高速な障害復旧に必要な思いやり

ソフトバンクテレコム
松嶋　聡

Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

背景::かれこれ３年前

- 新しいネットワークをつくりなさい
  - とあるアプリケーション用です
  - 高速な障害復旧，可能ならば sub-second で
  - メトロエリアのネットワークを紡いで、全国をカバーしなさい
    - 数千のルータ
    - 数百のメトロエリア
- 当時の前提条件
  - L1, L2 の故障検出メカニズムは使えません
    - え？ broadcast media なんすか？
  - 経路制御プロトコルは ＢＧＰ のみです
    - なんで？
  - ＢＦＤ有史以前です
    - あっても、「ＢＦＤ？　そんな高い箱使えません！」
      - あぁ、要は安くあげろと…

Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

ネットワーク設計してみる（１）

Physical view　　**Logical view**

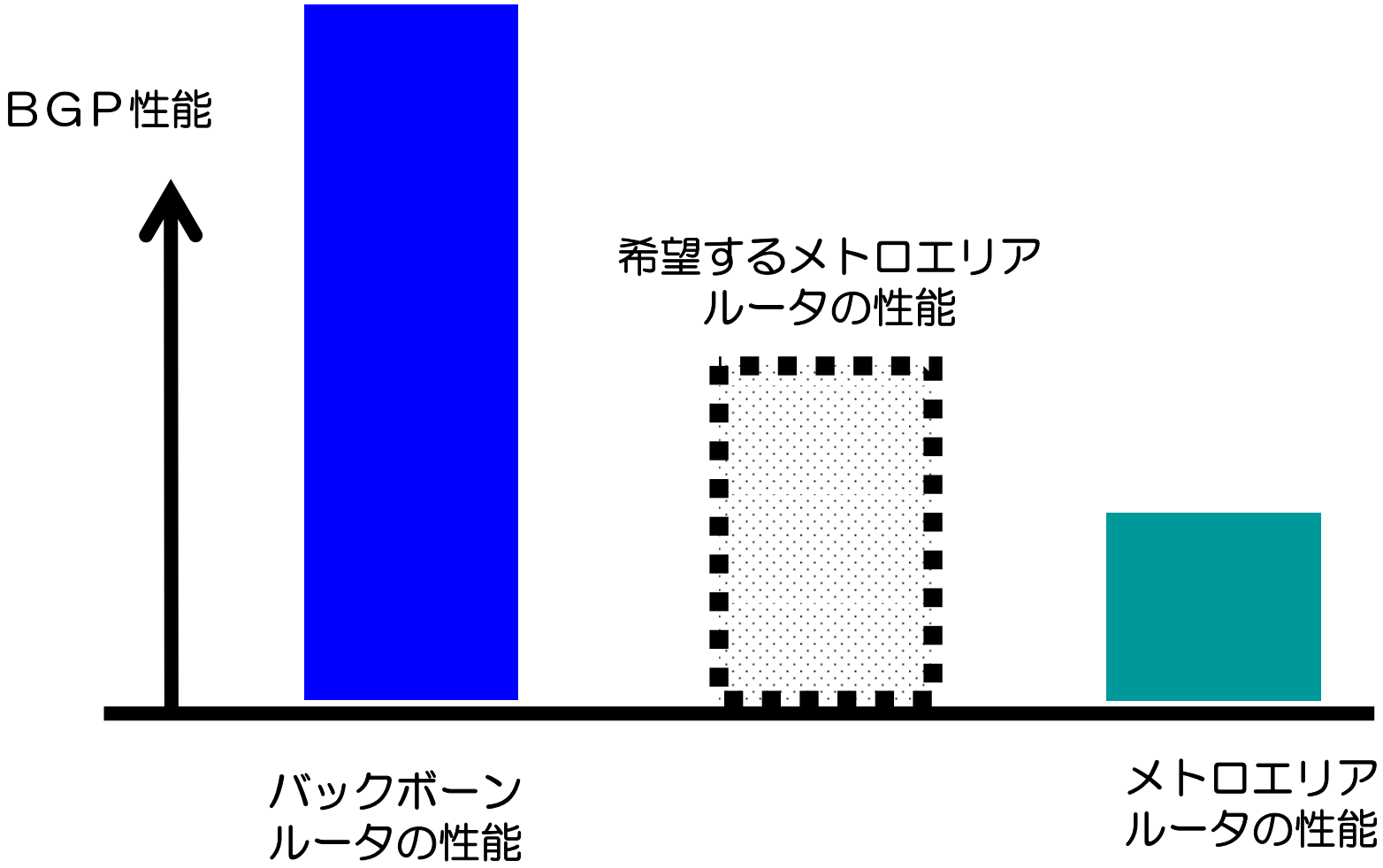
Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

ネットワーク設計してみる（２）

- 設計パラメータ::ＢＧＰ
  - 1バックボーンルータあたり、数百 ピア必要
  - ネットワーク全体で数万経路が必要
  - 経路（遅延）最適化のため、経路集約効率が限定される
- 設計パラメータ::ルータ能力
  - バックボーンルータのスペックはそれなり
  - メトロエリアルータは、しょぼしょぼ

Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

ネットワーク設計してみる（３）



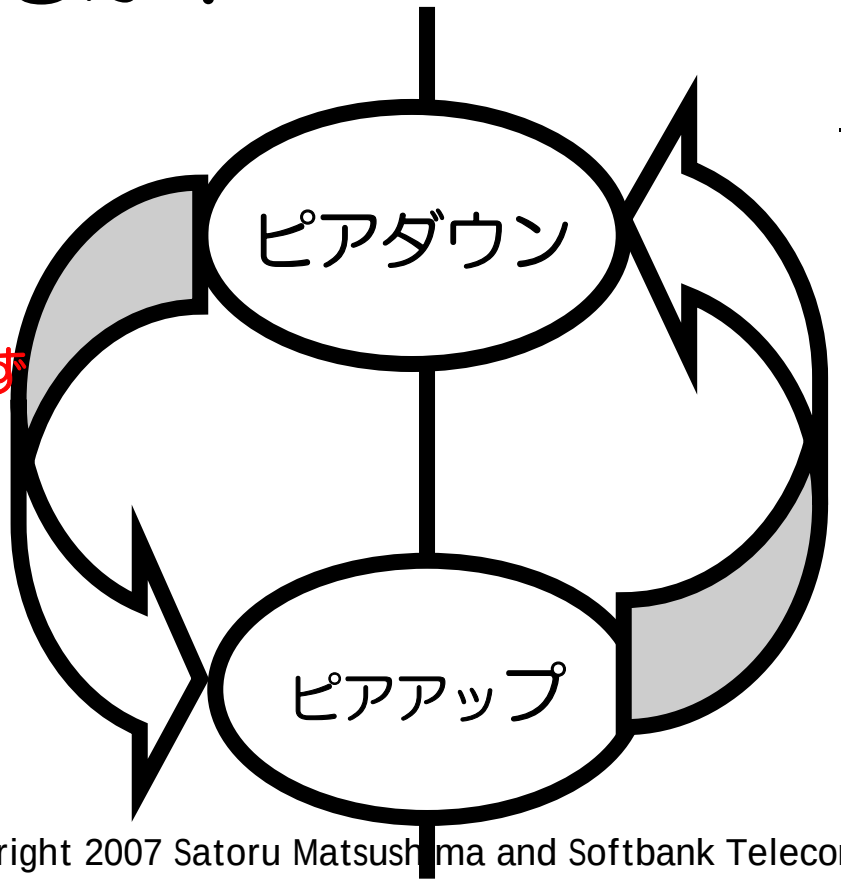
Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

ネットワーク設計してみる（４）

- 極端な性能差のＢＧＰピアが暴れだしたらどうなるか？

小さいルータ

- **CPU**パンパン
- 経路受け切れない
- ピアアップしきれずまた落ちる



大きいルータ

- ピアアップで全速アップデート
- 落ちた経路を都度アップデート

Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

こんなマシン・機能があったら...（１）

- ＢＧＰ性能の差をカバーするやさしい箱
  - keep-alive/hold-timer は default で、ピアの up/down 負荷をかけずに経路迂回したい
- でも、そんなの作ってくれないよねぇ
  - C3oさーん！
  - J5rさーん！

Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

こんなマシン・機能があったら…（２）

- ＢＦＤでなくていいから、ソコソコ速い障害検知メカニズム
  - マルチポイントメディアでも動作して
  - 相手ルータがサポートしていなくても使える一般的なプロトコルでプローブ

Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

こんな機能・マシンがあったら…（３）

- プローブと経路制御の連携どうしよう…
  - 検出したら telnet/snmp で config 変更？
  - 誰が？　人間？　NMS？　ひえー
- そうだ！ルートリフレクタ使おう！
  - ＲＲなら、ルータでがんばらなくても、ネットワークに外付けできるぢゃん！
  - 経路制御に直接作用させられるからプローブと連携させやすいかも！

Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

林さん！
よろしくお願いします

お願い１：普通のプロトコルでプローブ

お願い２：ルータに負荷をかけない高速障害復旧

Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

高速な障害復旧に必要な思いやり

～運用編～

Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

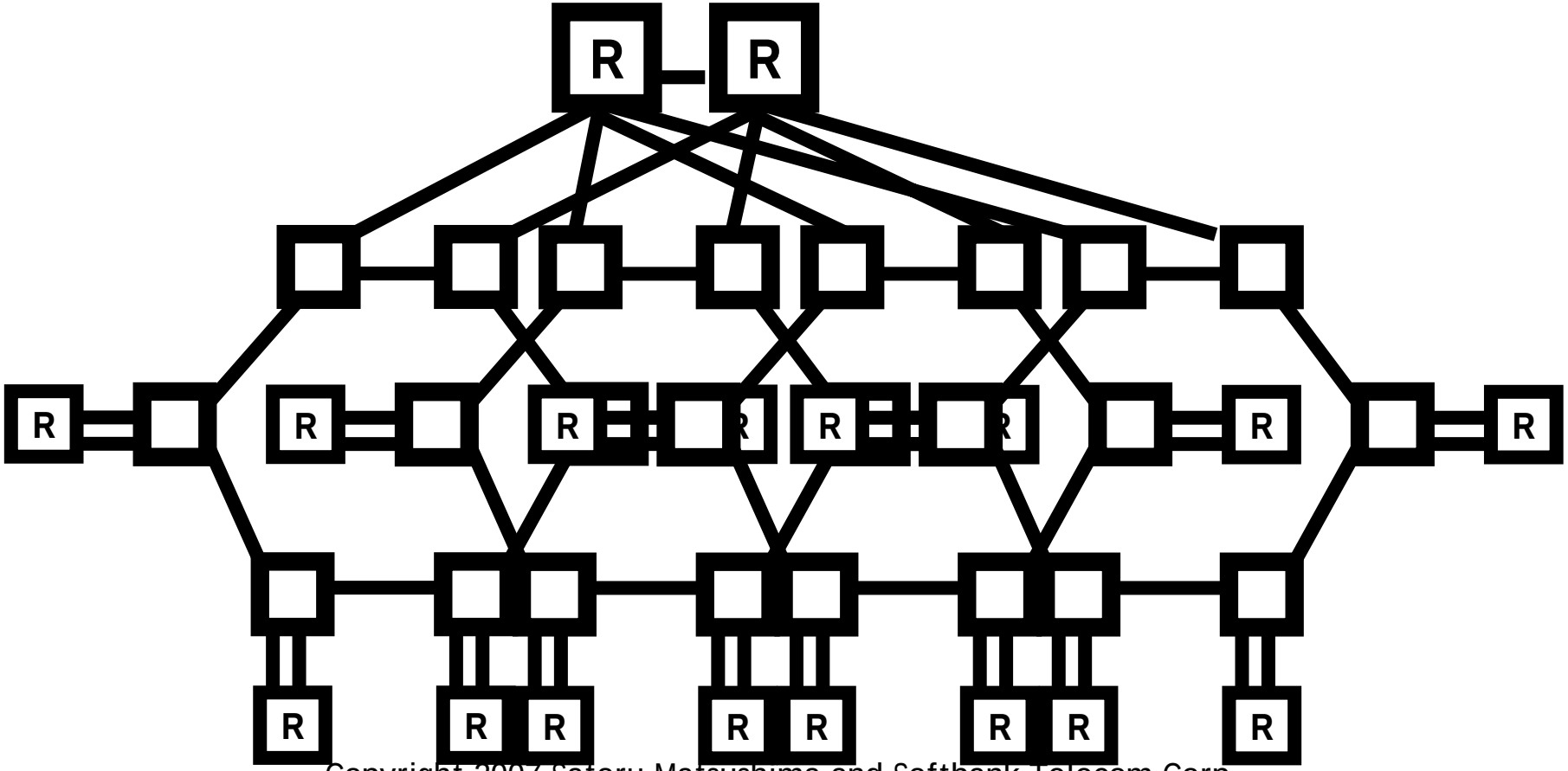
思いやりのある運用を目指そう！

- 誰に対して？
  - はじっこにいる、少し小さいルータたち
  - 数千～数万経路のピアを短時間で上げ下げするのは、かなりつらい
- 小さいルータに負荷をかけない
  - でも、障害検出と復旧はすばやく！
- ＲＲでどこまでできるか？
  - keepalive/hold-timer は default のままで（3分!）
  - プローブの間隔と保護時間の調整をやり
  - ピアをあげたままで、ＢＧＰメッセージ数を最小限にキープし
  - 高速な故障検出と復旧（迂回）を実現する！

Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

実際に構築してみよう

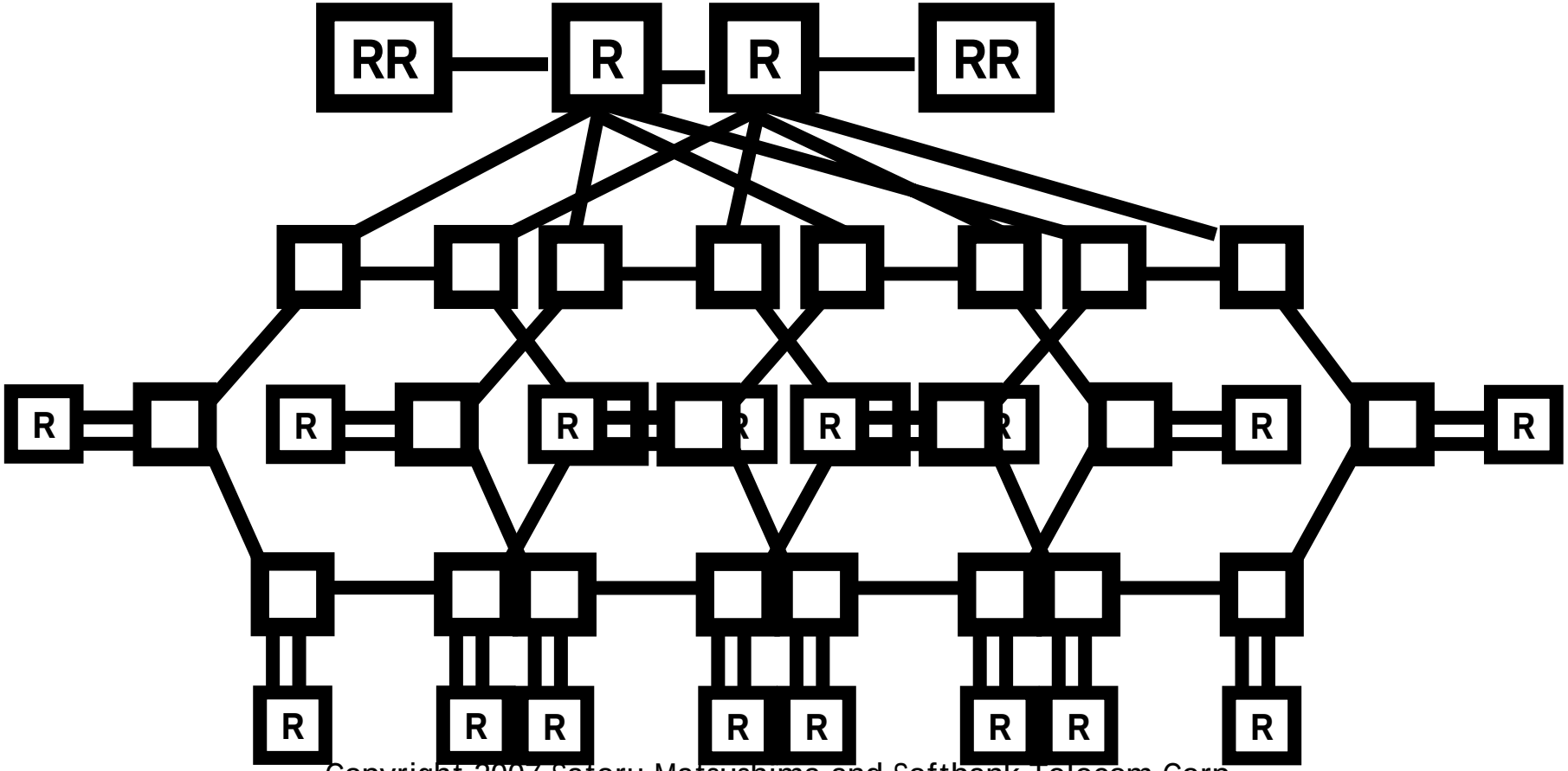
- メトロエリア集約ネットワークとＲＲ



Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

実際に構築してみよう

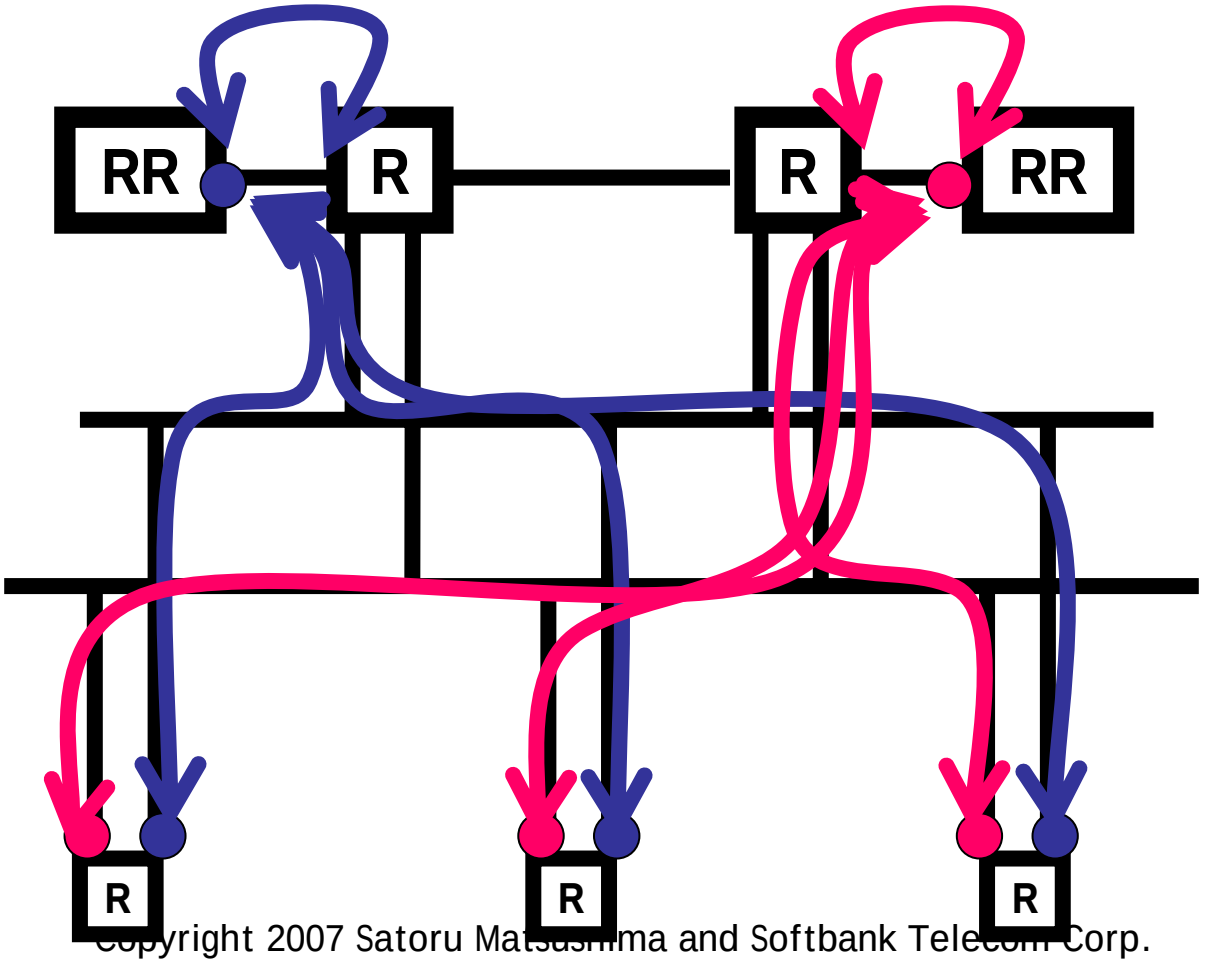
- メトロエリア集約ネットワークとＲＲ



Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

実際に構築してみよう

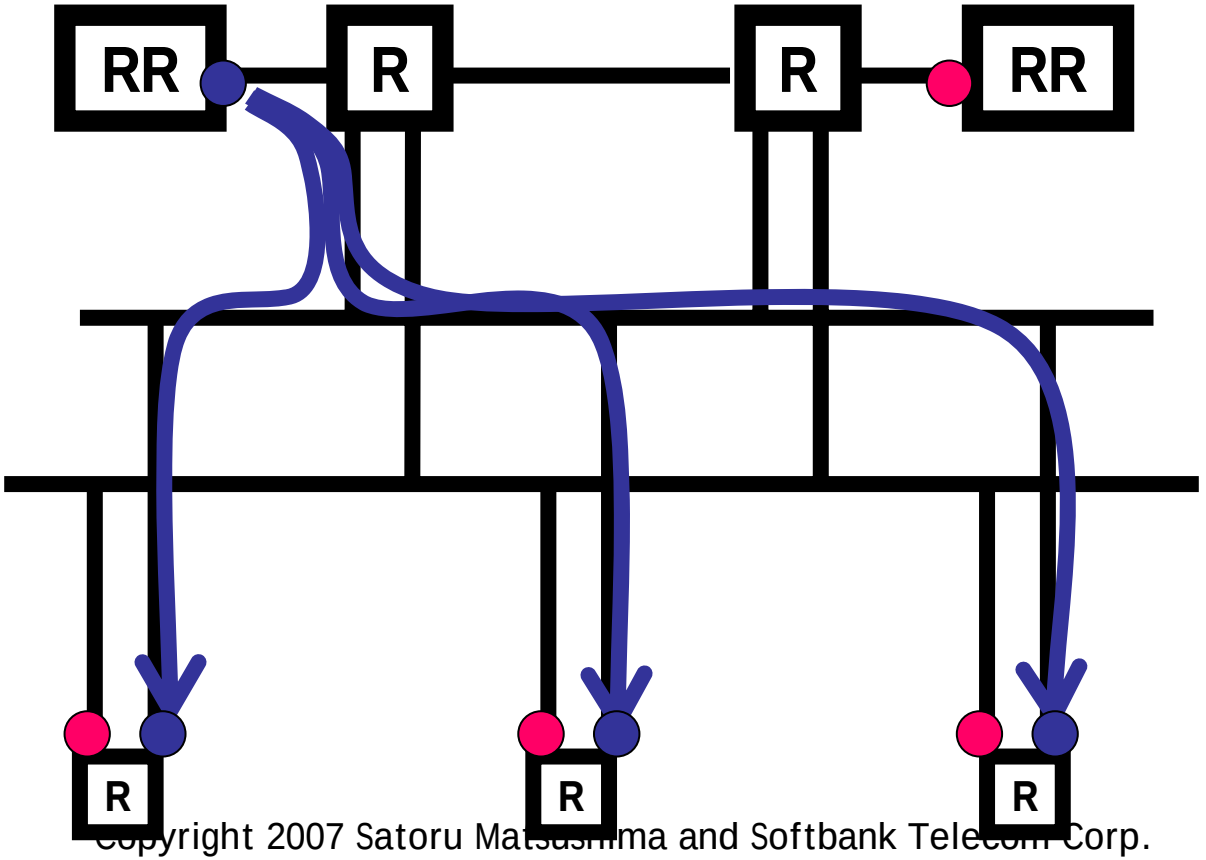
- ＢＧＰピアの設定



Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

実際に構築してみよう

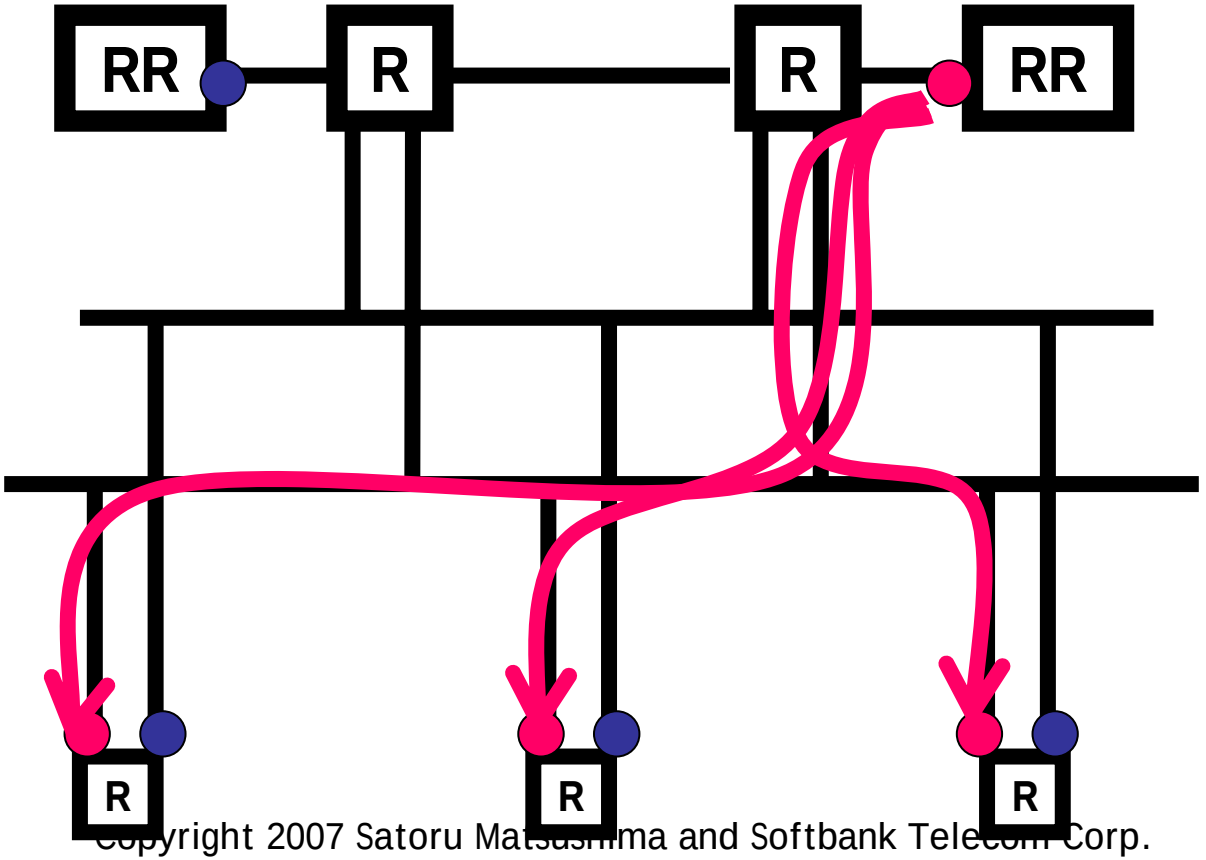
- プローブのかけ方



Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

実際に構築してみよう

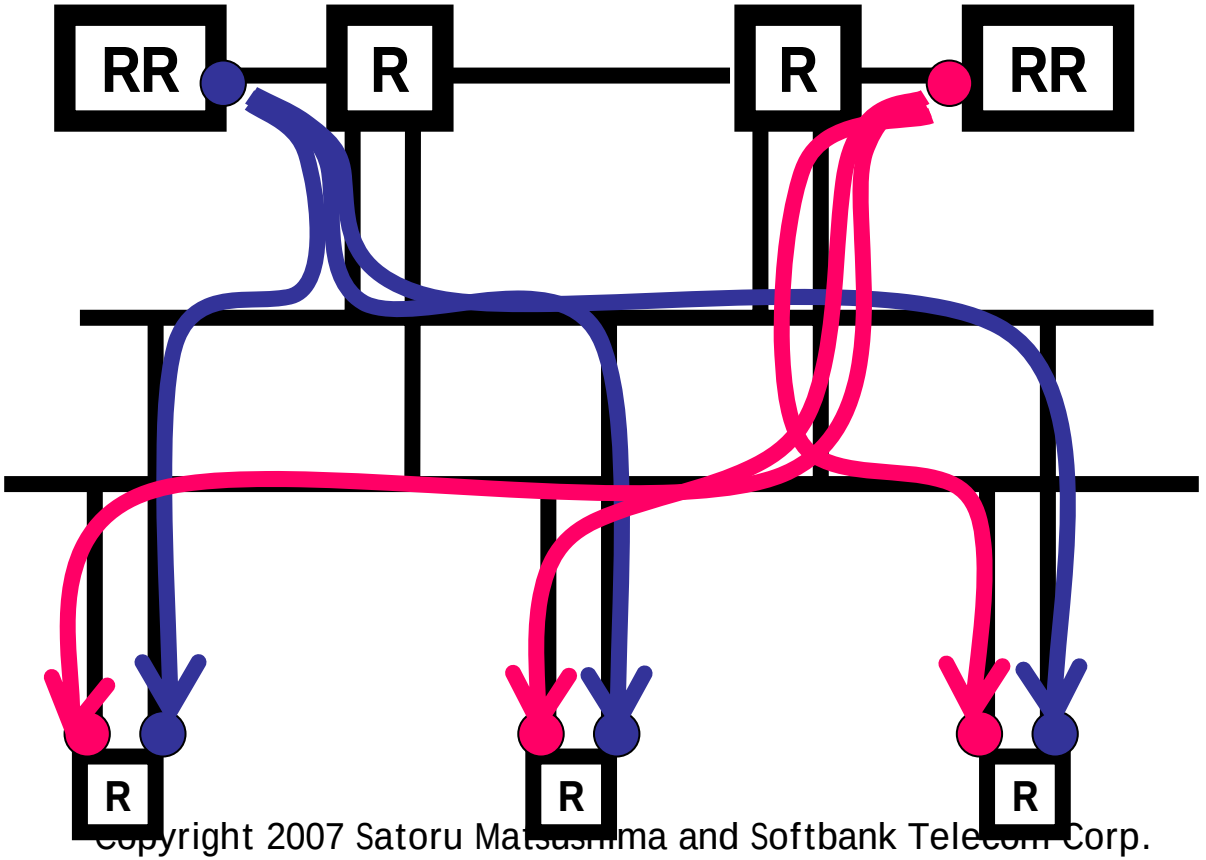
- プローブのかけ方



Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

実際に構築してみよう

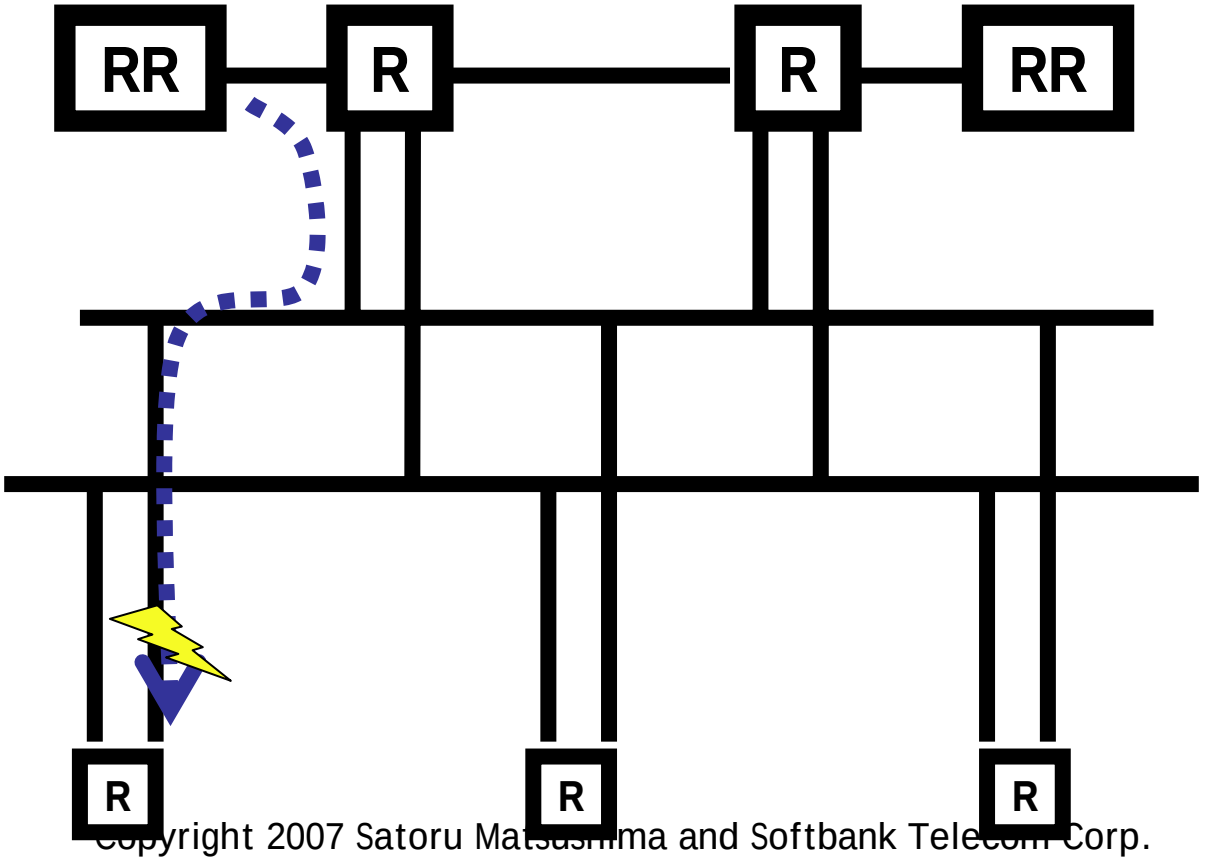
- プローブのかけ方（ＢＧＰピアと同じ）



Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

検証::障害検知時の動作

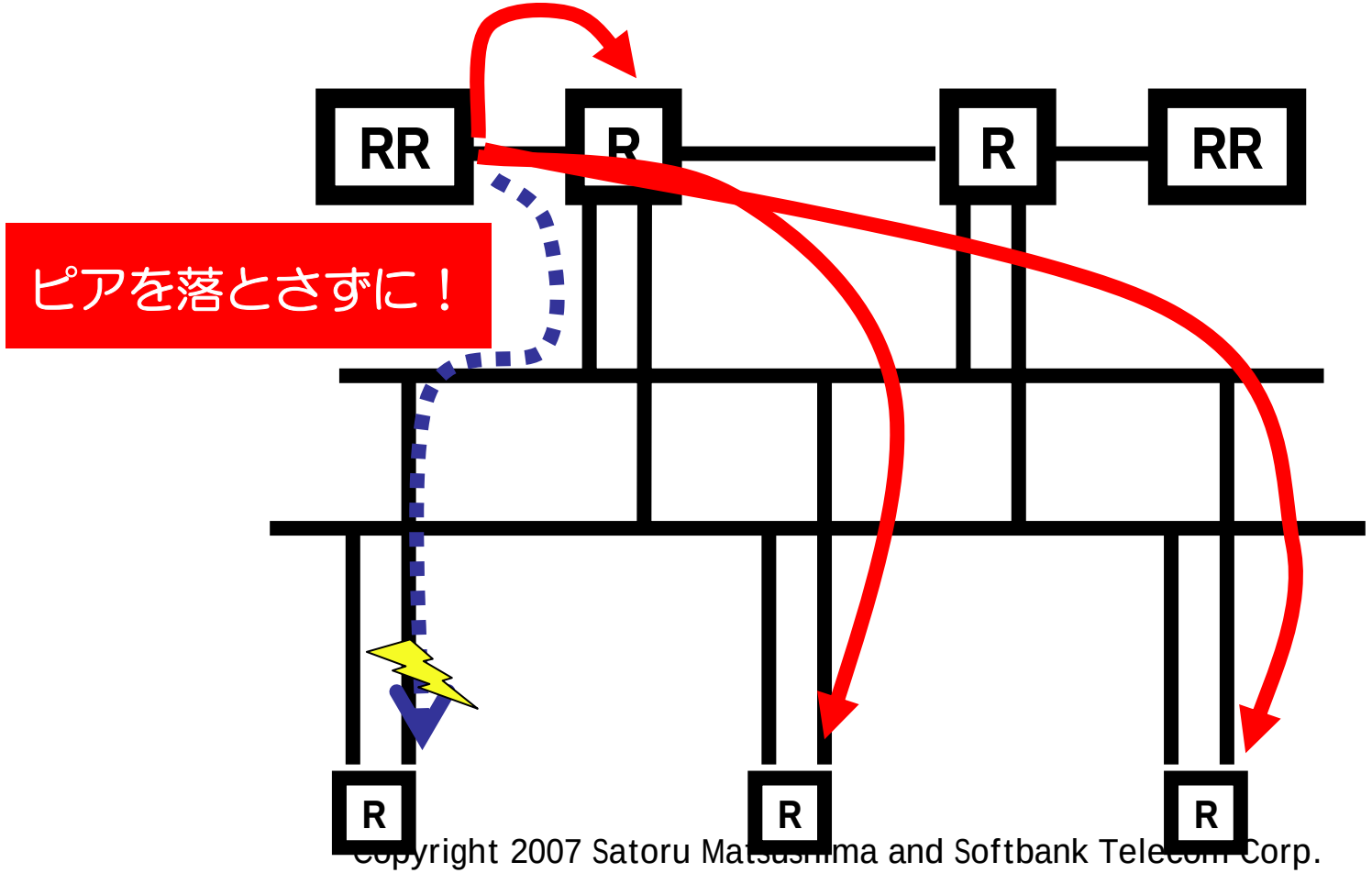
- プローブが通らなくなる



Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

検証::障害検知時の動作

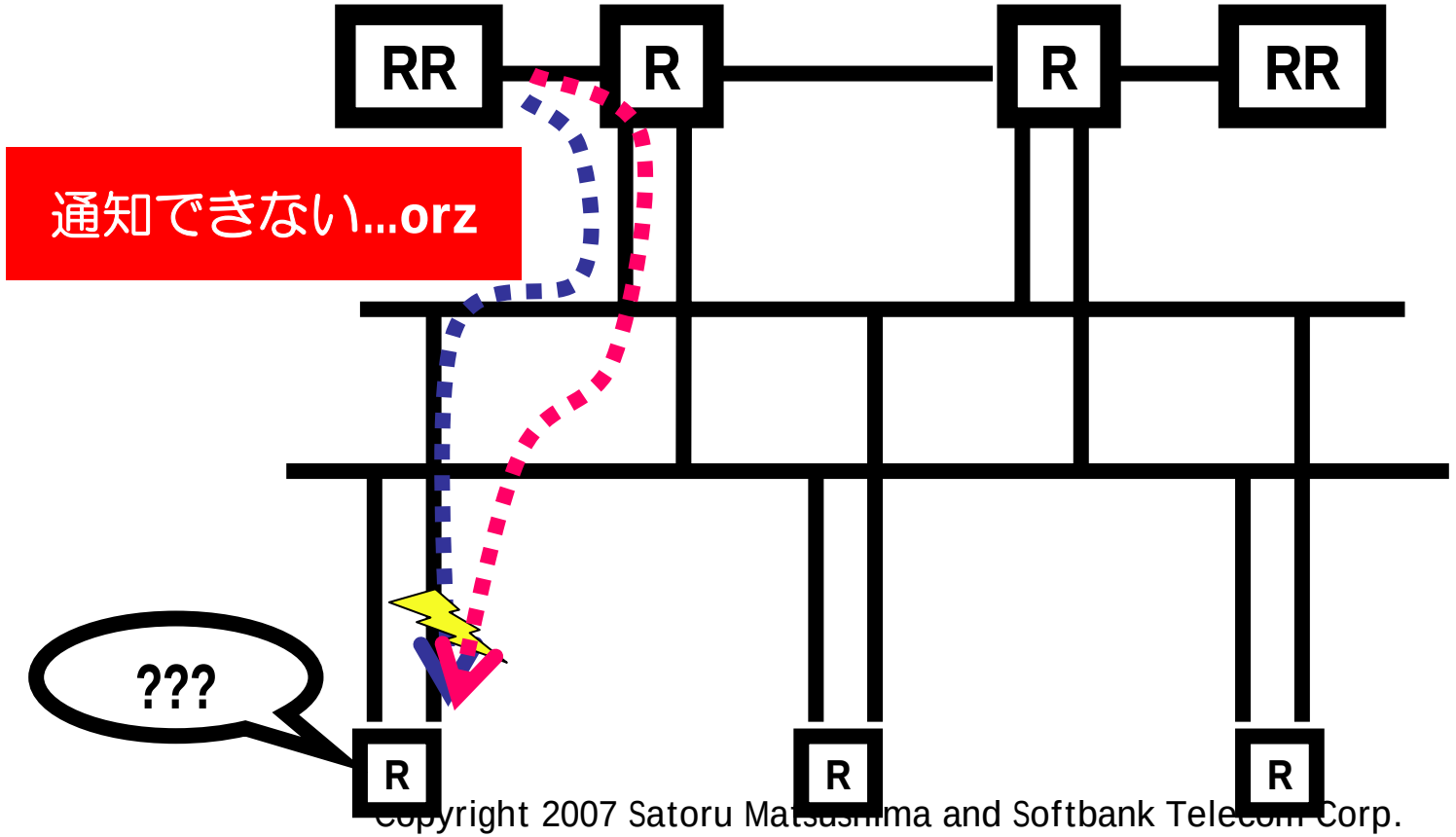
- すぐさま他のルータに通知（withdraw）



Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

検証::障害検知時の動作

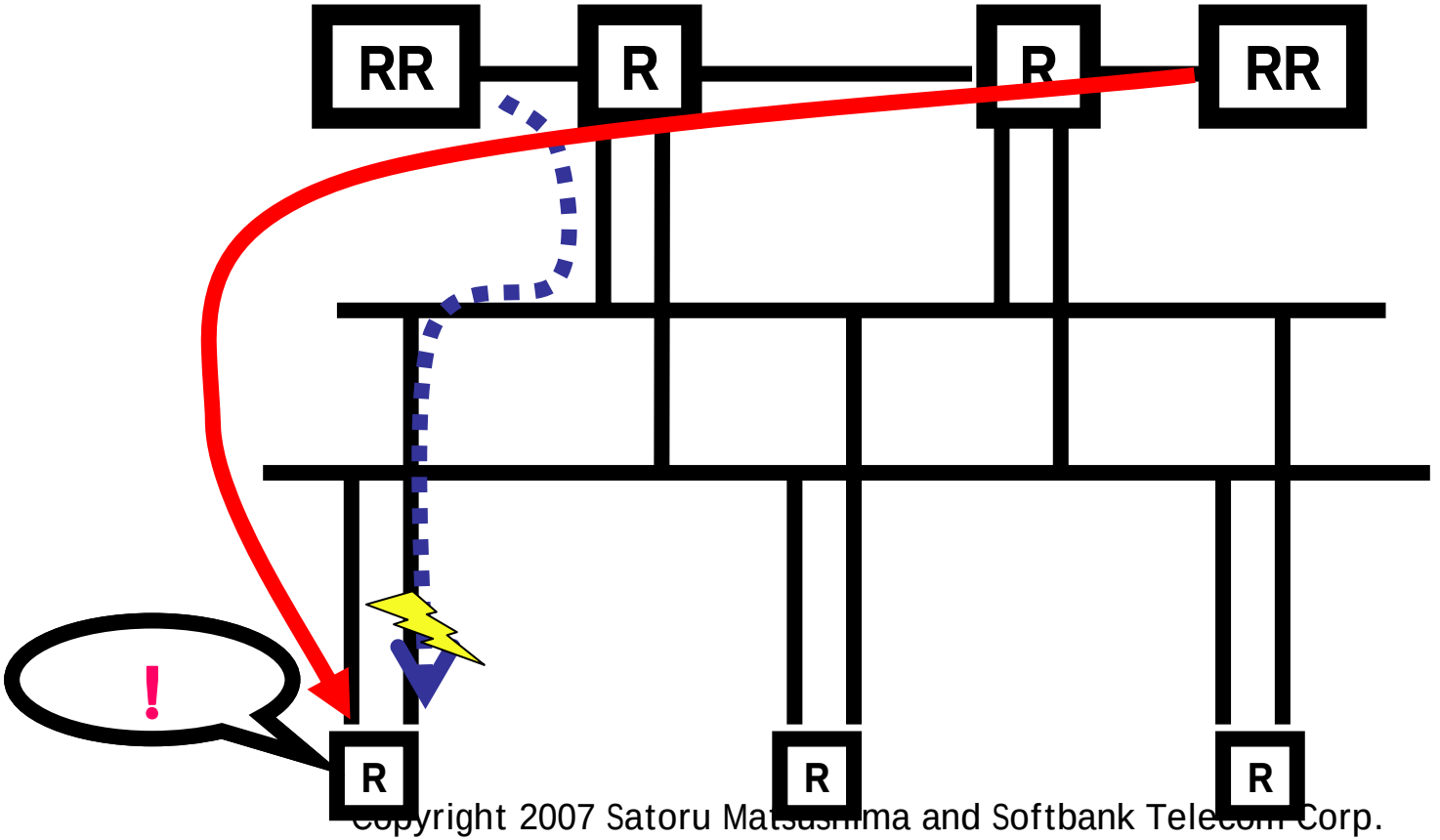
- 対象ルータが障害検知できないケース



Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

検証::障害検知時の動作

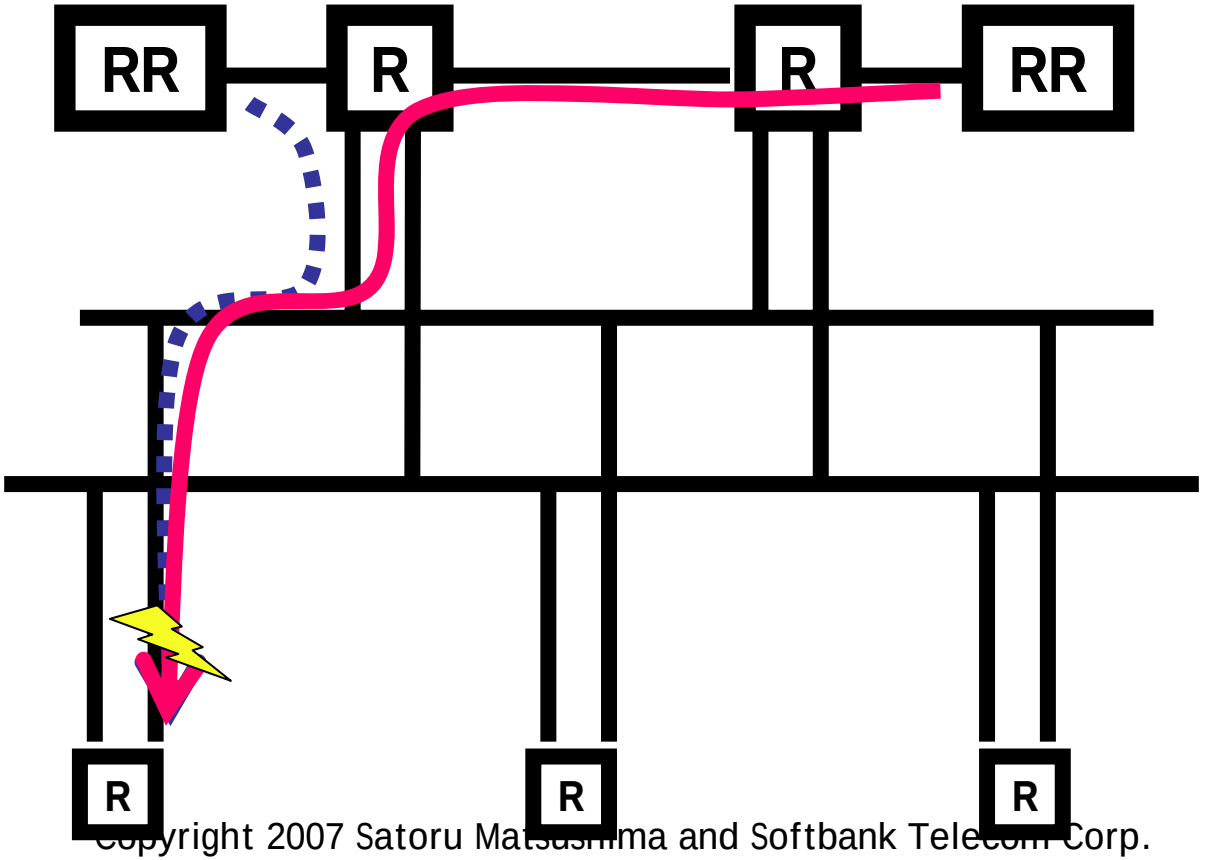
- 到達できるピアから通知してあげる
  （予備経路を強める）



Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

検証::障害検知時の動作

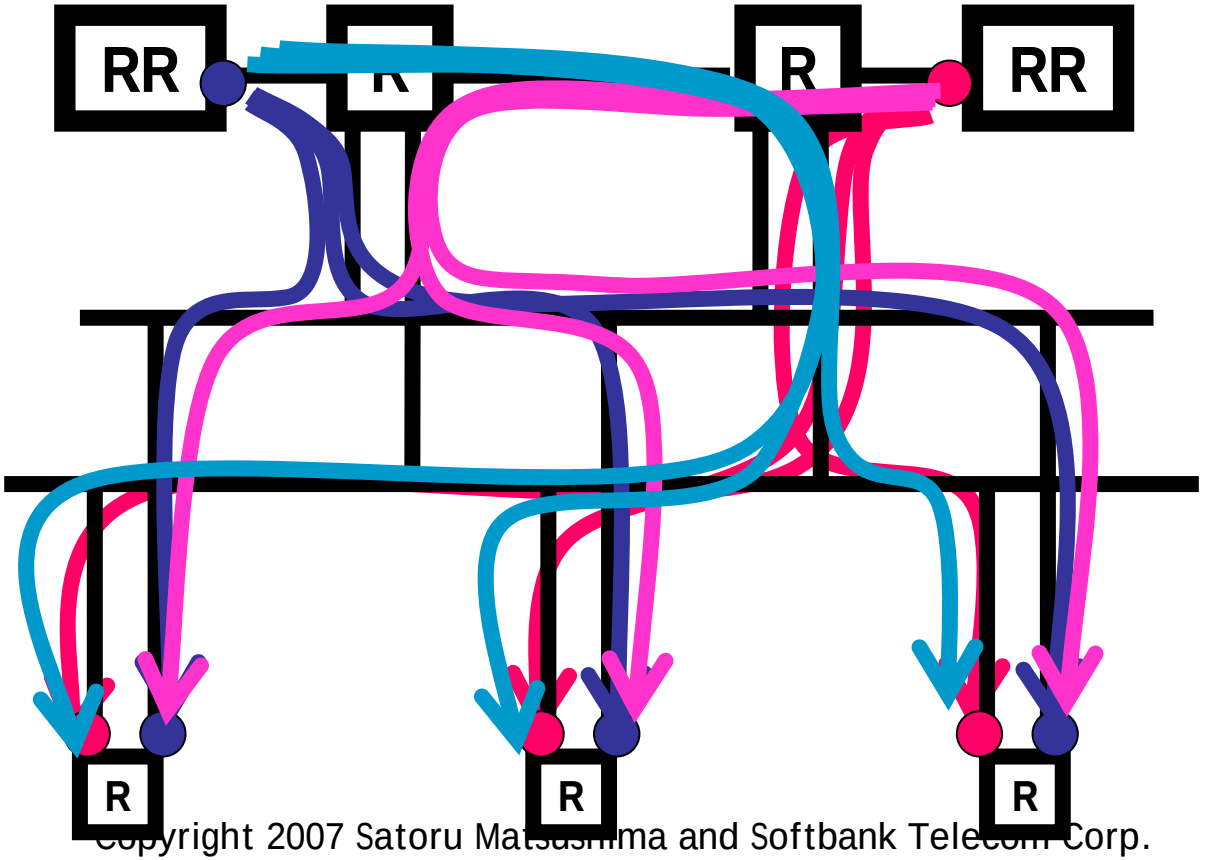
- プローブを追加してやる必要があった！



Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

検証::障害検知時の動作

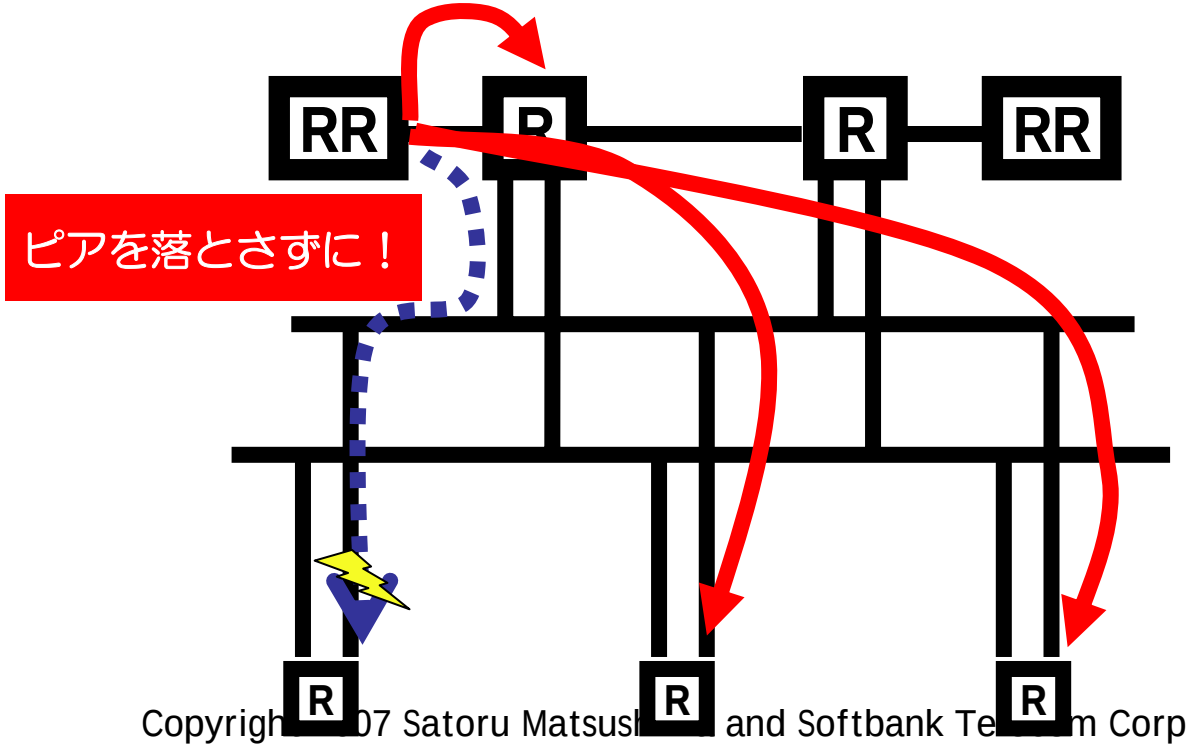
- やりなおし::プローブのかけ方



Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

検証後の動作ポリシ

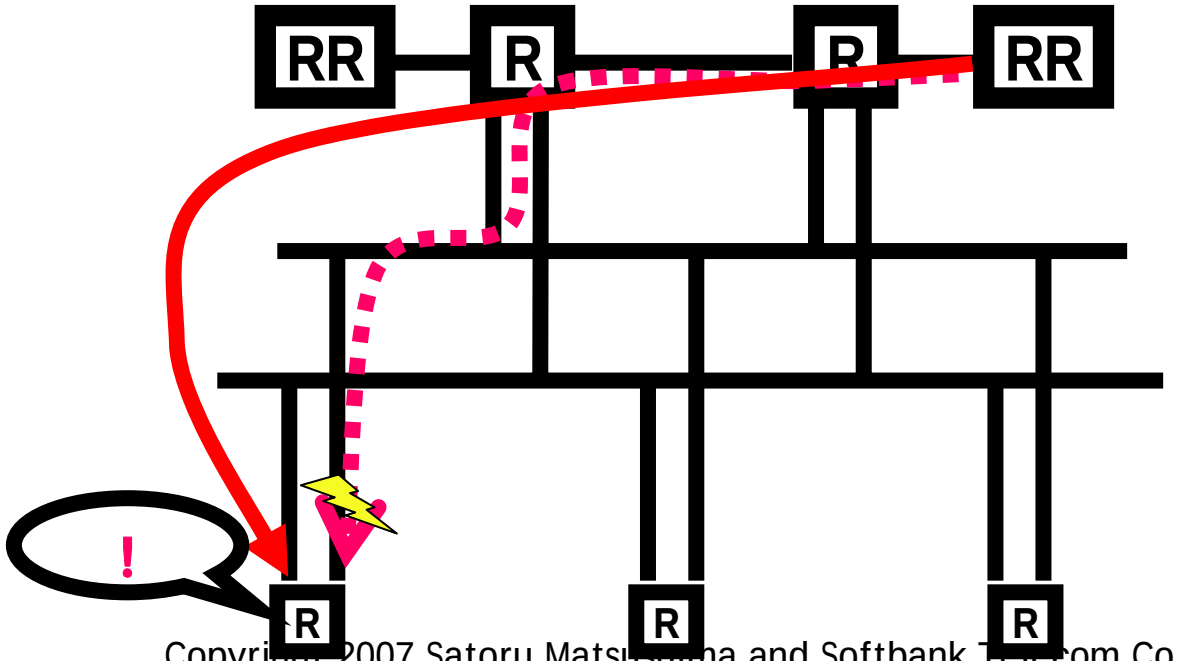
- 対象ルータ以外へwithdraw
  - ＢＧＰピアと同じパスのプローブで検出した場合



Copyright © 2007 Satoru Matsushima and Softbank Telecom Corp.

検証後の動作ポリシ

- 対象ルータへ 予備経路を強めるupdate
  - ＢＧＰピアと異なる（別のＲＲの）パスのプローブで検出した場合



Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

検証後の動作ポリシ

- プローブの保護時間
  - 障害検出するまで：
  - 障害復旧検出から実際に復旧させるまで：

プローブを受ける側と
プローブを送る側の
負荷をみながら慎重に

Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

まとめ::思いやる運用は実現できたか？

- BGP peer up/down はかなり抑制された
  - プローブがダウンしても、Peer downまで至らないケース多数
  - ICMPの負荷という話しもあるが、小さい箱にハードウェアBFDが載らない限り原理は同じ
- かなりアクロバティックな経路制御に…
  - プローブとBGPピアのパスを揃えるのが大変
    - マルチホップのBFDもこんな感じなのか…
  - 予備経路がいくつもある場合、優先順位を考えて設定するのが大変
- プローブはBFDでやれる事に越したことはない
  - 実際にパケットを転送するルータ自身が、プローブをもつ
  - でもやっぱり気になる、能力差の大きいピア
  - マルチホップ環境での高速検知が必要なケースがあったら、またこれをやることになるのだろう…
- RRで経路制御ポリシを作れるのはなかなかおもしろい
  - でも、ホントはRRはアトリビュートに変更を加えてはいけない

Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.

まとめ::終わりに

- BFD有史以前の、涙ぐましい高速障害復旧の実装・運用について紹介しました
  - 数千の小さい箱でも高速障害復旧
  - 汎用プロトコルをプローブにBFDライクな実装
  - アクロバティックな経路制御
- しかし、今回対象としたネットワークにBFDが導入されるのは果たしていつか？
  - BGPピアの能力差がはげしい環境
  - 低能力ルータには厳しい経路数を扱う環境
- 本質的には、経路数爆発状態における高速障害検知技術の運用を今から考えておきなさい、ということなのかも

Copyright 2007 Satoru Matsushima and Softbank Telecom Corp.