BGPのコンバージェンスとモニタリングに対するISP運用者的期待

JANOG27

NEC BIGLOBE, Ltd.

川村　聖一

コンバージェンスについて

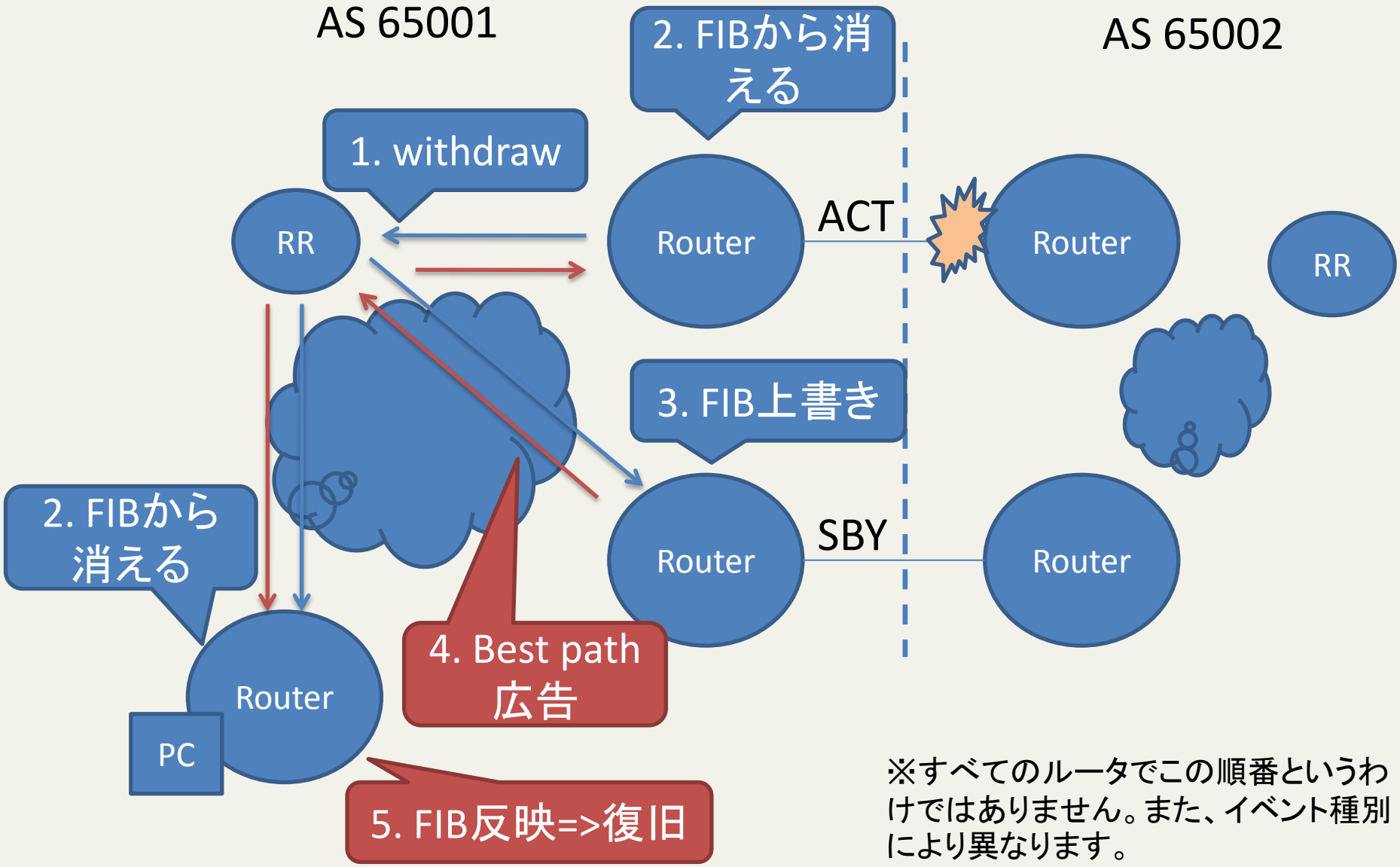
最近の動向

1. 経路が増えている(2011年1月で約330K IPv4)

2. たくさんの経路をやりとりしているPeer (Transitなど)がdownするとルータも忙しい
   - 忙しい＝CPU負荷だけが問題なわけではない
   - FIB update、iBGPとのやりとりが忙しい

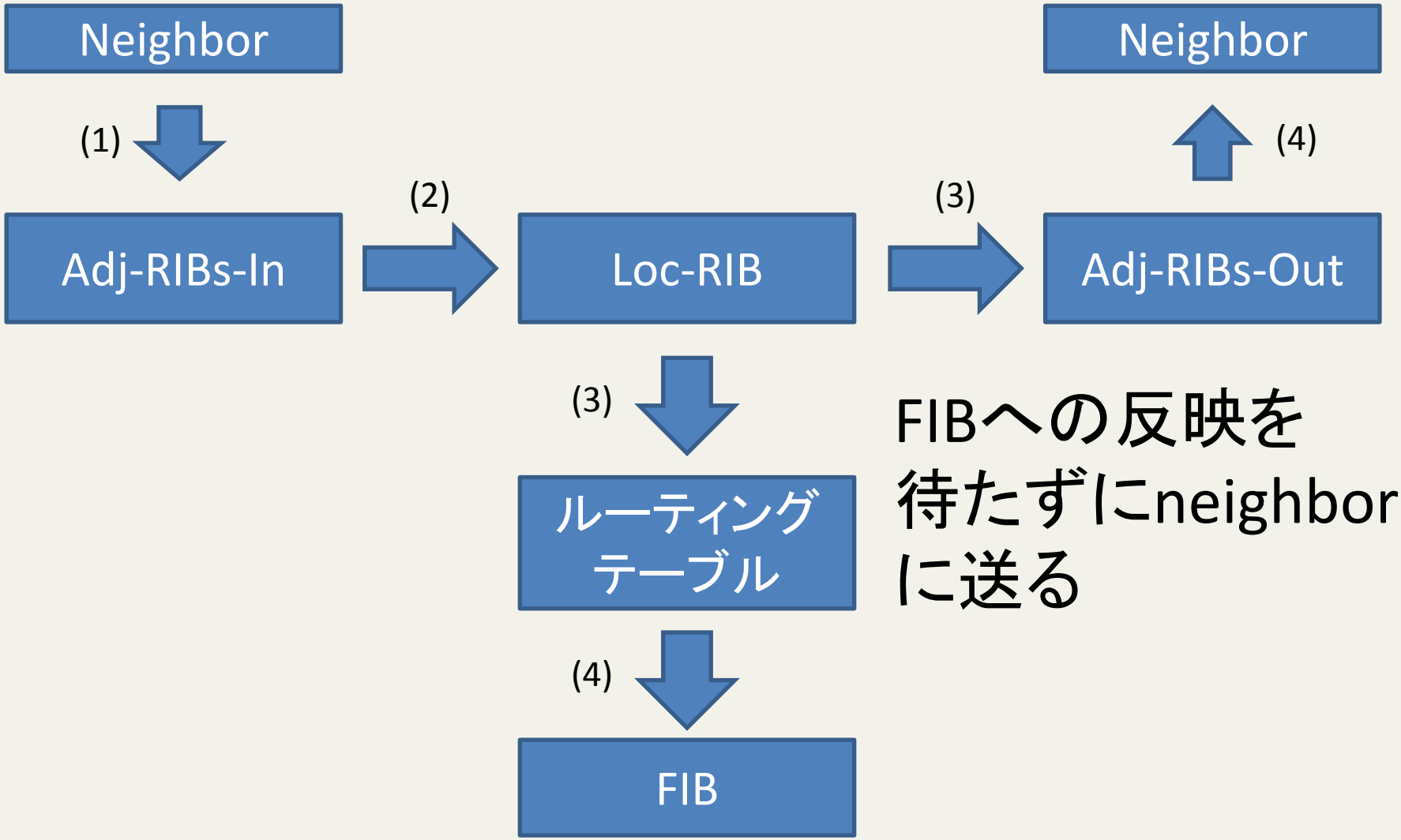
3. 実際にPeerが落ちると、どういう事がおきる？

BGP peer shut example



※すべてのルータでこの順番というわけではありません。また、イベント種別により異なります。

BGP Withdrawnメッセージ



FIBへの反映を
待たずにneighbor
に送る

出典:「ルータのFIB書き換え動作について」さくらインターネット(株)研究所　大久保さん
http://research.sakura.ad.jp/wp-content/uploads/2010/09/20100902-IRS22-LT.pdf

パケットロス

- FIBから経路が消える事により発生
- FIBから経路が消える理由は？
  - BGPでは経路が隠れる
    - Best Pathしか伝搬しない。Best Path＝Only One
    - Route Reflector構成の場合特に顕著
  - ベストパス1個しかもっていないからwithdrawされるとupdateが来るまで経路はない※
- BGPがshutされると？
  - Prefixがwithdrawされ
  - WithdrawされるとFIBから消え※
  - バックアップ経路が送られてくるとまたFIBに復活する

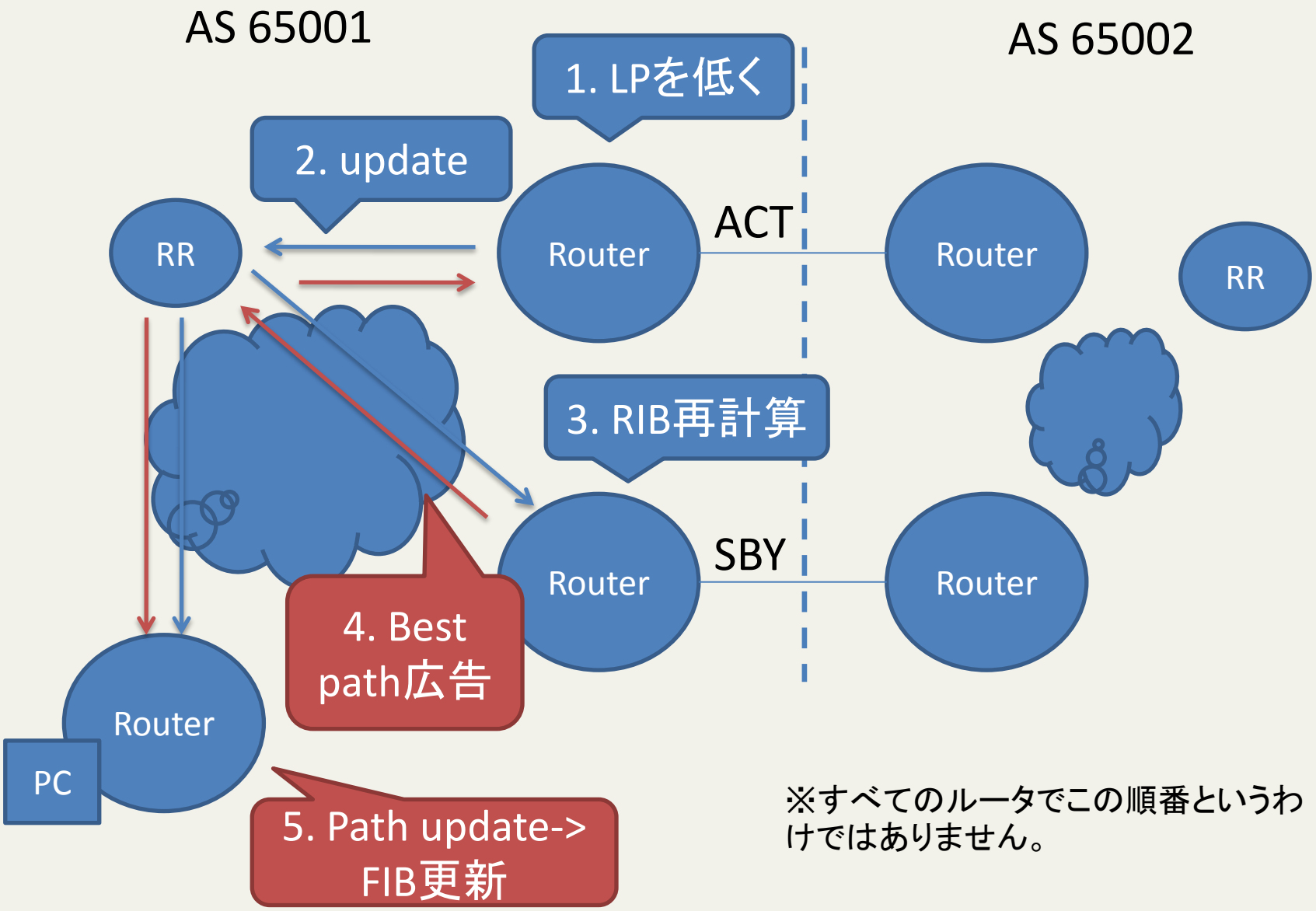
※ルータがFIBを書き換えるタイミングは実装によって異なるようです

障害 vs メンテナンス

- 障害発生時のPeer downはある意味仕方ない
  - それでもPIC/indirect next-hopでパケロスを改善したい（候補パスをふやしたい、FIBのupdateを最小限にしたい）
  - 後ほど土屋さんのところで

- メンテナンスでのPeer shutはせめてやさしく落としたい
  - Outbound: Shutする前にLocal preferenceを落としてみる
    - なるべくActiveな経路はWithdrawしない
  - Inbound: トラフィックは上手く制御できないかも
    - 最近MEDを聞かないのは一般的、Communityは横Peerでは使わないのも一般的
    - Peer groupの組み方によっては操作できない場合もある

Shut前にPreferenceを操作



※すべてのルータでこの順番というわけではありません。

Graceful shutdown

- IETFのGROW WGでDraftとして議論されている運用方法
  - http://tools.ietf.org/html/draft-ietf-grow-bgp-gshut-02
- やさしいBGP shutdownのやり方について
- 特質すべきは、BGPをshutするメンテナンスが発生する場合、ルータが相手側に「おとすよー」と通知することによって相手側ルータで迂回に必要なアクションが取れる事を想定
  - 「おとすよー」伝達手段1：電話
  - 「おとすよー」伝達手段2：Communityを付けて送ると、相手側がLPを下げてくれる
  - 相手側がCommunityを適切に処理してくれる事が前提

Question

- 「おとすよー」伝達プロトコル、ほしい？
  - Graceful shutdownのCommunityがWell knownで定義された場合、送信側としてはこれを使いたいですか？
  - MEDを操作される事が運用に影響を与える可能性があるような環境の場合でも、受信側としてgraceful shutdownは許容できる？
- shutdownをかける側として、(graceful shutdownに限らず）機器にはどういう実装を求める？

モニタリングについて

BGPはモニタリング機能が弱い

- 制御レイヤ(BGPセッションそのもの)
  - syslogやエンタープライズMIBで取れる情報はそこそこあるけど、up/down以外はあまりとらない
    - 機器によってバラバラ
    - そもそもsyslogやMIBが有効活用できるかどうかは受け側のシステムに依存する
  - 各Peerが送ってくるCapability、Holdtimerなどの情報履歴はあまり管理されてない
    - Graceful shutdown本当にやるなら、把握しておかないとね
  - できないからしてない？ やりにくいからしてない？

BGPはモニタリング機能が弱い

- データレイヤ(経路)
  - 相手のPeerからいつ、何時に、どれだけの経路を流されているか管理できていますか？
  - 受信経路(Adj-RIBs-In)数と、フィルタ後の経路(Loc-RIB)数は管理できていますか？
  - 流れている経路そのものの記録(dump)はとれていますか？

BGPモニタリングへの期待

- 動向把握
  - 相手の送ってくるCapabilityやTimerは意味がある
    - 4octet対応度合い、どれだけ品質を気にしているかどうか
  - Peer毎の受信経路の増加（増えてるPeer、減ってるPeer）
- 障害防止
  - Adj-RIBs-InとLoc-RIBに必要以上の差分があるか確認
    - ポリシーは適切かチェック
- 障害対応
  - x時間前の経路を調査したい、と思った時に対応できる
- 証拠
  - 経路Hijackされた時に確証が残る

手段

- MRT
  - http://tools.ietf.org/html/draft-ietf-grow-mrt-13
  - Loc-RIBをファイルにexportするフォーマット
  - quaggaで動く
    - dump bgp routes-mrt /home/mucho/route/dump%Y.%m.%d.%H%M 60m
  - 出来上がったbinファイルを読むためのツール
    - libbgpdump
- BMP
  - http://tools.ietf.org/html/draft-ietf-grow-bmp-05
  - Adj-RIBs-Inをモニターして、BMPコレクターに送る
    - 受け取った経路のDump
    - Peer Downの理由を記録
    - 統計情報の記録
      - フィルタにひっかかった経路の数など
    - Peer Upとその付属情報

現状

- MRT
  - 商用ルータでは実装されてない
- BMP
  - まだコレクター(monitoring station)がなさそう
  - 基本機能は最小限。拡張に期待
  - “Many researchers wish to have access to the contents of routers' BGP RIBs as well as a view of protocol updates that the router is receiving”
  - 研究者向け・・・？

Question

- MRT、BMPにどういう機能を期待しますか？
- 実装に求める事は？
- Peer毎の統計で管理したい情報は何？