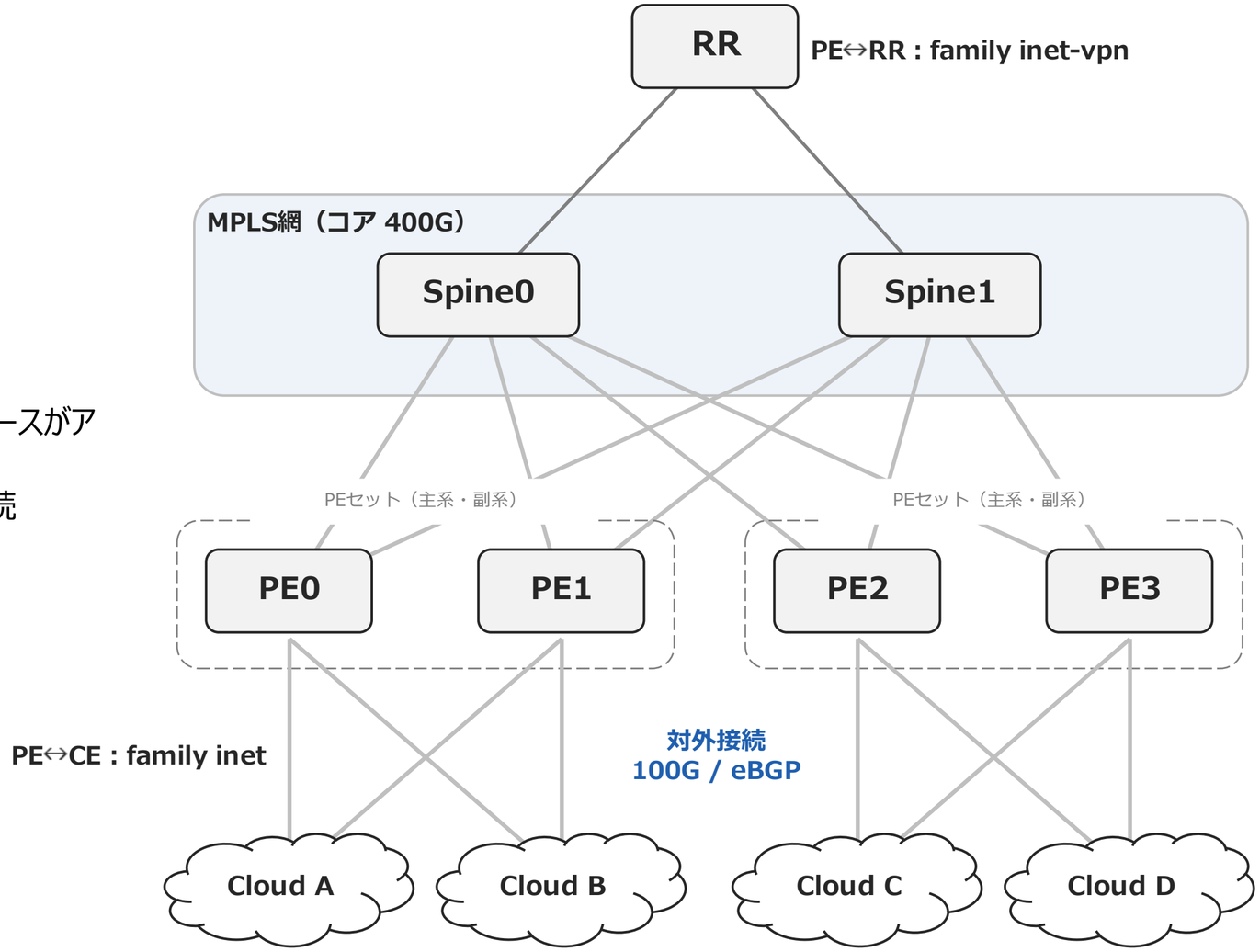
つなごう。驚きを。幸せを。

L3VPNの設計とラベル付与則

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

Flexible InterConnect (FIC) プレミアム

- クラウド事業者間を接続するL3VPN網
  - 基本は400Gで構成、対外接続は主として100G
  - 想定する収容物理回線総数: 数10～100+
  - 想定するPE数: 100セット+
  - 想定する経路数: 100万+/PE
- 本発表でのL3VPN: BGP/MPLSベースのIP-VPN
  - 概ね1ユーザが網内の1VRFに対応
  - PEのVRFには物理回線にvlanで重畳された論理インターフェースがアタッチされる
  - VRF内論理インターフェースにてBGP family inetで対外接続
  - PEはVRFの経路をBGP family inet-vpnでRRに広告



© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

設計にあたっての悩み

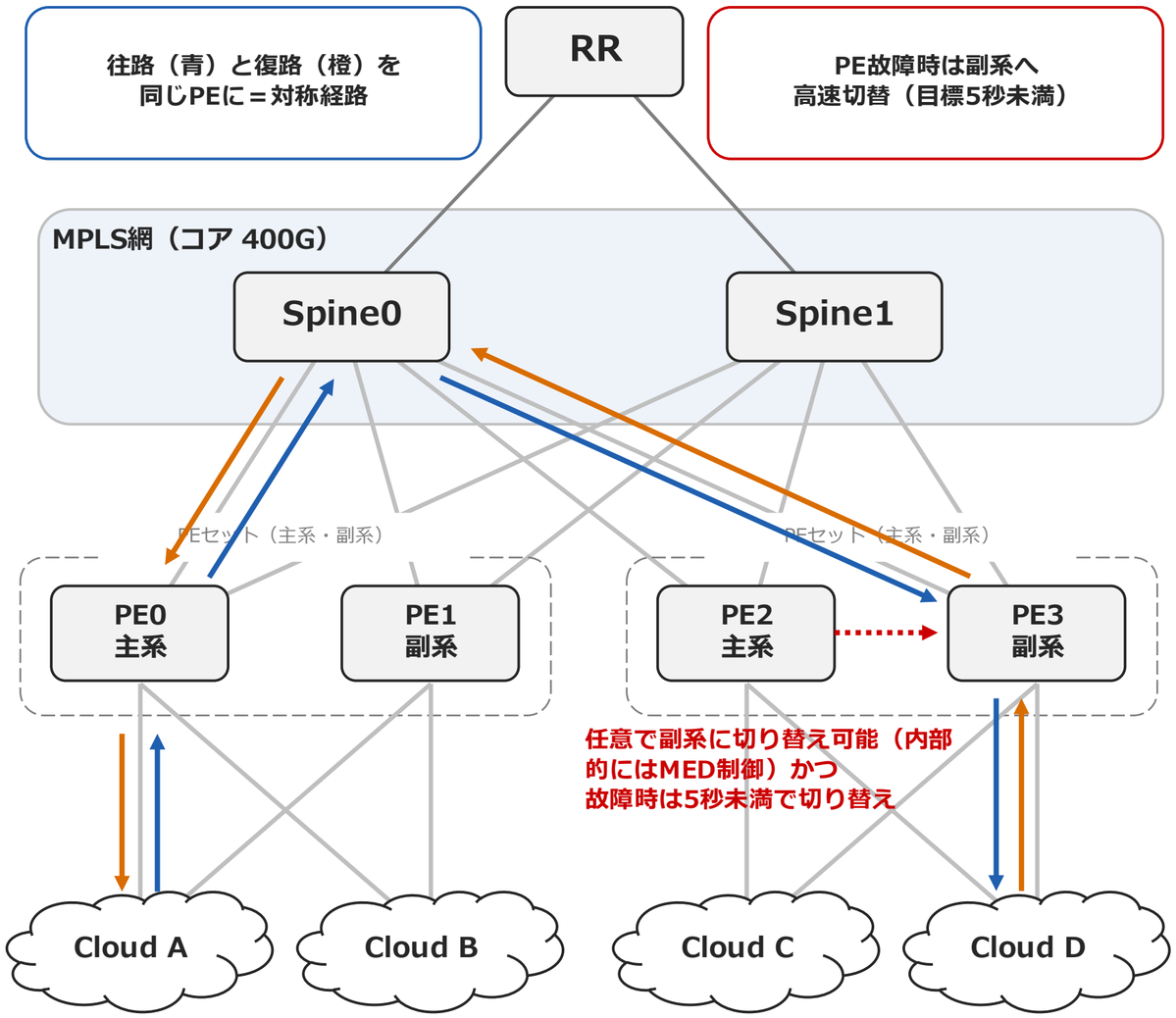
つなごう。驚きを。幸せを。
NTT docomo Business

- ロバストでスケーラブルなL3VPNを設計しようとすると、意外とまとまった情報がない…
- well-knownだから？もしくはなんとなく設計してもとりあえず動いてしまうから？
- 本日は、商用L3VPN網であるFICプレミアムの設計を通して得られた知見の中から、
  個人的に重要だと思うポイントを説明します

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

本日のテーマ

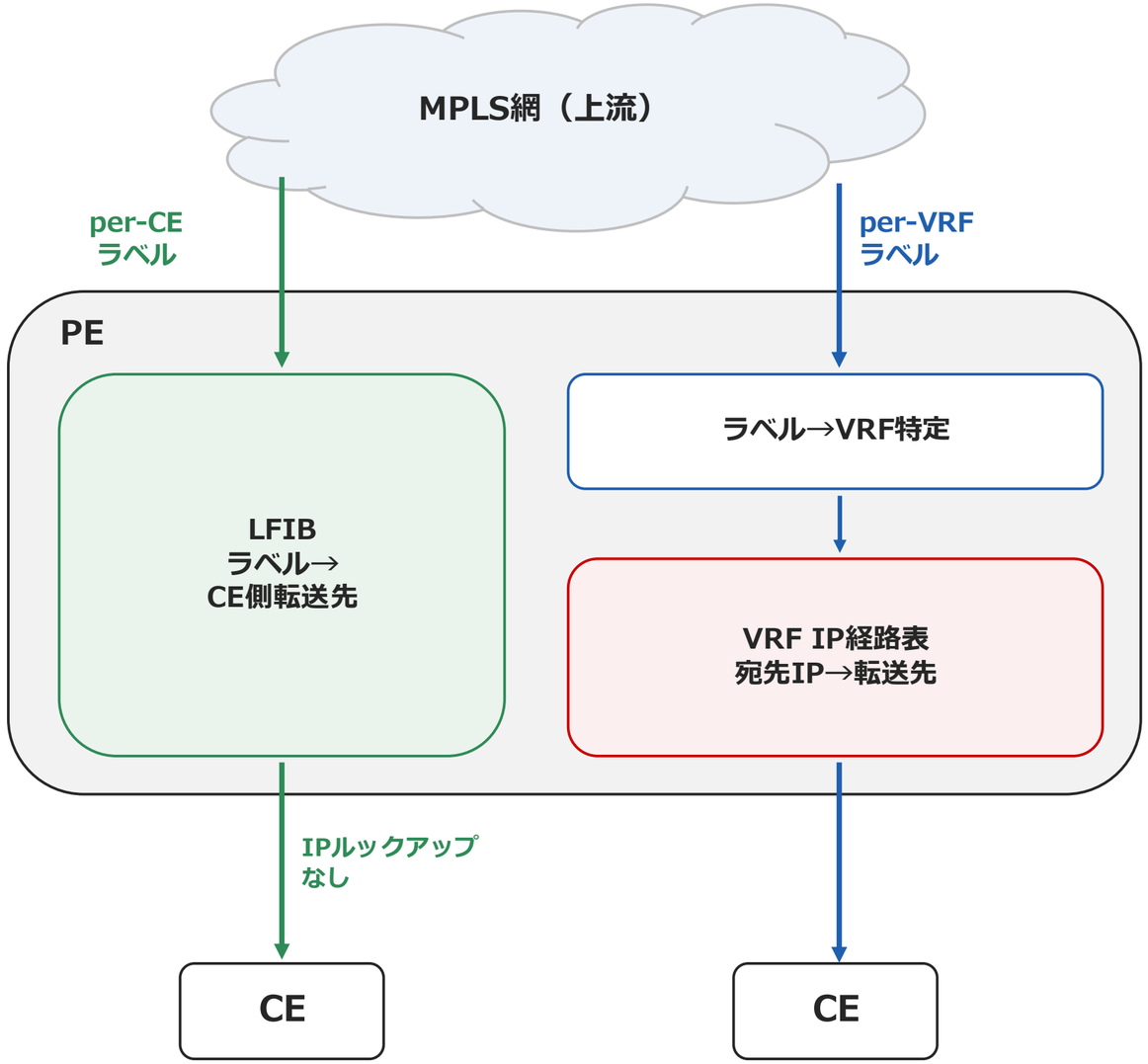
- L3VPNとして以下の要件をどうやって実現する？
  1. すべてのコネクション（VRF内でのBGP接続）はPEセット単位で冗長化（主系・副系）される
  2. ユーザは任意のタイミングで優先するPEを選択できる
     - さらに、往復での対称経路を担保する必要がある
       - 非対称経路が発生すると、クラウド内でステートフルファイアウォールなどを使用するユーザにとって不都合
  3. PEに故障が発生した際は対向の系へ高速に切り替えたい
     - 目標断時間：5秒未満
     - ただし、経路数の増加に対してスケーラブルである必要がある



© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

L3VPNのラベル付与則

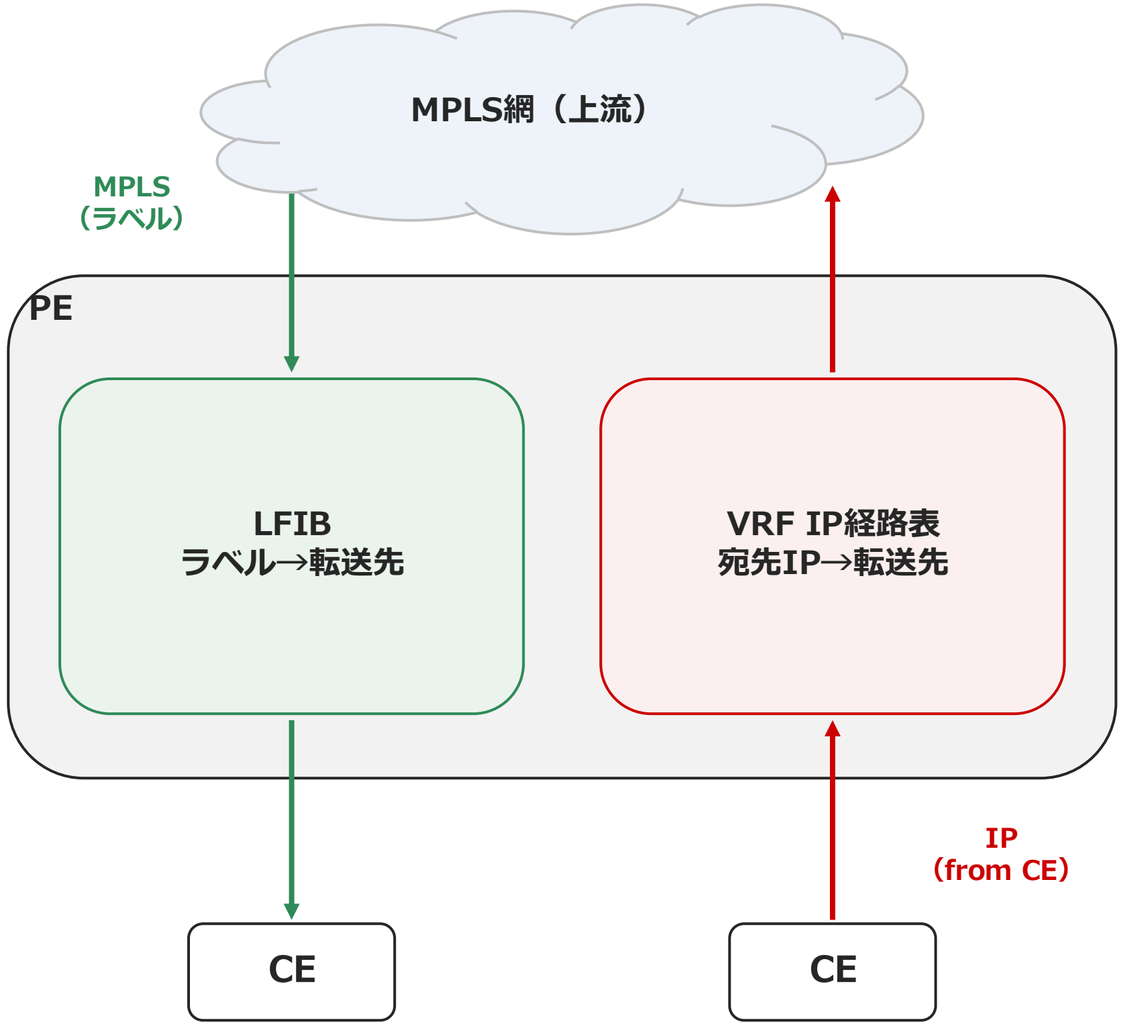
- ラベル付与則はL3VPNの挙動に大きく影響するため、設計上の最重要ポイント
  - per-CE / per-Prefix / per-VRF
  - 一般的にはラベル付与則選択の要素として「網内で扱うラベル数」と「IP的な経路制御の粒度」のトレードオフが意識されるが、それだけを考えて選択すると落とし穴がある
- 特にper-CE/Prefixとper-VRFで動作が大きく異なる
  - per-VRF：ラベルからVRFを特定し、宛先IPでIPルックアップ
    - 手軽かつIPネットワーク観点でもわかりやすいが、制約も大きい
  - per-CE：受信ラベルに対応するCE側転送先へLFIBに従い送出（IPルックアップはしない）
    - 慎重に設計すればMPLSのメリットをフルに享受できるが、CEへの送出処理はIPネットワーク観点では直感に反する



© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

PE内部での転送処理

- 上流から着信したMPLSパケットは各ラベルに対応するLFIBによって転送先を決める
- CEから受信したIPパケットは、VRFのIP経路表で転送先を決める
- つまり、同じ宛先でもIPで入る（from ダウンリンク）かラベルで入る（from アップリンク）かによって転送処理が異なる

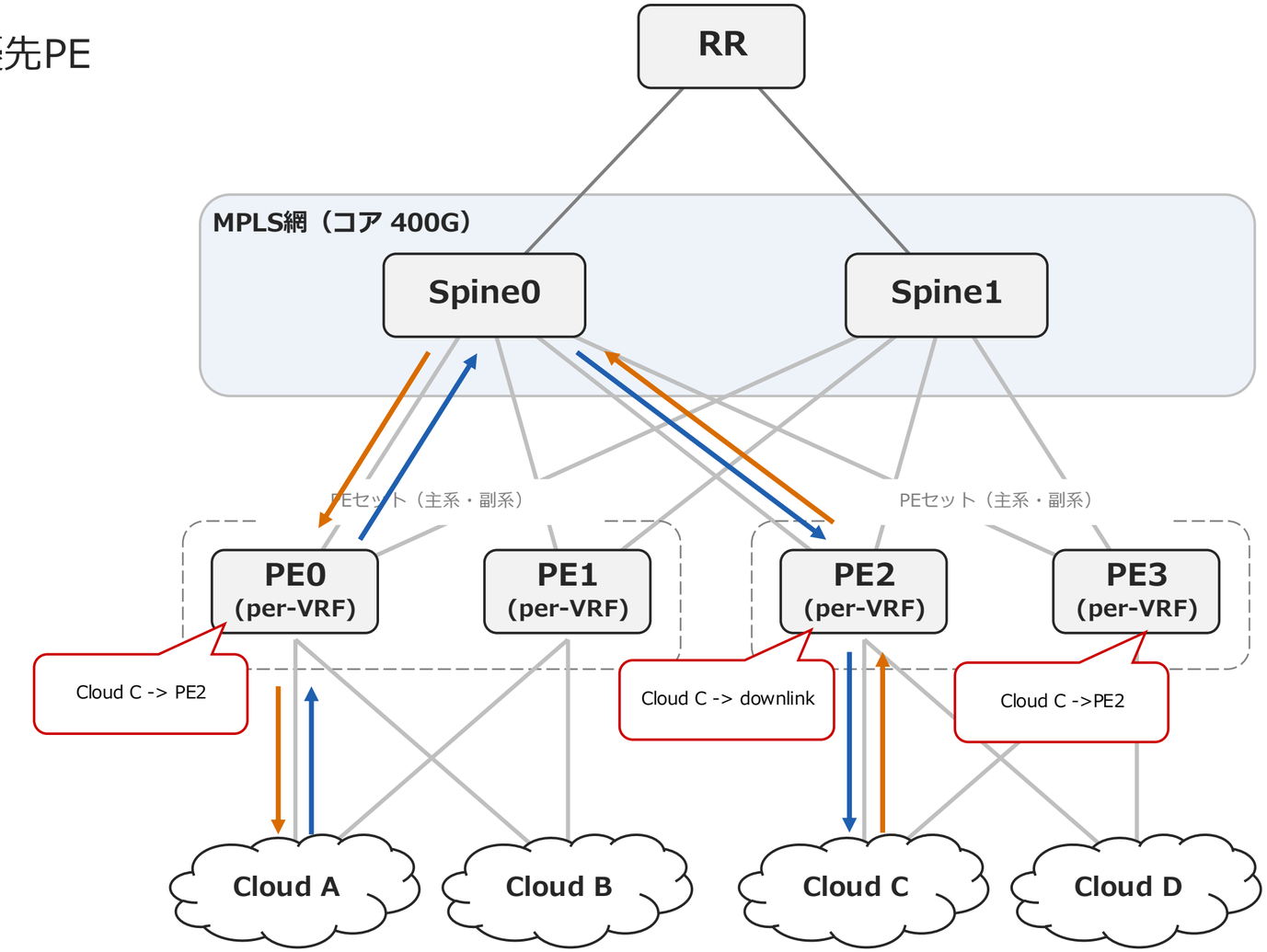


同じ宛先でも入口とラベル付与則で転送処理が変わる

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

per-VRFラベルを使った典型的な設計例

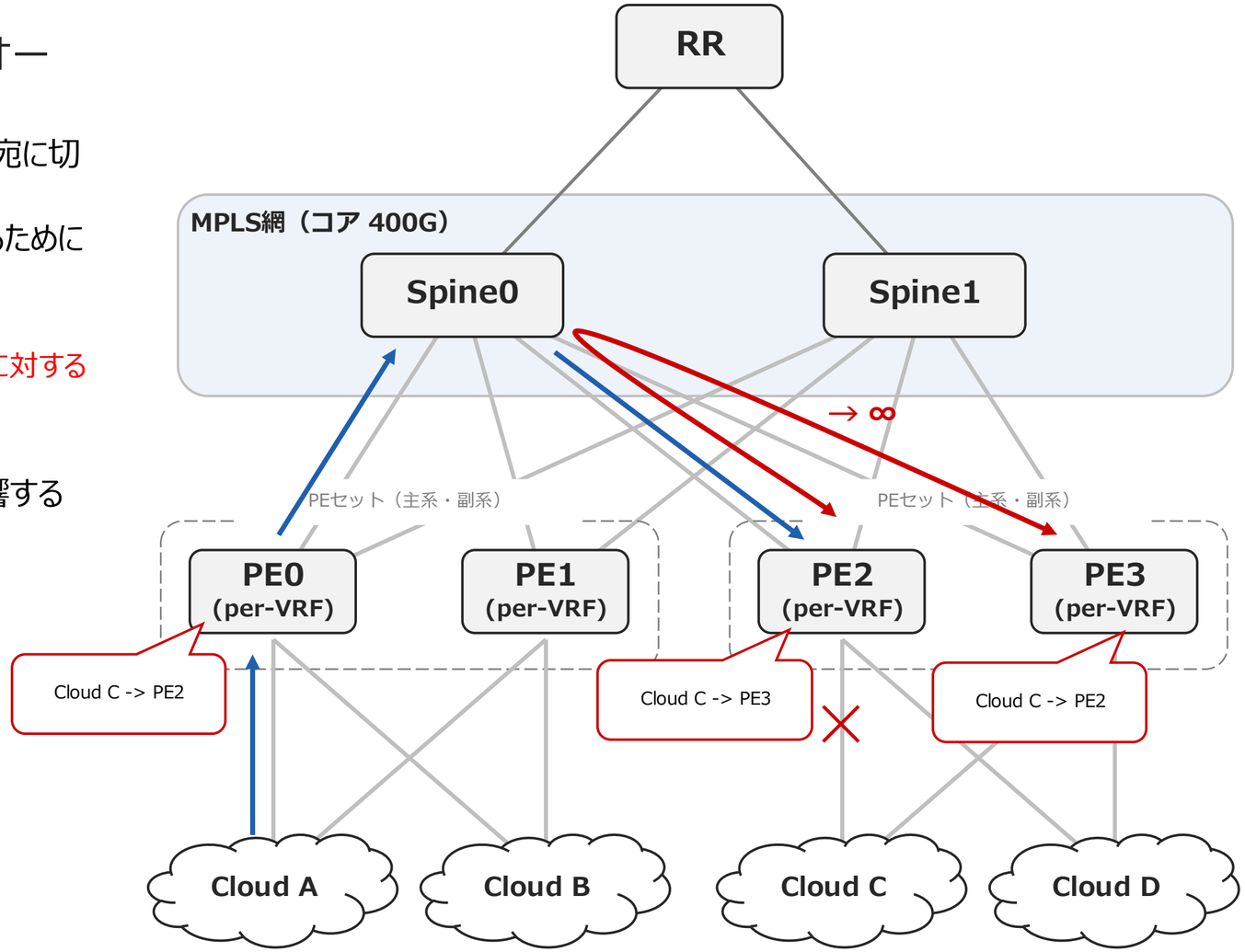
- 対向PEの経路を相互に保持し、両PEで接続ごとの優先PEを選ぶ



© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

per-VRFラベル最大の問題は経路数に対するスケーラビリティ

- 単純にper-VRFラベルを採用すると、PEのフェイルオーバーに要する時間が経路数によって増大する
  - PE2は自筐体ダウンリンクの断により比較的高速に経路をPE3宛に切り替えられる
  - 一方、PE0が優先をPE3へ、PE3が優先を自筐体に切り替えるために要する時間は長い
    - BGP Update + FIB Update
    - 経路数に依存して切り替えに要する時間は増加 -> 経路数に対するスケーラビリティが確保できない
  - 収束までの間、PE2 → PE3 → PE2の転送ループが発生する
  - ループによる輻輳が、故障した回線を使っていない通信にも影響する



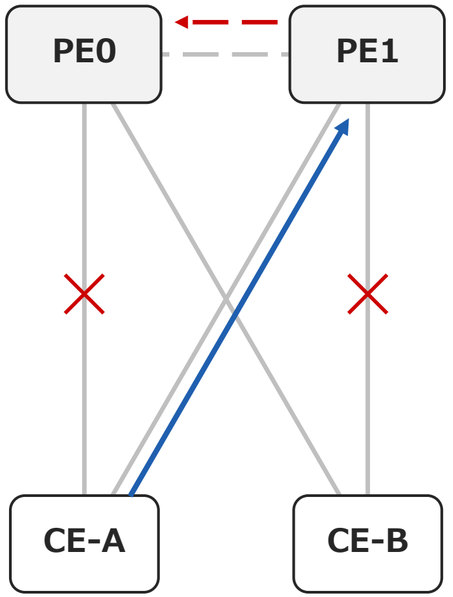
© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

PE主系・副系で、互いがeBGPから学習した経路をrejectすればいいのでは？

- 対向PE由来の経路を除外し、対向PEへの転送を制限する
  - 一見よさそうに思えるが、非対称経路・耐障害性の面で問題が出るため結果として解決にならない
- 耐障害性に問題が出るパターン
  - CE-AのPE0側回線と、CE-BのPE1側回線が同時に故障する
    - これはたとえば、別個のクラウド事業者の回線メンテナンスが同時刻に被る場合などで実際に発生しうる
    - この場合CE-AとCE-Bの間の通信には、CE-A → PE1 → PE0 → CE-Bの経路が必要になる
    - しかしPE1にPE0由来のCE-B宛経路がないため、通信断となる
- 非対称経路になるパターン
  - ユーザが優先度切り替え機能によりCE-BについてはPE1側を優先しているパターン
  - その場合であっても、CE-AからPE0に入ったCE-B宛通信はPE0側の回線へ出る
  - 結果としてCE-AとCE-Bの間の通信が非対称となる
  - ユーザからこのような内部構成は見えないため、「CE-BについてはPE1優先のつもりで運用しているはずが、特定のCEからは引き続きPE0から通信が入ってくる」ように見えてしまう

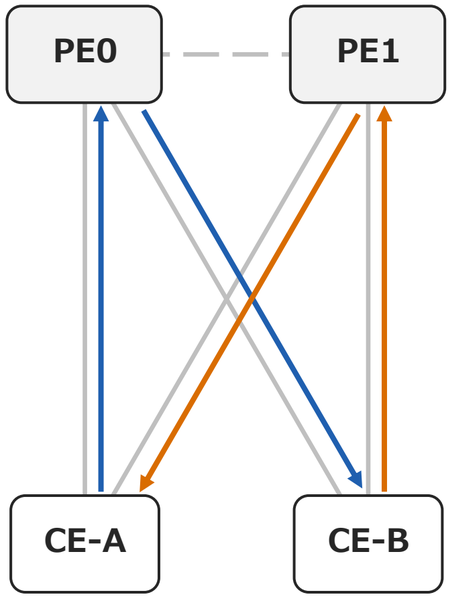
耐障害性の問題

PE1→PE0が必要



PE1にCE-B宛経路なし
（reject済）→ 不通

非対称経路の問題

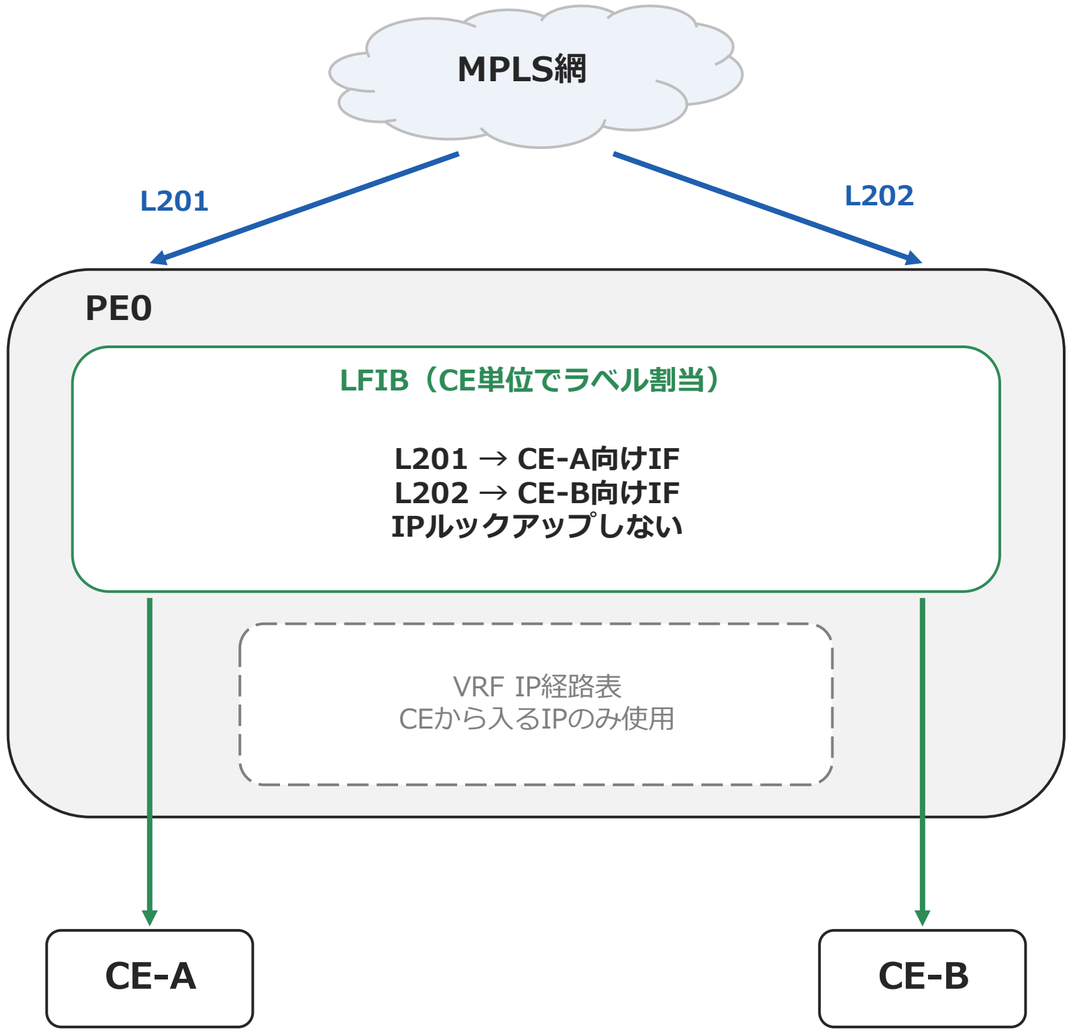


往路はPE0、復路はPE1
→ 非対称

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

per-CEラベル

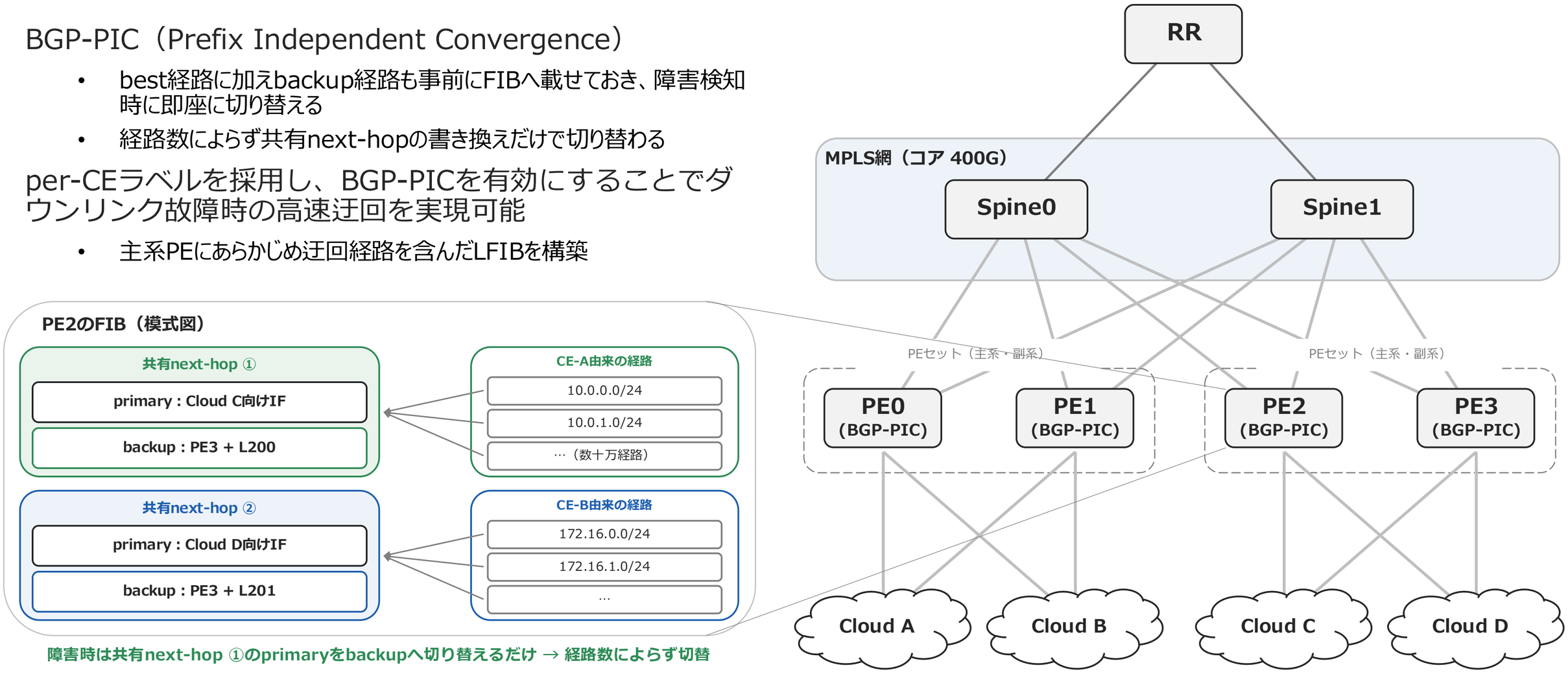
- 1 CEにつきラベルを1つアサインする
- ラベルつきのパケットが着信した際、受信ラベルに対応するCE側転送先へLFIBに従い送出する
  - IPルックアップはしない



© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

per-CEラベル + BGP-PICによる高速迂回

- BGP-PIC（Prefix Independent Convergence）
  - best経路に加えbackup経路も事前にFIBへ載せておき、障害検知時に即座に切り替える
  - 経路数によらず共有next-hopの書き換えだけで切り替わる
- per-CEラベルを採用し、BGP-PICを有効にすることでダウンリンク故障時の高速迂回を実現可能
  - 主系PEにあらかじめ迂回経路を含んだLFIBを構築

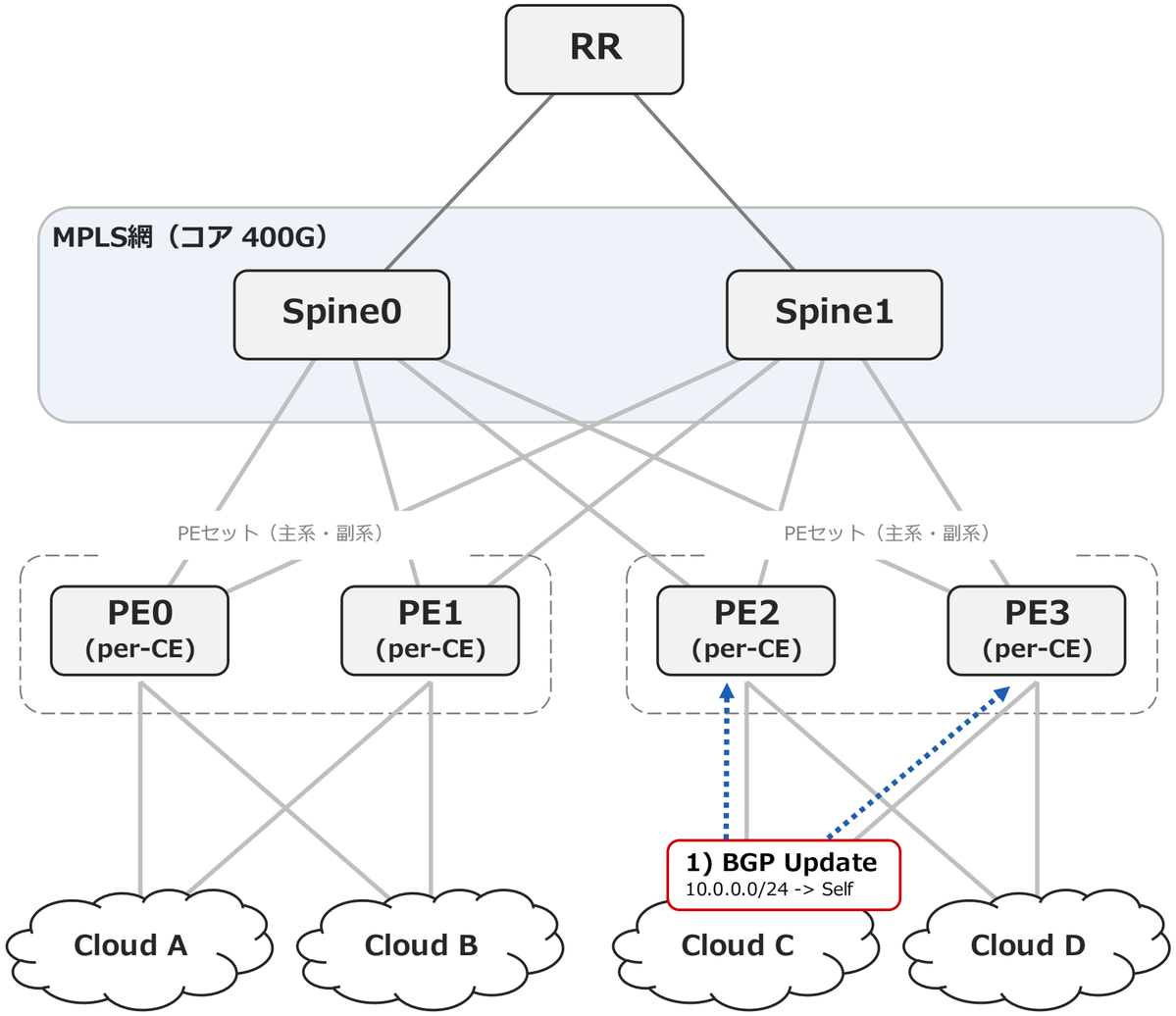


障害時は共有next-hop ①のprimaryをbackupへ切り替えるだけ → 経路数によらず切替

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

per-CEラベル + BGP-PICによる高速迂回

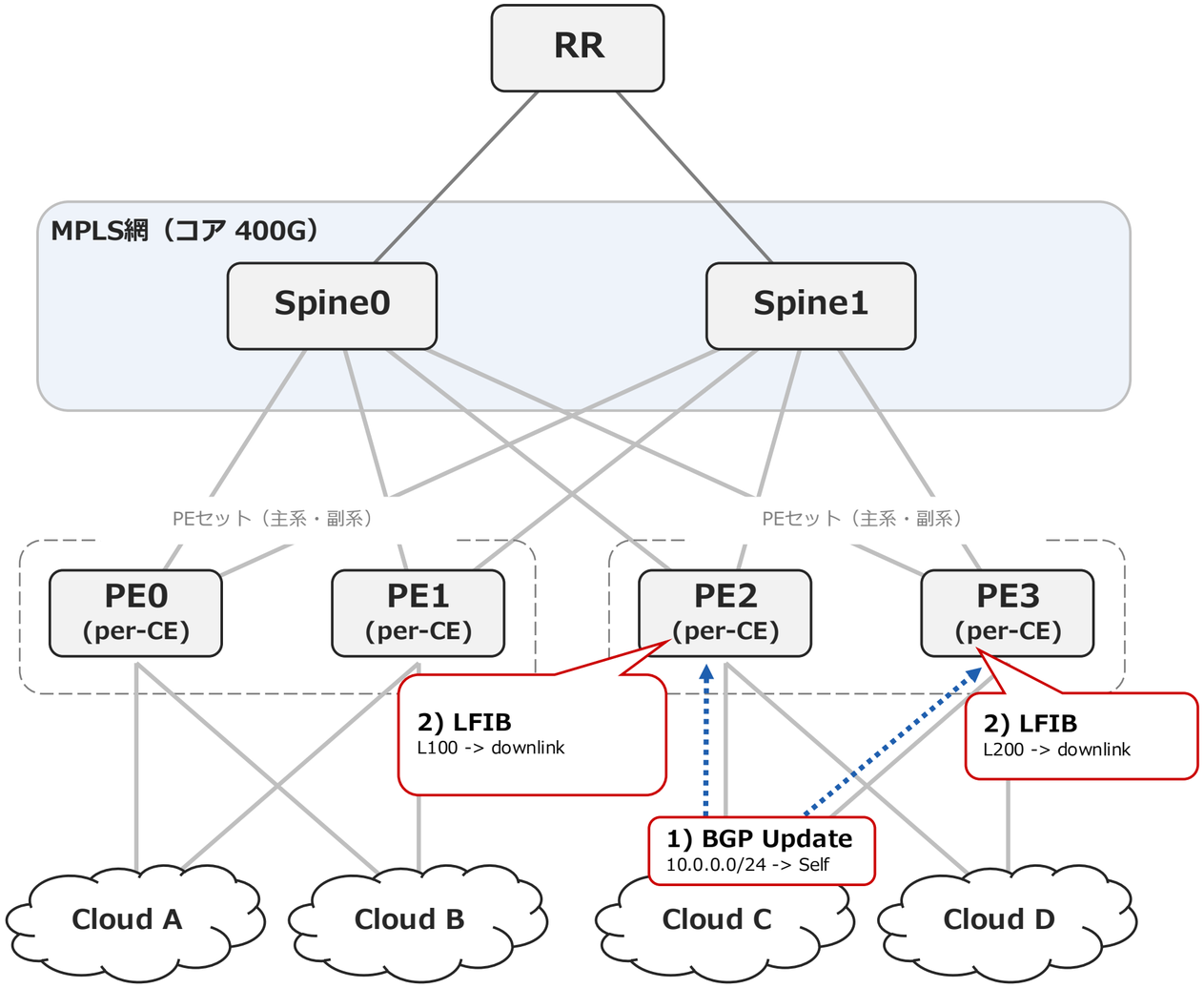
- 初期収束の流れは以下
    1. Cloudから経路広告を受ける（10.0.0.0/24）



© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

per-CEラベル + BGP-PICによる高速迂回

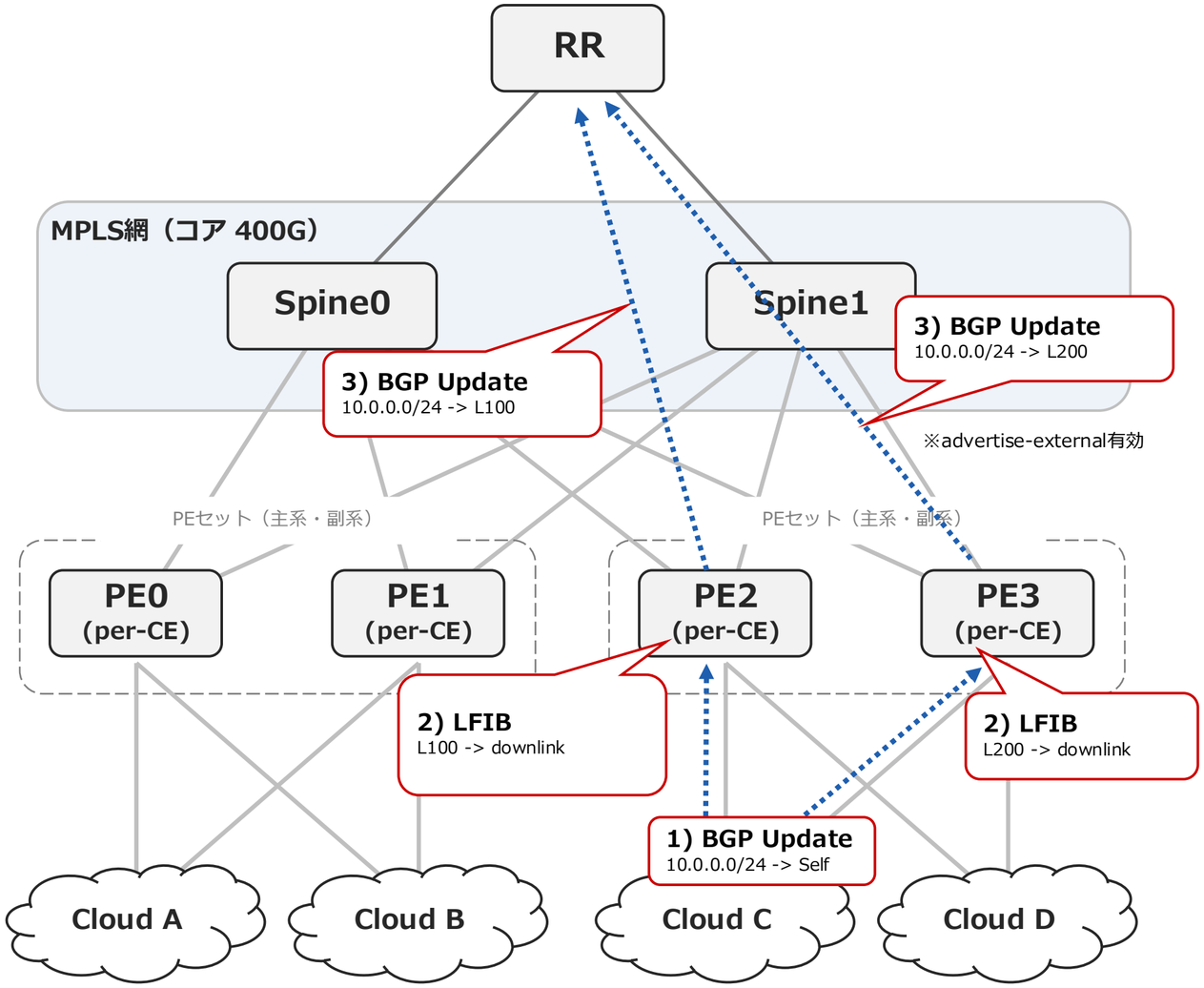
- 初期収束の流れは以下
    1. Cloudから経路広告を受ける（10.0.0.0/24）
    2. PEそれぞれがLFIBを構築（PICでないダウンリンク単体行きのもの）



© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

per-CEラベル + BGP-PICによる高速迂回

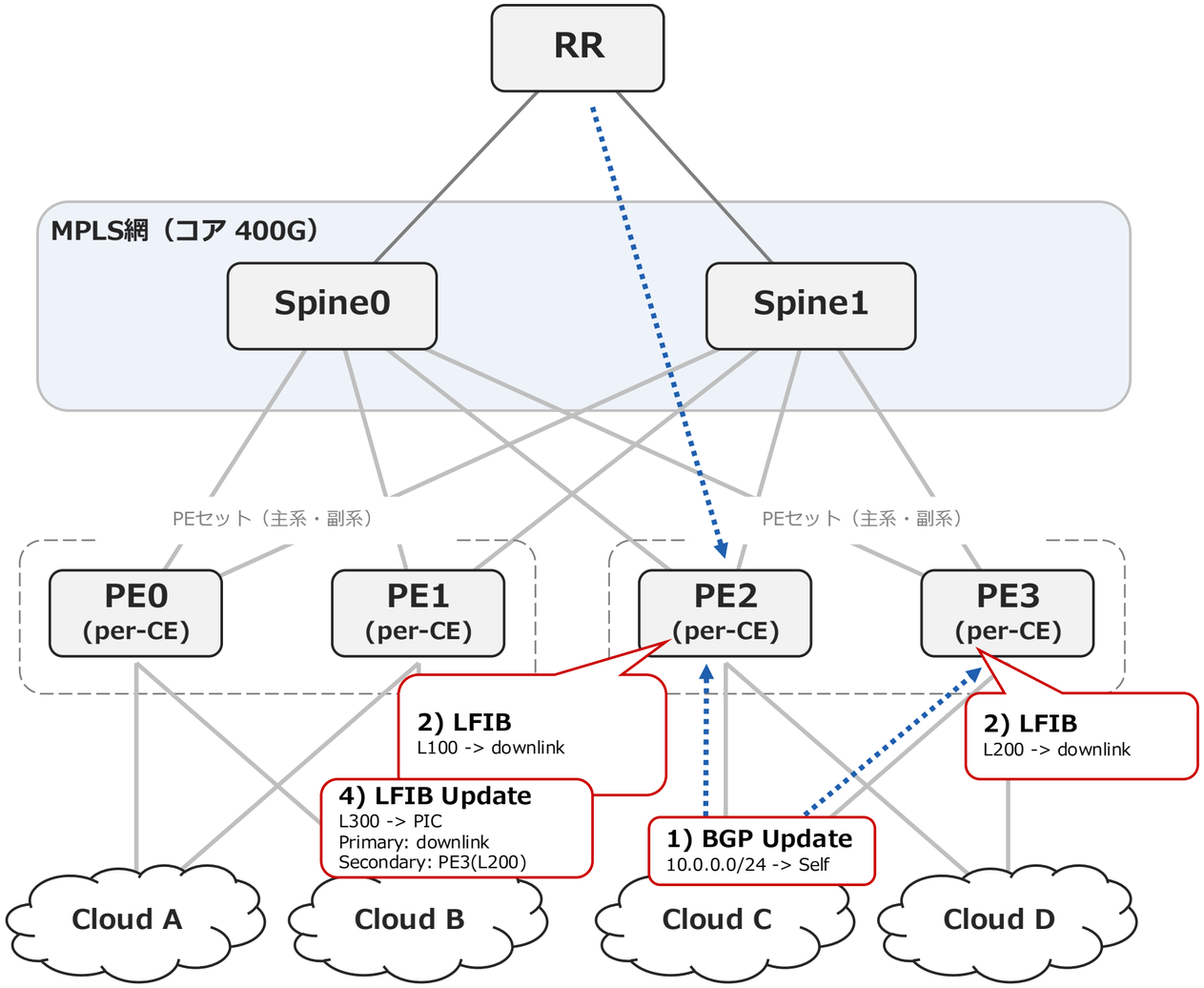
- 初期収束の流れは以下
    1. Cloudから経路広告を受ける（10.0.0.0/24）
    2. PEそれぞれがLFIBを構築（PICでないダウンリンク単体行きのもの）
    3. PEそれぞれが10.0.0.0/24にラベルをつけてBGP Update



© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

per-CEラベル + BGP-PICによる高速迂回

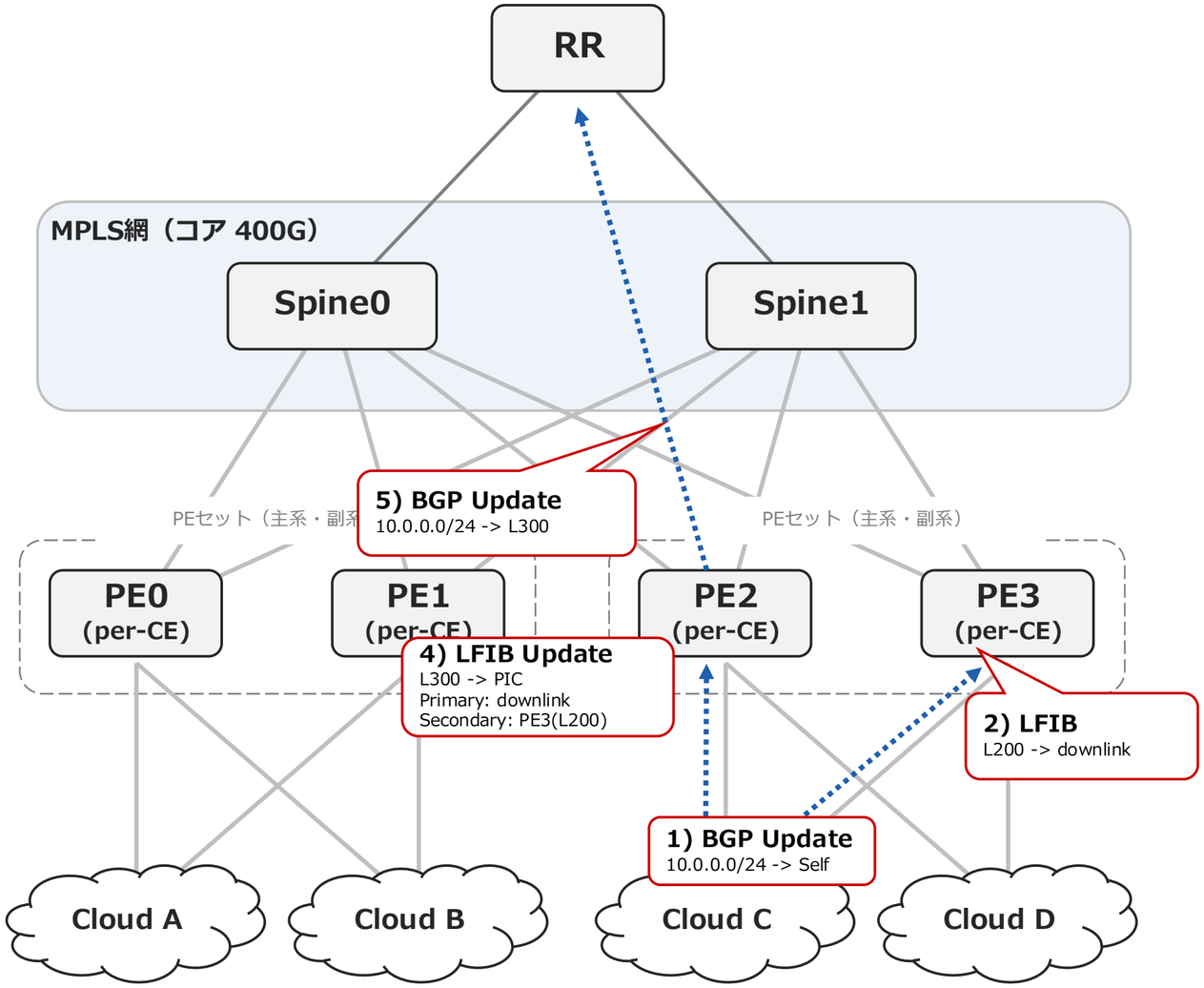
- 初期収束の流れは以下
    1. Cloudから経路広告を受ける（10.0.0.0/24）
    2. PEそれぞれがLFIBを構築（PICでないダウンリンク単体行きのもの）
    3. PEそれぞれが10.0.0.0/24にラベルをつけてBGP Update
    4. 優先側のPEが非優先側PEの経路を受け取ってPIC用ラベルをアサインしLFIBを更新



© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

per-CEラベル + BGP-PICによる高速迂回

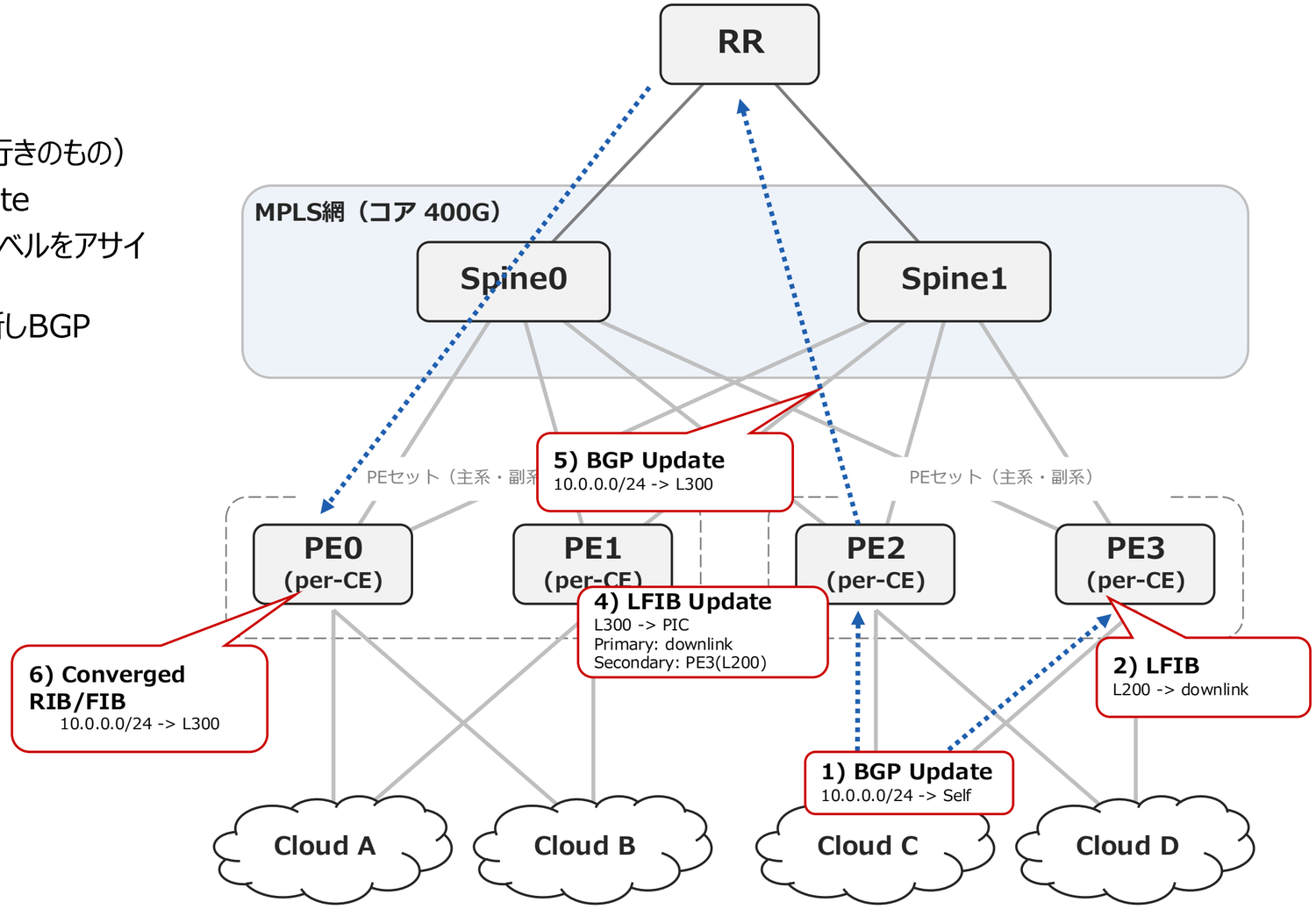
- 初期収束の流れは以下
    1. Cloudから経路広告を受ける（10.0.0.0/24）
    2. PEそれぞれがLFIBを構築（PICでないダウンリンク単体行きのもの）
    3. PEそれぞれが10.0.0.0/24にラベルをつけてBGP Update
    4. 優先側のPEが非優先側PEの経路を受け取ってPIC用ラベルをアサインしLFIBを更新
    5. 優先側のPEが10.0.0.0/24のラベルをPICのものに更新しBGP Update



© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

per-CEラベル + BGP-PICによる高速迂回

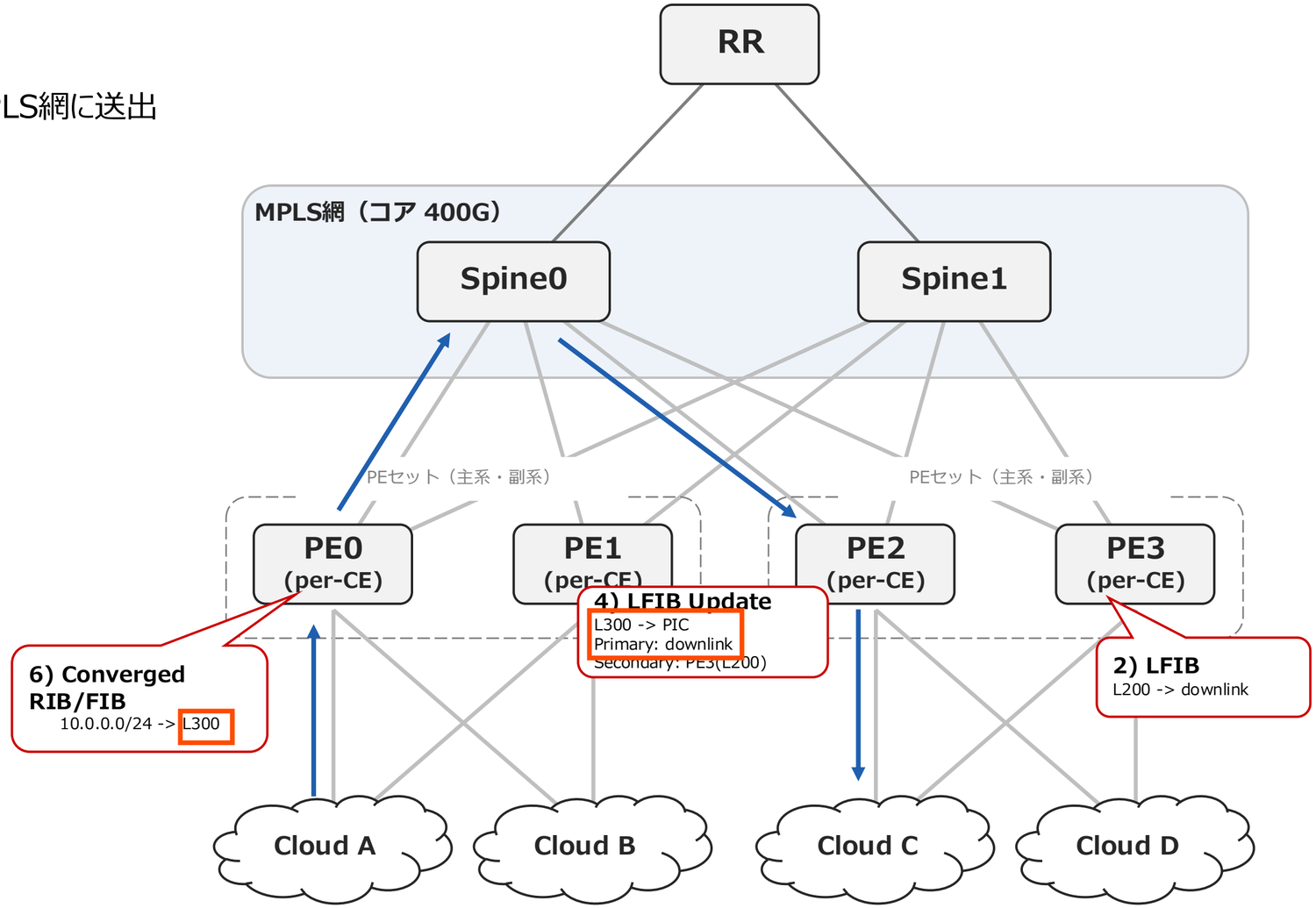
- 初期収束の流れは以下
    1. Cloudから経路広告を受ける（10.0.0.0/24）
    2. PEそれぞれがLFIBを構築（PICでないダウンリンク単体行きのもの）
    3. PEそれぞれが10.0.0.0/24にラベルをつけてBGP Update
    4. 優先側のPEが非優先側PEの経路を受け取ってPIC用ラベルをアサインしLFIBを更新
    5. 優先側のPEが10.0.0.0/24のラベルをPICのものに更新しBGP Update
    6. 送信側PEで経路収束



© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

per-CEラベル + BGP-PICによる高速迂回

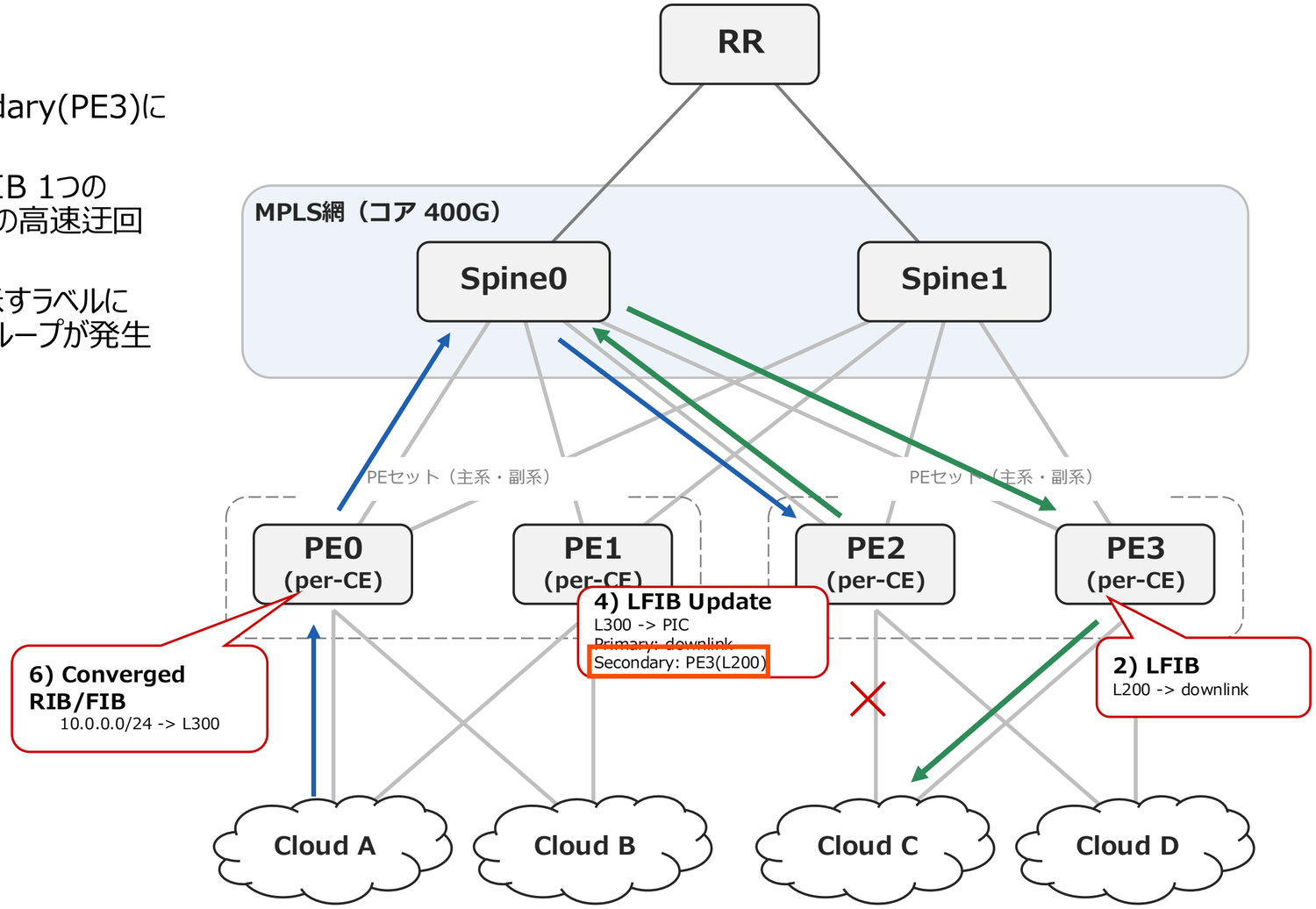
- 正常時
  - PE0は10.0.0.0/24宛のパケットをL300でEncapしMPLS網に送出
  - PE2はPICのLFIBに従いパケットをダウンリンクに送出



© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

per-CEラベル + BGP-PICによる高速迂回

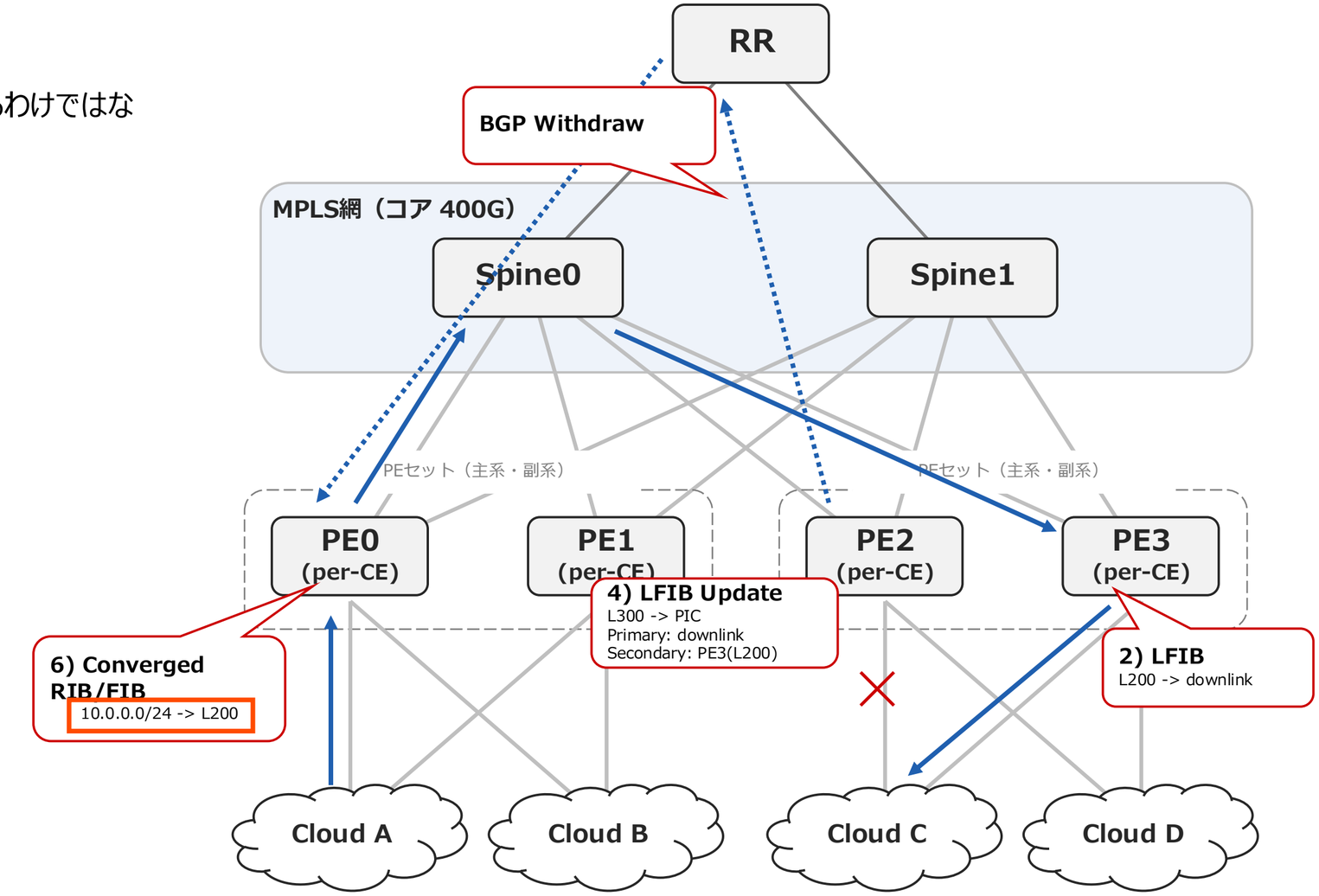
- ダウンリンクの故障発生時
  - ダウンリンク故障時はPE2があらかじめ設定されたSecondary(PE3)に向かってパケットを高速迂回
  - 同じCEの全経路が単一のラベルを参照しているため、LFIB 1つのPrimaryを切り替えるだけで済み、主系PE -> 副系PEの高速迂回を実現可能
  - 迂回先の副系PEではIPルックアップをせずdownlinkを示すラベルに従ってパケットを送出するのみであるため、経路収束中のループが発生しない



© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

per-CEラベル + BGP-PICによる高速迂回

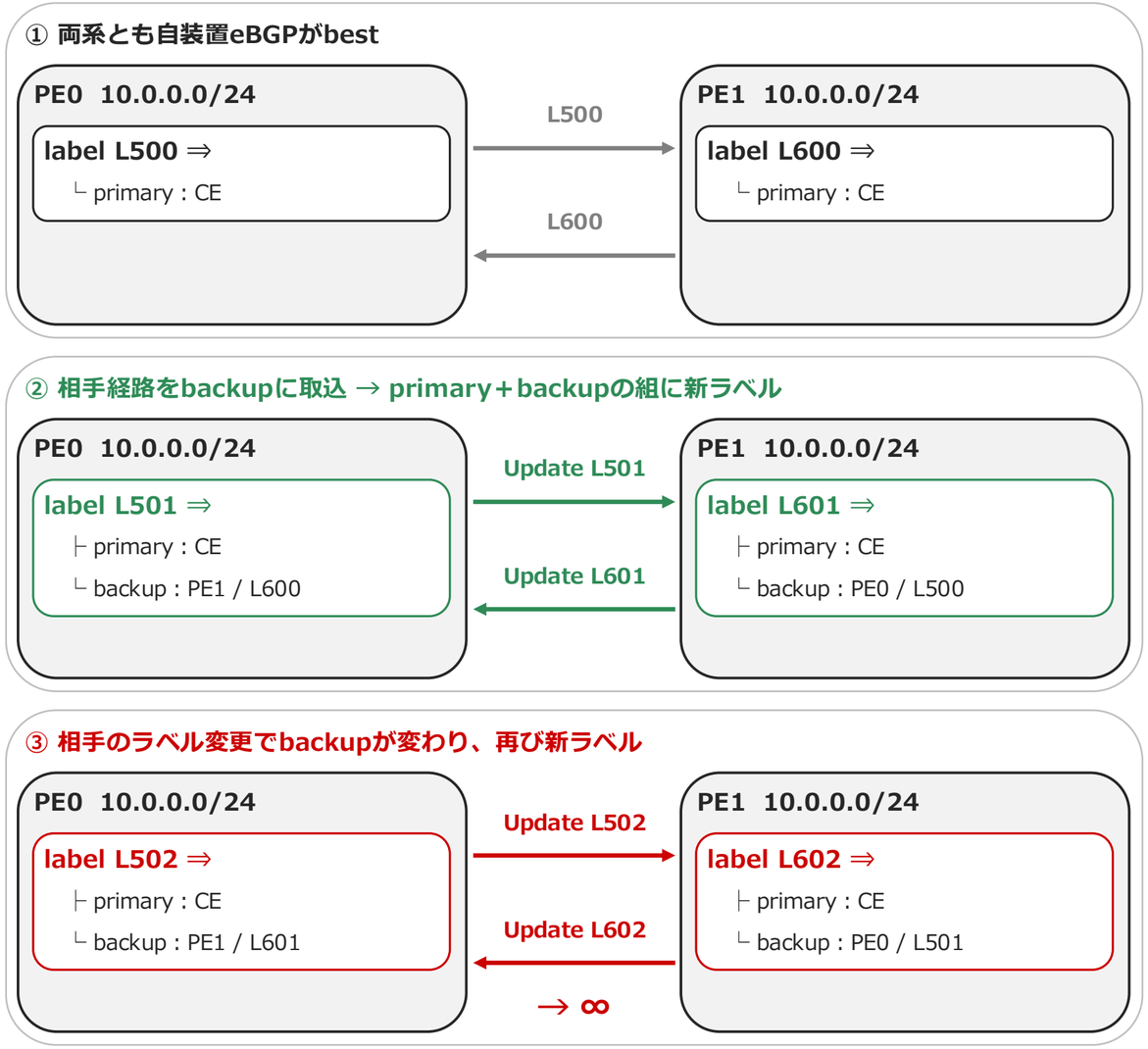
- ダウンリンクの故障発生時
  - 経路収束後は副系PEが優先になる（迂回が永続するわけではない）



© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

ただし、運用難度はやや上がる

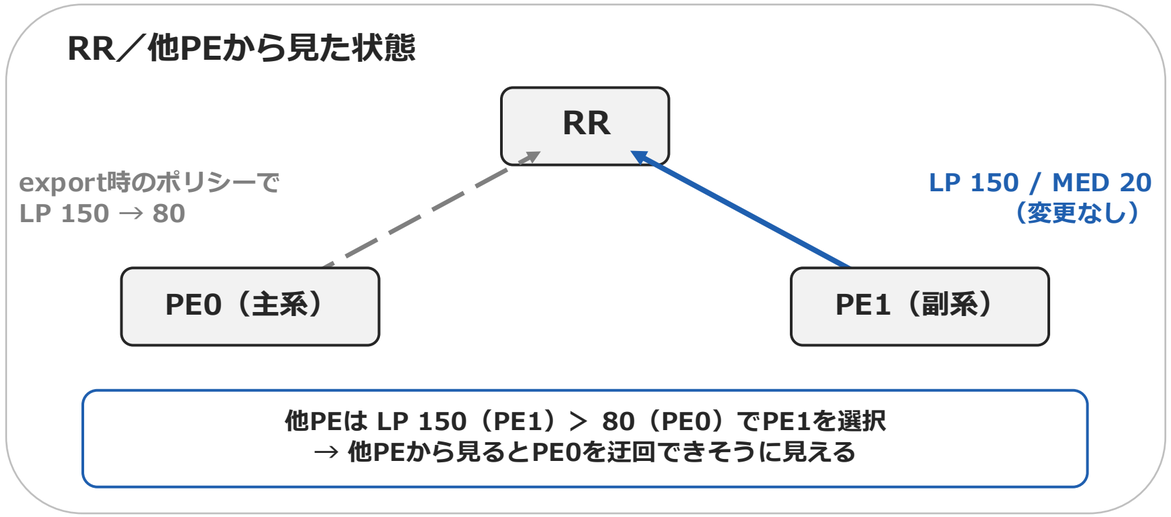
- 例えば、per-CE＋BGP-PICの両系Act構成でのラベル更新循環問題
- per-CEラベル ＋ BGP-PICを使いつつ、両系のPEで自身のeBGP経路を優先するとコントロールプレーンでBGP Updateのループが発生する
  - per-CEラベル ＋ BGP-PIC特有の落とし穴？
  - Juniper MXの実装依存である可能性
- つまり、この構成では事実上Act / Sby構成が必須
  - PE0／PE1が、それぞれ自装置のeBGP経路をベストに選ぶようなメトリック設計は不可

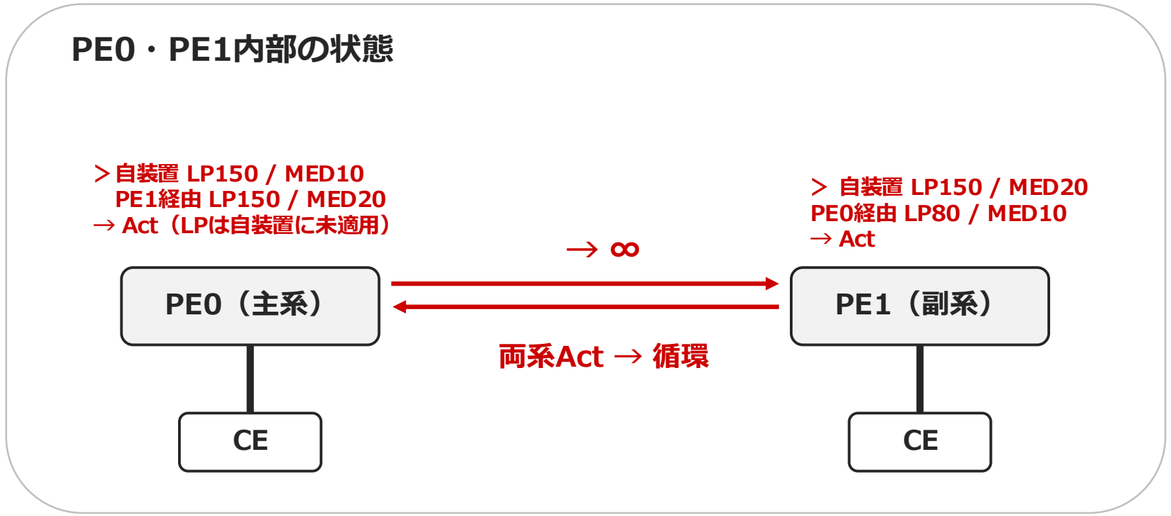


© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

ラベル更新循環の発生例

- 迂闊にBGPのメトリックを変えると両系Actとなり、経路広告のループが発生する
    1. PE0を迂回するためにRR向けのexportフィルタでLPを下げる
    2. PE0的には自分が優先のままだが、PE1から見るとPE0の経路のLPが下がり、PE1的にも新たに自分が優先となる
    3. 結果、両系PEで自身のダウンリンクが優先されてしまい、前項のラベル更新循環が発生
    - PE迂回・復帰操作の途中も含め、両PEが自装置側を優先する状態を避ける必要がある





© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

まとめ

つなごう。驚きを。幸せを。
NTT docomo Business

- BGP/MPLSベースのIP-VPNは広く普及する一方、その設計手法については体系的な情報が少ない
- ラベル付与則はL3VPNの挙動に大きく影響するため、設計上の最重要ポイント
    - 特にper-CE/Prefixラベルとper-VRFラベルの挙動差は大きい
- per-CEラベルとBGP-PICにより経路数に依存しない高速迂回と経路の対称性を両立できるが、運用難度はやや上昇する

-> 同じような点で苦労した経験、他にこういう技術がある、など幅広いコメントをいただけると幸いです。

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.

つなごう。驚きを。幸せを。

NTT docomo Business

© NTT DOCOMO BUSINESS, Inc. All Rights Reserved.