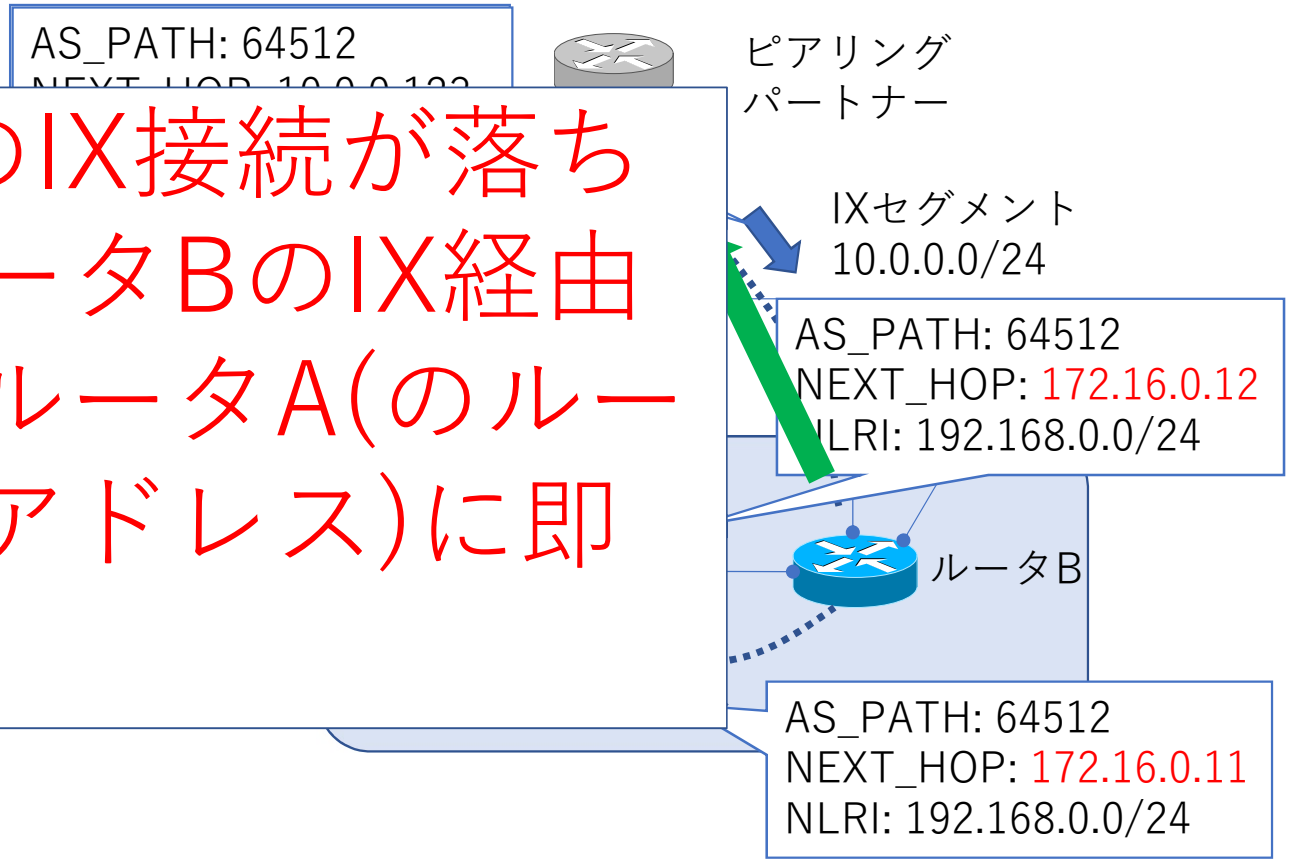
```
192.168.0.0/24
        @from 10.0.0.123
ベストパス  AS path
        > to 10

        from 17
バックアップ AS path
実はルータA  > to 17
でのIX経由

        from 10
        AS path
        > to 10

        #[Multi
        > to 10.0.0.123 via xe-0/0/0.0
        to 172.16.0.11 via ae0.0
```

ルータBのIX接続が落ちると、ルータBのIX経由の経路はルータA(のループバックアドレス)に即座に向く



例: next-hop selfでBGP PICのバックアップパスがいまいちに: next-hop selfする場合

192.168.0.0/24

@from 10.0.0.123

ベストパス AS path

> to 10

from 17

バックアップ AS path

実はルータA > to 17

でのIX経由

from 10

AS path

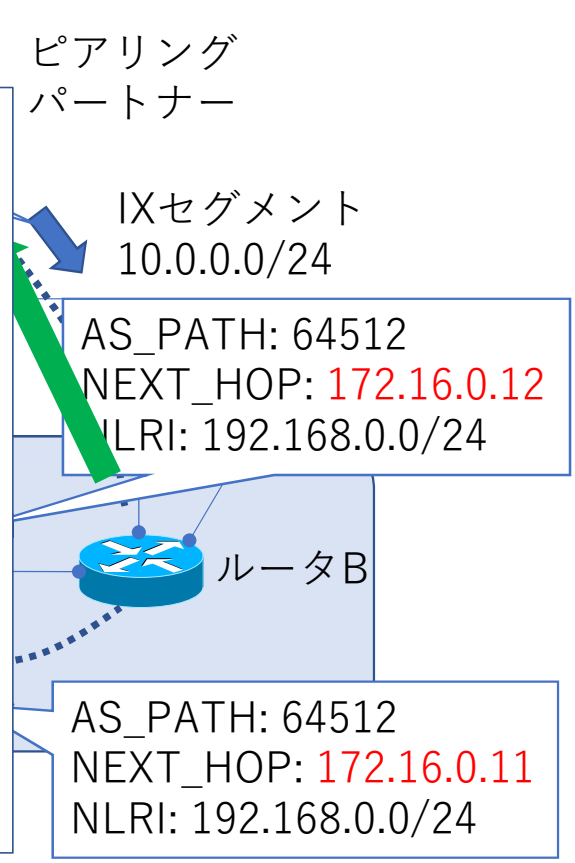
> to 10

#[Multi

> to 10

to 172.

ルータA、Bどちらも同じ動きをするので、もし、IXとの接続が2本同時に切れると、超高速に経路ループ完成!
(うれしくない)



終わり