BGPオペレータよりモノ申す
経路収束問題におけるBGP ADD-PATHの適用方法について考える

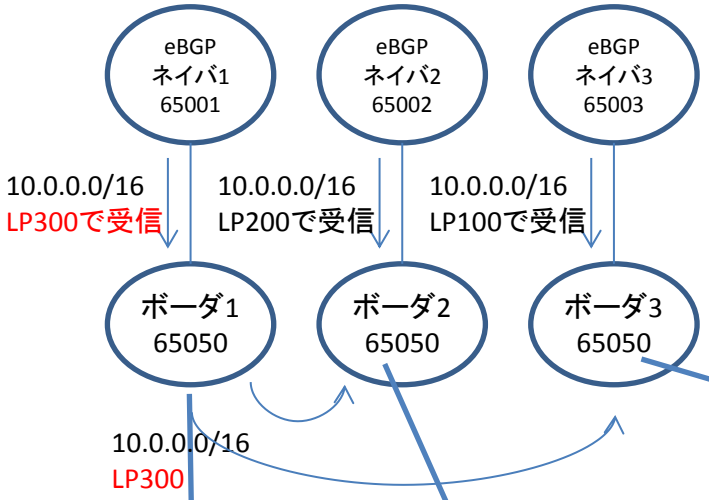
自称ネットワークエンジニア
篠宮 俊輔
shino@fornext.jp

発表の背景

- BGPの収束時間の短縮、収束中のブラックホール、過渡状態のループ回避の構成、技術は常に興味を持っています
- トランジットのBGPセッションダウンで数十秒～数分間ルーティング出来ない宛先があるのは、どうにかしたいです
- 多量の経路書き換えがあると、OSPFで用いているBFDが落ちる実装に関わっていたりしました
- 効率が良いのは好き☺

- そんなことで、(ADD-PATHではなく)BGP advertise best externalには以前からチャレンジしています。

話題とするBGPセッションダウン時の収束までの流れ
最初の状態(通常時)



AS65050での10.0.0.0/16宛てのベストパスは
ボーダ1がeBGPで受け取っているパス。

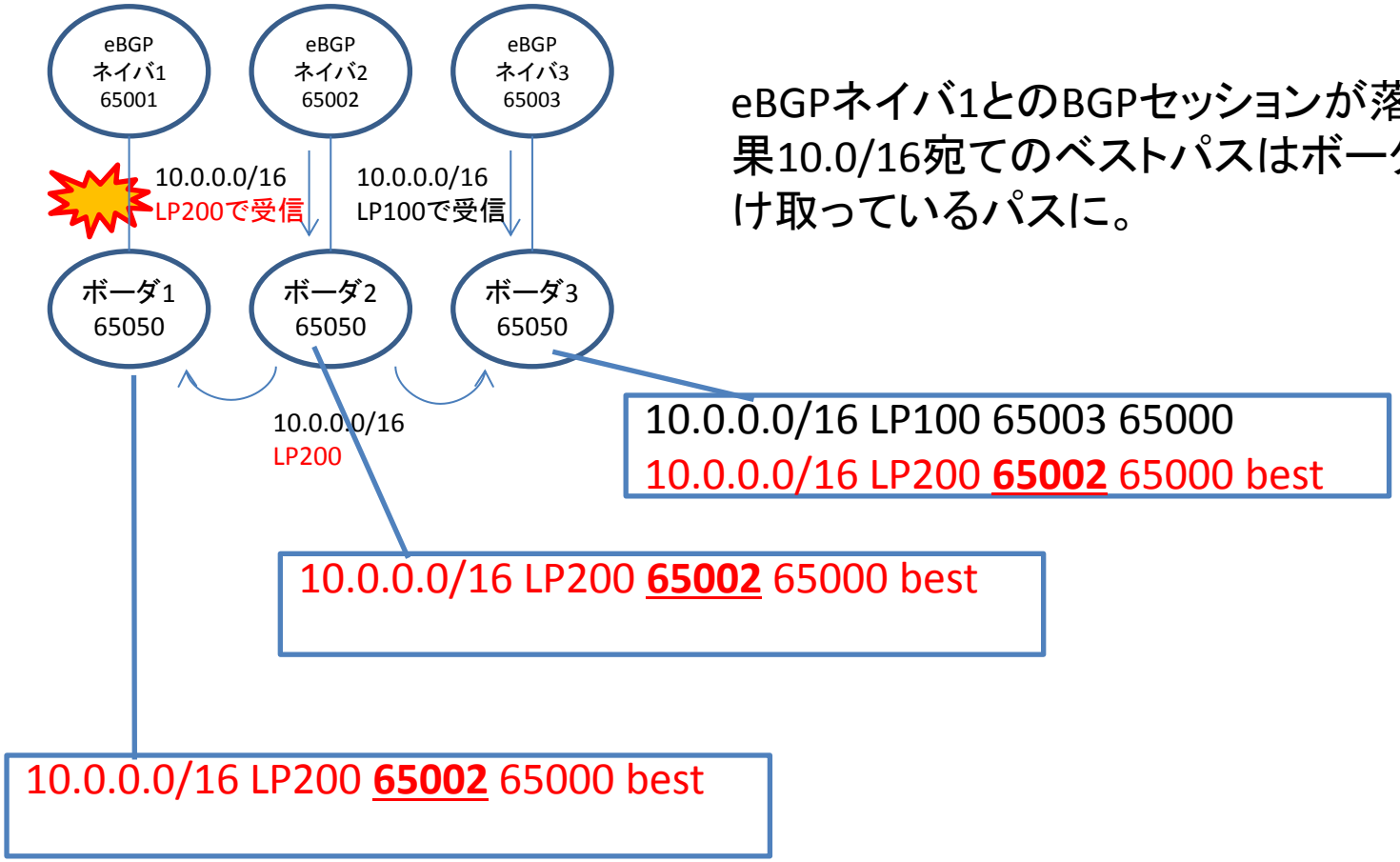
10.0.0.0/16 LP300 **65001** 65000 best
10.0.0.0/16 LP100 65003 65000

10.0.0.0/16 LP300 **65001** 65000 best
10.0.0.0/16 LP200 65002 65000

10.0.0.0/16 LP300 **65001** 65000 best

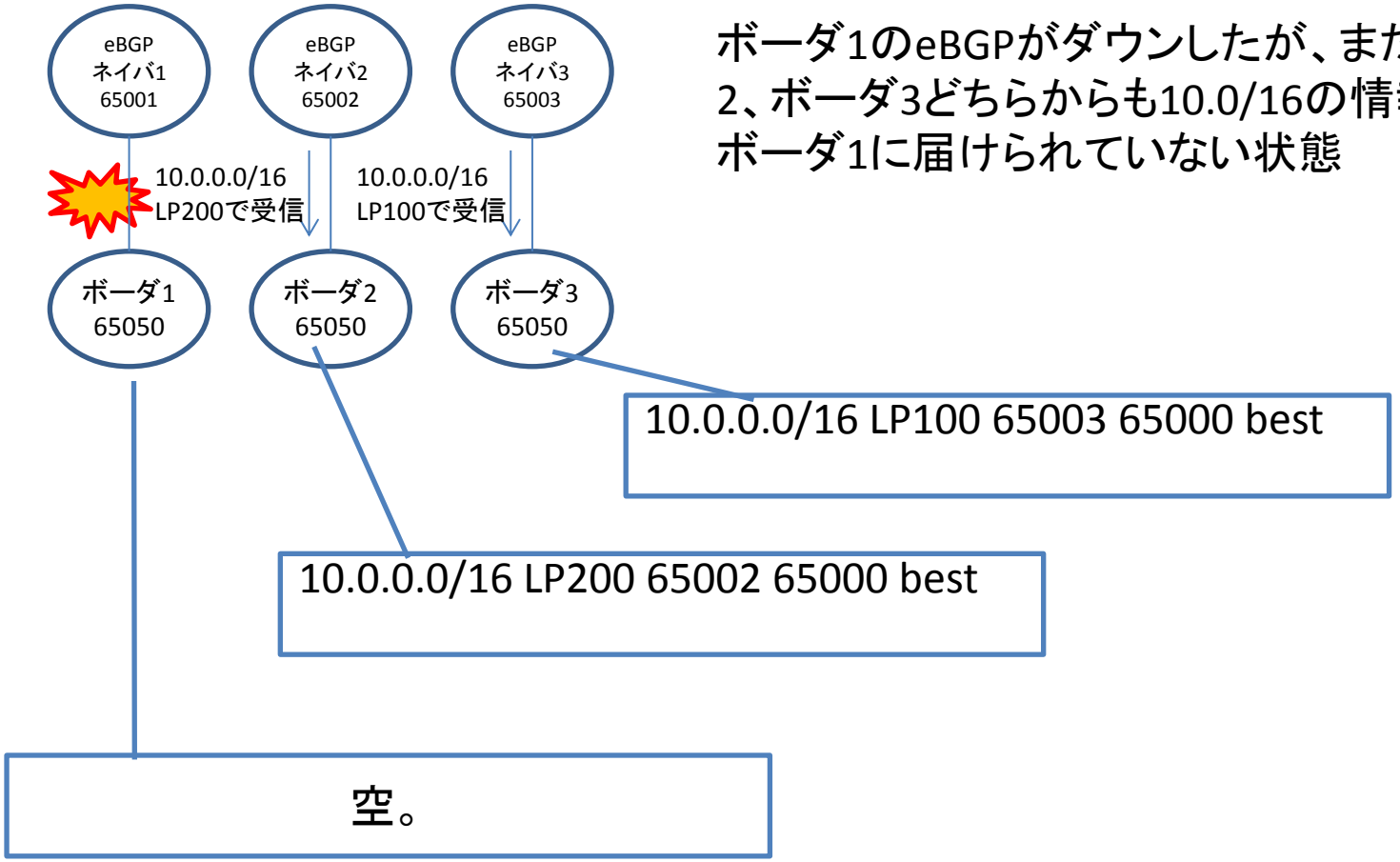
- 図のボーダ1～3間のiBGPはフルメッシュ

話題とするBGPセッションダウン時の収束までの流れ 終わりの状態(収束完了)



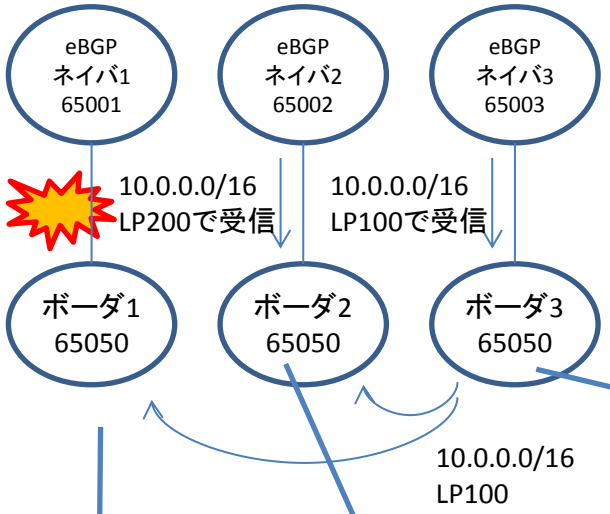
eBGPネイバ1とのBGPセッションが落ちて、結果10.0/16宛てのベストパスはボーダ2が受け取っているパスに。

話題とするBGPセッションダウン時の収束までの流れ 途中でこのような状態も(その1)



ボーダ1のeBGPがダウンしたが、まだボーダ2、ボーダ3どちらからも10.0/16の情報がボーダ1に届けられていない状態

話題とするBGPセッションダウン時の収束までの流れ 途中でこのような状態も(その2)



ボーダ1がボーダ2、ボーダ3へ10.0/16のwithdrawを送り、その結果(最終的にはベストではない)ボーダ3が受けているパスがボーダ1に届いた状態。

MRAIやルータの負荷次第で。

10.0.0.0/16 LP100 65003 65000 best

10.0.0.0/16 LP200 65002 65000 best
10.0.0.0/16 LP100 65003 65000

10.0.0.0/16 LP100 **65003** 65000 best

経路収束までにある意味無駄な状態、動作がある

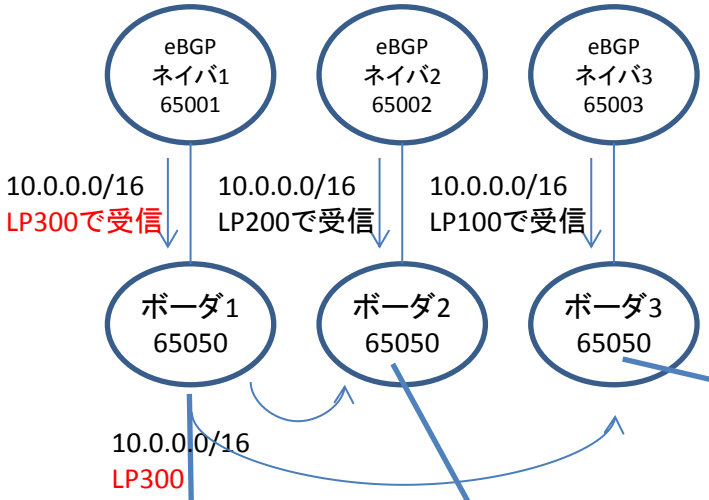
- プレフィックスに対応する有効なネクストホップがなくなるため、FIBから削除

- 最終的にはベストパスにはならないのにタイミング次第でベストパスになってしまうパス

そんなことで
BGP advertise best external

- advertise best externalはADD-PATHと同様、ベストパス以外もiBGPで他へ渡す方法です。
  - BGPへメッセージの追加などはない
  - ベストパスではなくてもbest externalであればiBGPで渡すところが通常のiBGPと違う
    - best external --- ベストパスではないが、iBGPで貰っている経路がなかった場合、ベストパスになるeBGP経路
- ベストパス以外も渡す点がADD-PATH、advertise best externalで同じ

先ほどの例でのbest external



best external --- ベストパスではないが、iBGPで貰っている経路がなかった場合、ベストパスになるeBGP経路

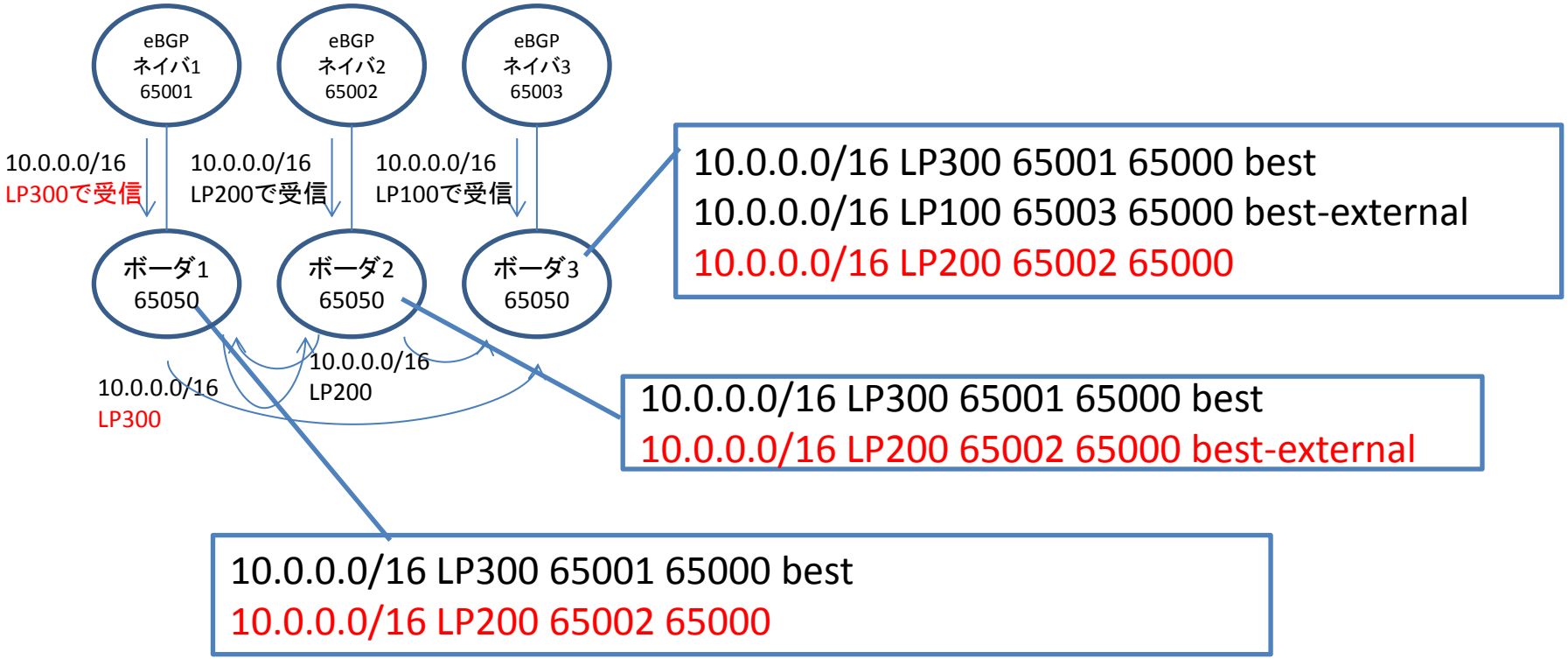
10.0.0.0/16 LP300 65001 65000 best
10.0.0.0/16 LP100 65003 65000 best-external

10.0.0.0/16 LP300 65001 65000 best
10.0.0.0/16 LP200 65002 65000 best-external

10.0.0.0/16 LP300 65001 65000 best

- 図のボーダ1～3間のiBGPはフルメッシュ

先ほどの例で、ボーダ2がbest-externalをadvertiseした状態



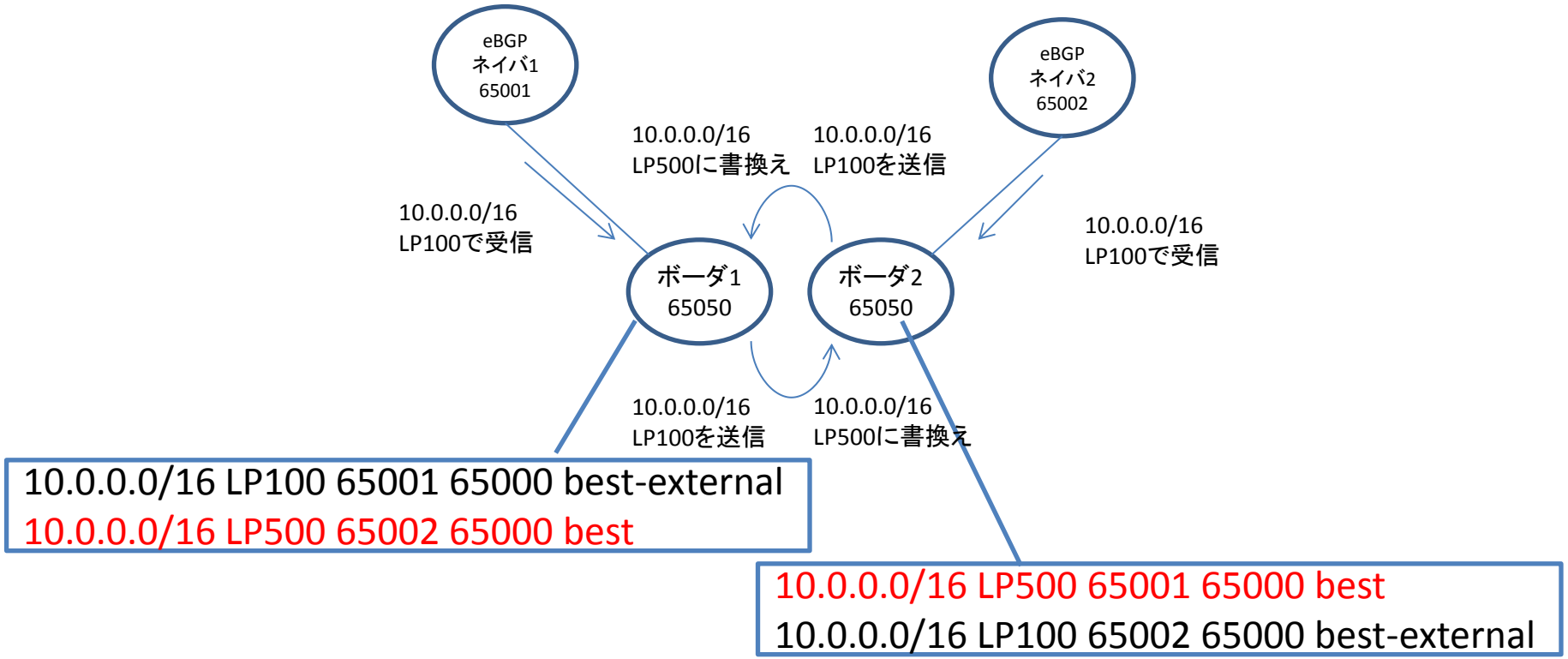
結果、eBGPネイバ1とのセッションが落ちても

- ボーダ1の10.0/16のエントリがRIBから無くなることはなく
- ボーダ3がeBGPネイバ3から受けている経路がベストだと勘違いすることもない

ベストパス以外もiBGPで渡されるがための注意

- iBGPでの(定常的な)ルーティングループが出来やすい
  - iBGPで受け取った経路の属性をいじっている場合
  - 異なる実装の複数台のルータで構成している場合のベストパス選択アルゴリズムの違い
  - 経路の存在時間でベストパスが決まると危険
    - ルータIDの比較が無効の実装は注意
  - ルータのトポロジ上の位置によってベストパスが決まると注意
    - 基本ルータ間リンクはフルメッシュか。
  - 元々怪しかったが、ベストパス以外も広告されることにより、ループ完成。

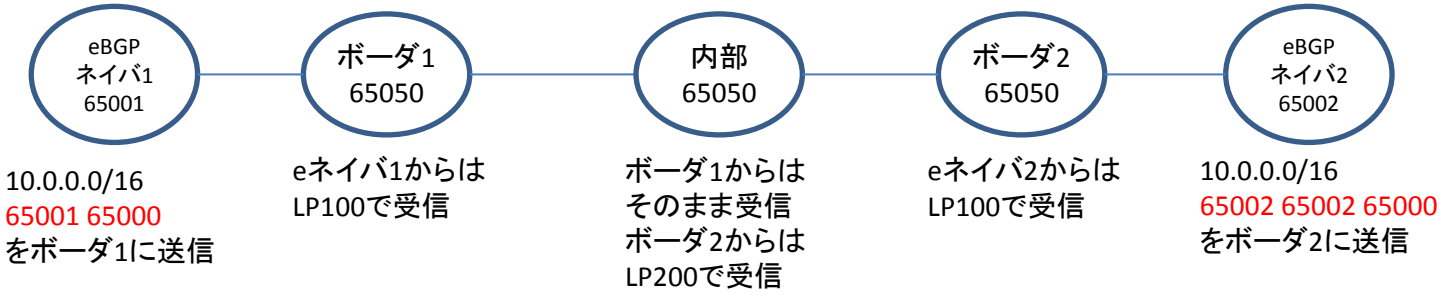
ベストパス以外を渡すがためループになる例 iBGPでのLocalPreferenceの書き換え その1



- ボーダ1は、ボーダ2からiBGPで受け取った経路のLocalPreferenceを500(大きな値)へ書き換え
- ボーダ2は、ボーダ1からiBGPで受け取った経路のLocalPreferenceを500(大きな値)へ書き換え
- 結果、ボーダ1、ボーダ2はお互いに相手のbest-externalをベストパスと選択。
- advertise best externalが有効ではなければ、ベストパスは65001経由か65002経由のどちらかに決まる。

ベストパス以外を渡すがためループになる例 iBGPでのLocalPreferenceの書き換え その2

構成、設定(iBGPはフルメッシュ。リンクは下記の様に一直線)



best external設定なし	ボーダ1	内部	ボーダ2
best	10.0/16 ネイバ1 LP100	10.0/16ボーダ1 LP100	10.0/16ボーダ1 LP100
			10.0/16ネイバ2 LP100

adv. best external設定あり	ボーダ1	内部	ボーダ2
best	10.0/16 ネイバ1 LP100	10.0/16ボーダ2 LP200	10.0/16ボーダ1 LP100
best external			10.0/16ネイバ2 LP100
		10.0/16ボーダ1 LP100	

ボーダ2ではベストではないパスが内部ではベストに!

次の話題

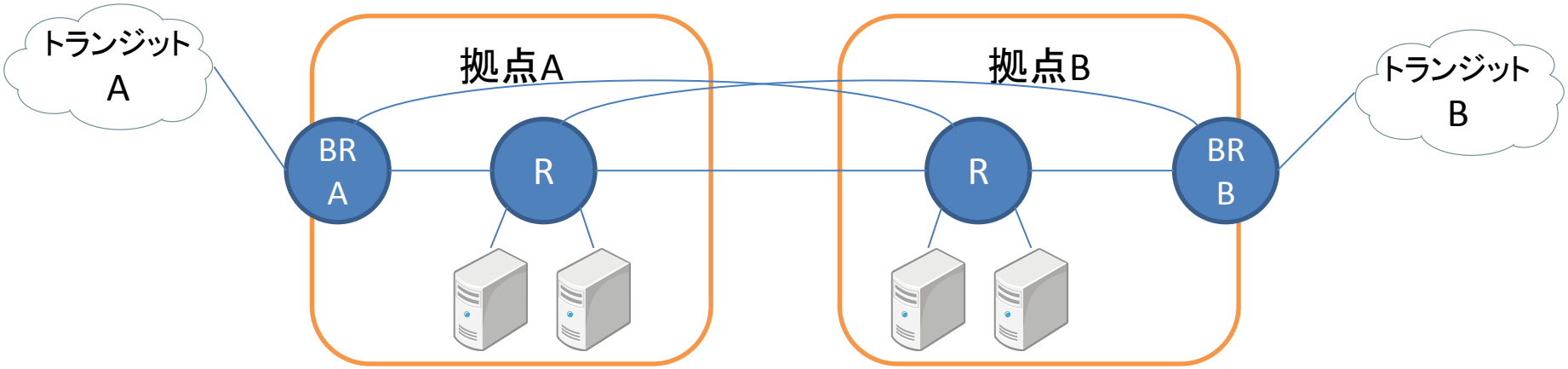
- ここまでのは、収束問題を相手にするけどADD-PATHではなくadvertise best externalをネタとしていました

- 次からは、ADD-PATHを扱うけど、収束問題ではない話題です<(_ _)>

そんなあなたにADD-PATH

- RR構成を取っている
  - でも、RRの台数は少なくしたい
- ホットポテト(closest exit)ルーティングをしている
  - ベストパスはRRが決めるので、ホットポテトしたいルータの集合毎にRRが必要

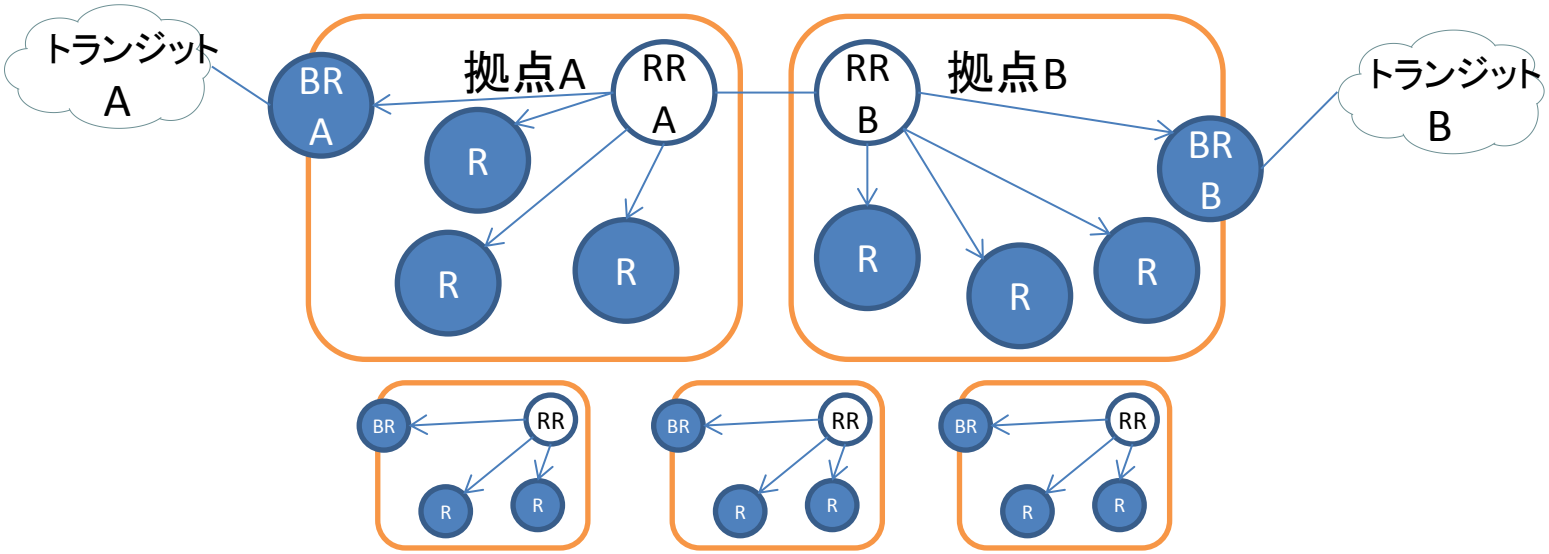
複数拠点構成でホットポテト



- 拠点Aのノードはトランジット Aから出やすく、拠点Bのノードはトランジット Bから出やすくするホットポテト構成(実装方法は話題の外)
  = 拠点Aのルータと拠点Bのルータのベストパスは異なる
- ルートリフレクタ(RR)なしでiBGPフルメッシュで良ければ、各ルータ(R)がそれぞれベストパスを決めるので実現できる。

- ボーダルータ(BR)、Rがたくさんある等の理由で、RR構成としようとすると、拠点毎にRR(群)を用意することになる

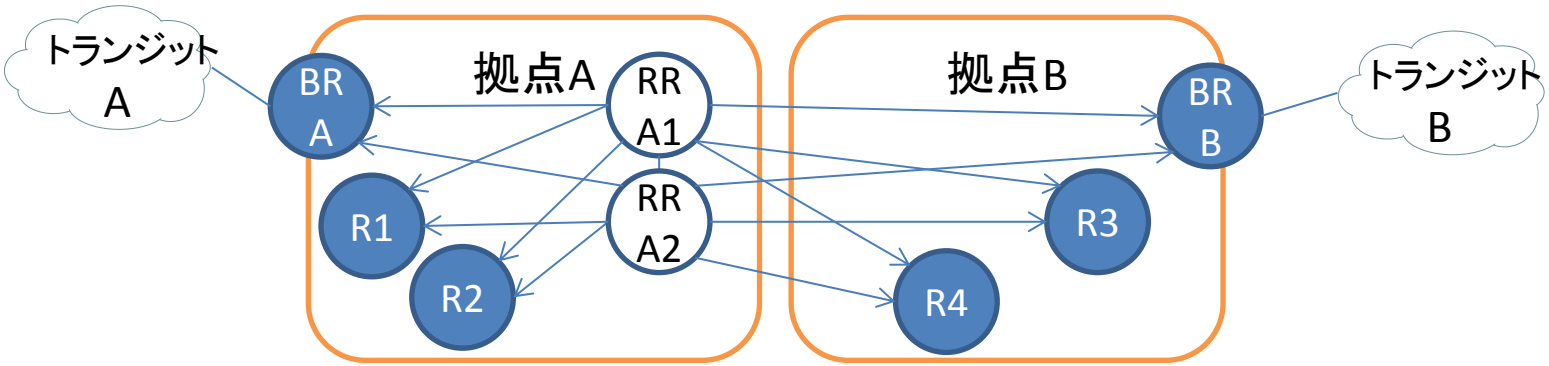
複数拠点構成でホットポテト



- RR構成とすると、拠点毎にRR(群)を置かなければならない
- 拠点の数も多かったりすると、大変。
- RRも1拠点に1台ではなく、冗長性、実装の多様性を考慮すると
  複数台 & 複数実装欲しかったりする。

- 少ない台数のRRで構成したい!!
- 台数が多い理由の一つはRRがベストパスしかRRCに渡さないから

ADD-PATHでの対応

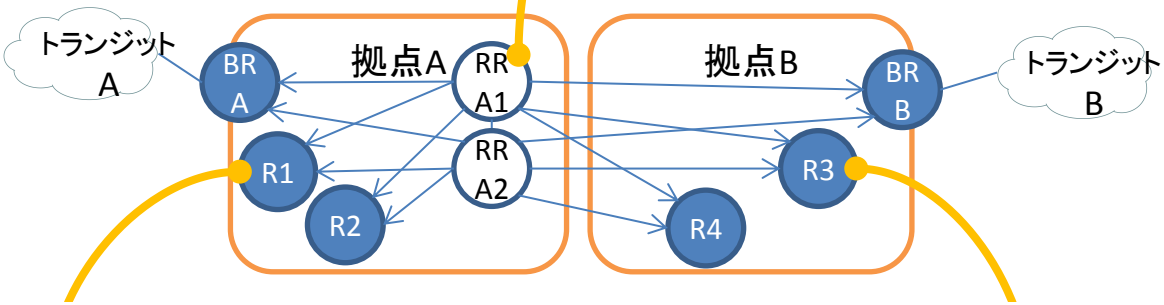


- RRを拠点Aに2台配置。
- RR A1、A2はホットポテトでトランジットA経由をベストパスとしている。
- RR A1、A2はADD-PATHで拠点A、拠点BのRRCへベストパス+もう1経路第2ベストパスを渡す。

ADD-PATHでの対応
(IGP metricでのホットポテト)

prefix	next-hop	IGP metric	status
10.0.0.0/16	BR_A	(600)	(best)
10.0.0.0/16	BR_B	(1600)	(additional-path)

RR A1、A2のRIB(これがRRCへ送られる)



RRから受けた経路情報でのR1のRIB

prefix	next-hop	IGP metric	status
10.0.0.0/16	BR_A	100	best
10.0.0.0/16	BR_B	1100	

RRから受けた経路情報でのR3のRIB

prefix	next-hop	IGP metric	status
10.0.0.0/16	BR_A	1100	
10.0.0.0/16	BR_B	100	best

もう少し工夫したい

- 良さそうであるが、全プレフィックスについてADD PATHすると渡す経路情報はその分増える。送信経路数を2にすれば経路数は倍。3にすれば3倍。
- ADDする経路を減らしたい
  - ピアから貰ったのはADDするけどトランジットからのはしない
  - 国内向けパーシャルトランジットから貰ったのはADDする
  - AS-PATH長が長いのADDしない(国外からの経路の可能性)

ADD-PATHする経路選別 (Cisco IOS)

Cisco IOS

```
router bgp 65001
 neighbor 192.168.0.2 remote-as 65001
 address-family ipv4
  bgp additional-paths select all best 3
  neighbor 192.168.0.2 activate
  neighbor 192.168.0.2 send-community
  neighbor 192.168.0.2 route-reflector-client
  neighbor 192.168.0.2 additional-paths send
  neighbor 192.168.0.2 advertise additional-paths best 3
  neighbor 192.168.0.2 route-map rmap_bgp_rrc_out out
 exit-address-family
!

route-map rmap_bgp_rrc_out permit 10
 match additional-paths advertise-set best 1
!
route-map rmap_bgp_rrc_out deny 20
 match additional-paths advertise-set best-range 2 3
 match community com_fulltransit
!
route-map rmap_bgp_rrc_out permit 100
```

- ADDする経路の選択はneighbor routemap outで設定。普段使うroute-mapと同じ。
- route-mapにmatch additional-pathsが拡張されている。
  - best 3でベスト1、ベスト2、ベスト3の経路にマッチ。
  - best-range 2 3でベスト2とベスト3にマッチ
- RRとして反射する経路の属性の書換えには制限有り。2nd以降を弱くしたい、など出来ず。

ADD-PATHする経路選別 (Juniper JUNOS)

JUNOS(Junos OS)

```
neighbor 192.168.0.2 {
family inet {
      unicast {
         add-path {
            send {
               prefix-policy add_peers-routes;
               path-count 6;
            }
         }
      }
   }
}
policy-statement add_peers-routes {
   term peers {
      from {
         route-filter 192.168.0.0/16 exact;
         route-filter 172.16.4.0/24 exact;
         route-filter 172.16.8.0/22 exact;
      }
      then accept;
   }
   then reject;
}
```

- ADDする経路の選択はadd-path send prefix-policyにポリシを指定して設定。
- ただし、今のところ、経路選択にプレフィックスしか使えない。
  - AS-PATH、communityなど、BGPらしい属性での選択が出来ない。
  - JUNOS 12.1R5で確認。

さぁ、皆さまも

- best externalやADD-PATHはまだあまり使われていないのか、バグ踏んだり、必要な機能がなかったり
  - 皆さまも試してみて、メーカへレポートしてみてください。
  - Cisco C892J(15.3(3)M)、Juniper SRX100でも動きます。
    - 1台 10万円しない程度、です。
    - SRXのIPv4、IPv6でルータの動作
      - security forwarding-options family mpls mode packet-based
      - security forwarding-options family inet6 mode packet-based