インターネットの経路あれこれ

JANOG28

NTT Communications

益子 直樹

吉村 知夏

自己紹介

OCN

益子

吉村

自己紹介

OCN

益子

吉村

本日のゴール

- BGP運用における悩みを共有させてください
- よりよい運用への議論をしたい

Facebookアンケート

Facebookアンケート

Facebookアンケート

『お仕事や学校でBGPを使う機会はどれくらいありますか？』

Facebookアンケート

『お仕事や学校でBGPを使う機会はどれくらいありますか？』

- もっとBGPを知りたいでござる：46votes
- AS運用/設計者：29votes
- 良く使います！：21votes
- 知ってはいるけどあんまり使わないかなぁ：14votes
- 不良になる：5votes
- 男は黙ってIGP：4votes
- Static教：4votes
- なんの略でしたっけ？：1votes

計163votes。たくさんのご回答ありがとうございました☺

ISPの悩み

@JANOG27

- トラヒックの増加
- 経路のいろいろ

ISPの悩み

@JANOG27

- トラヒックの増加

Today's Focus!

- 経路のいろいろ

お題　2つ

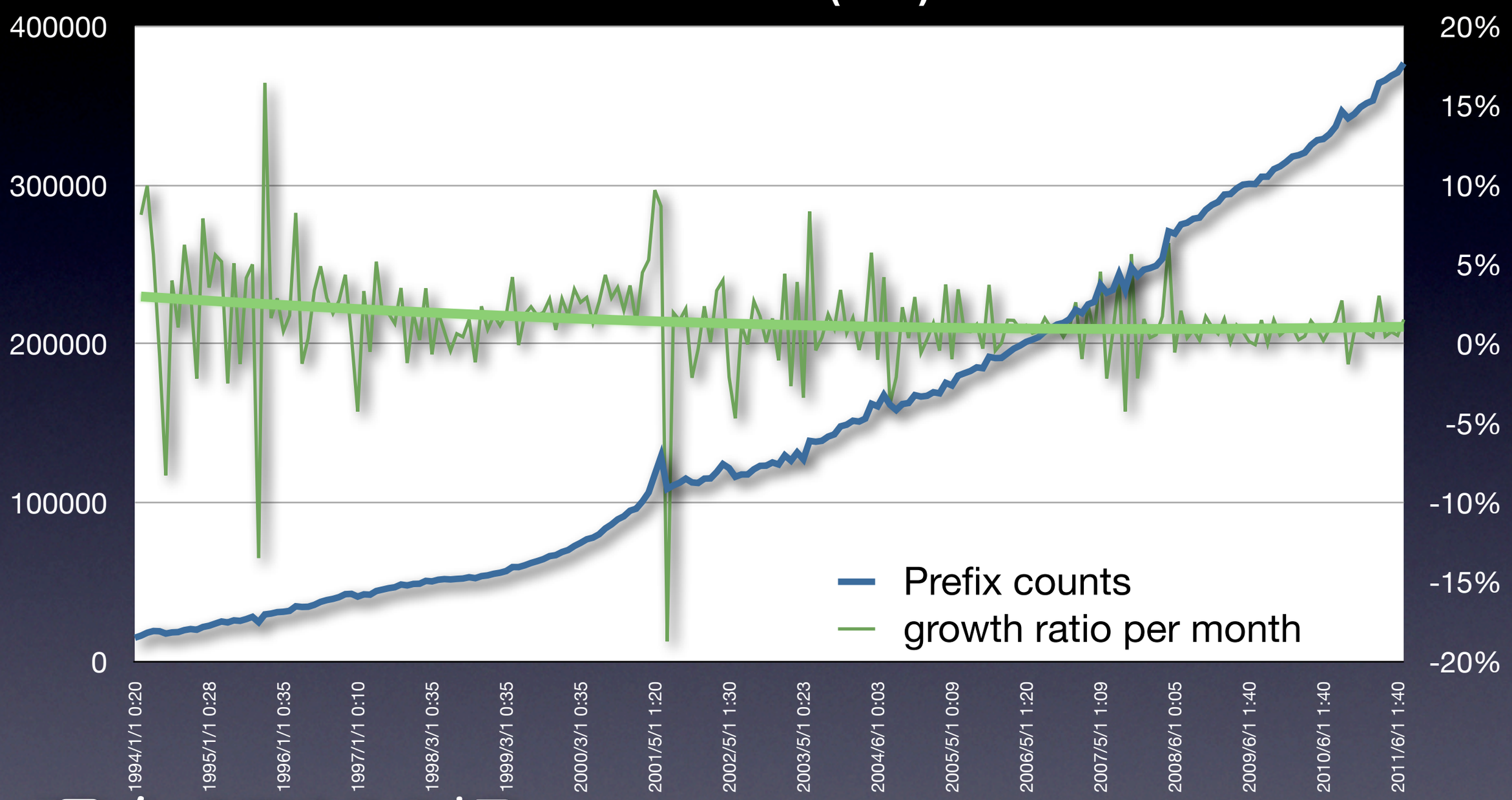
- 経路数増大とその対処法
- 異常経路とその対処法
- （番外編）東日本大震災による経路変動

経路数増大

370000 : 10%

6400 : 60%

BGP経路数(IPv4)

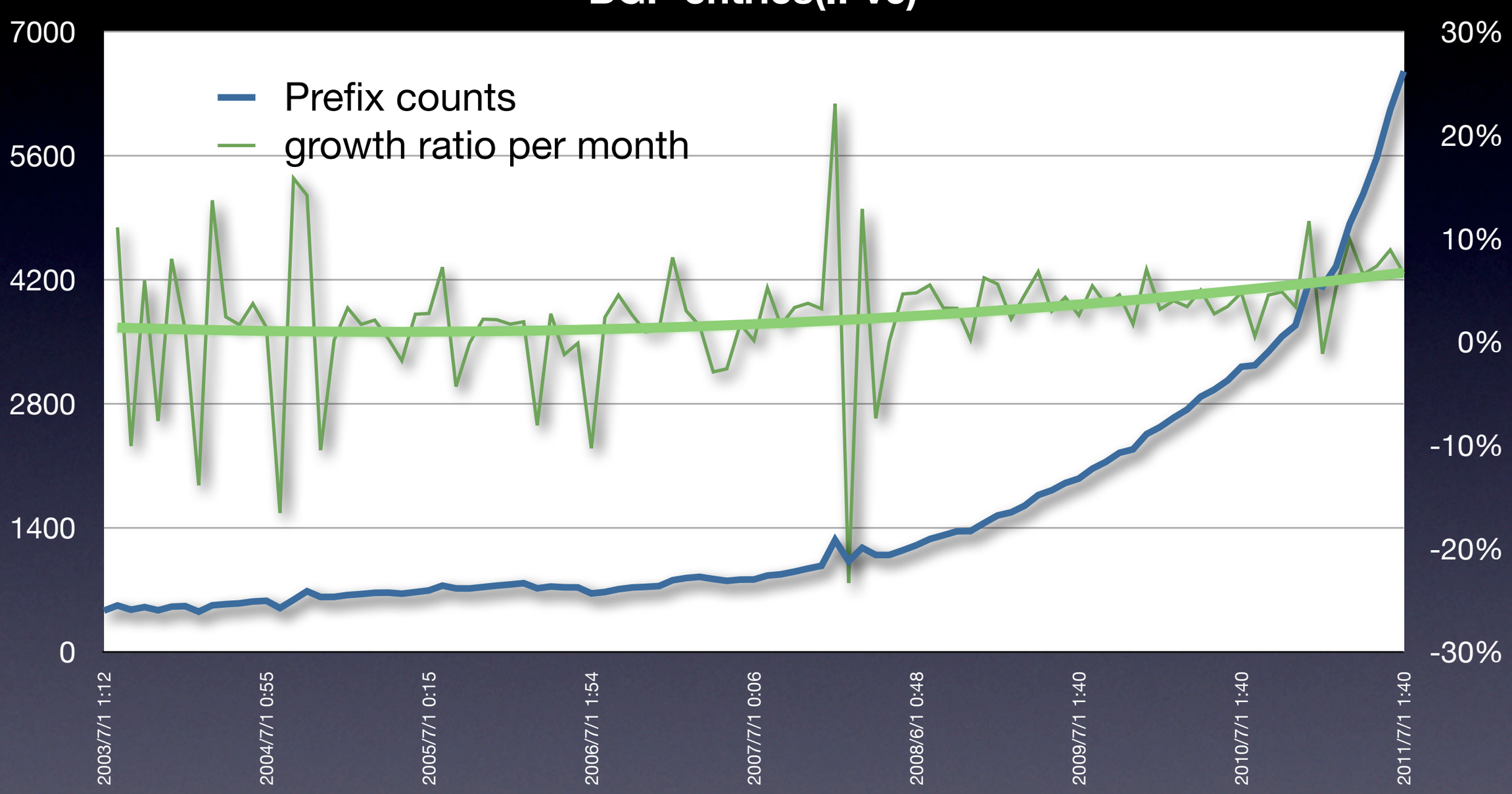


現在37000超

このままの伸びで推移するのではないか?

データ引用 http://bgp.potaroo.net

BGP経路数(IPv6)



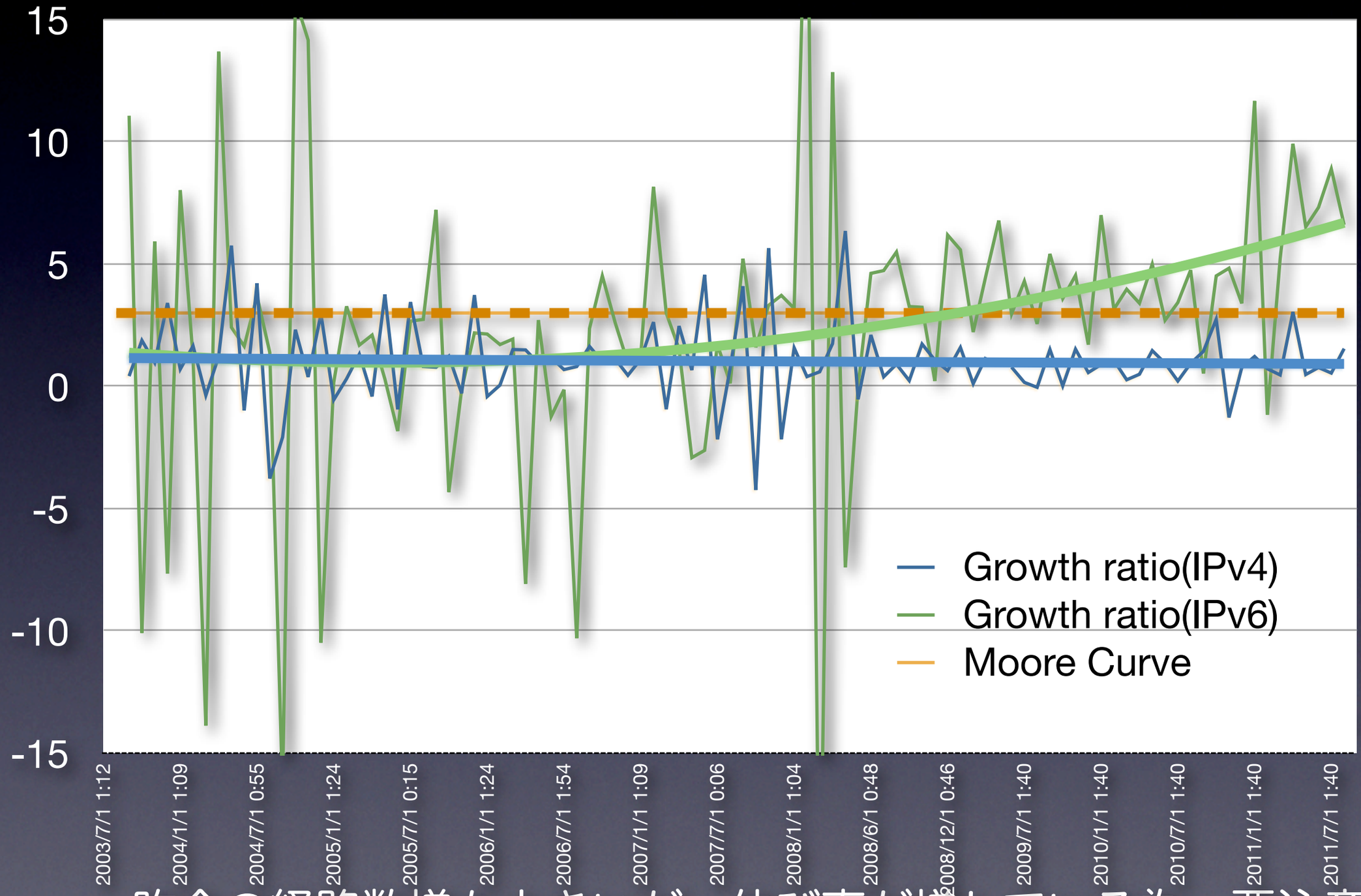
現在6400超

2008年頃から急激な立ち上がり

データ引用 http://bgp.potaroo.net

伸び率の比較

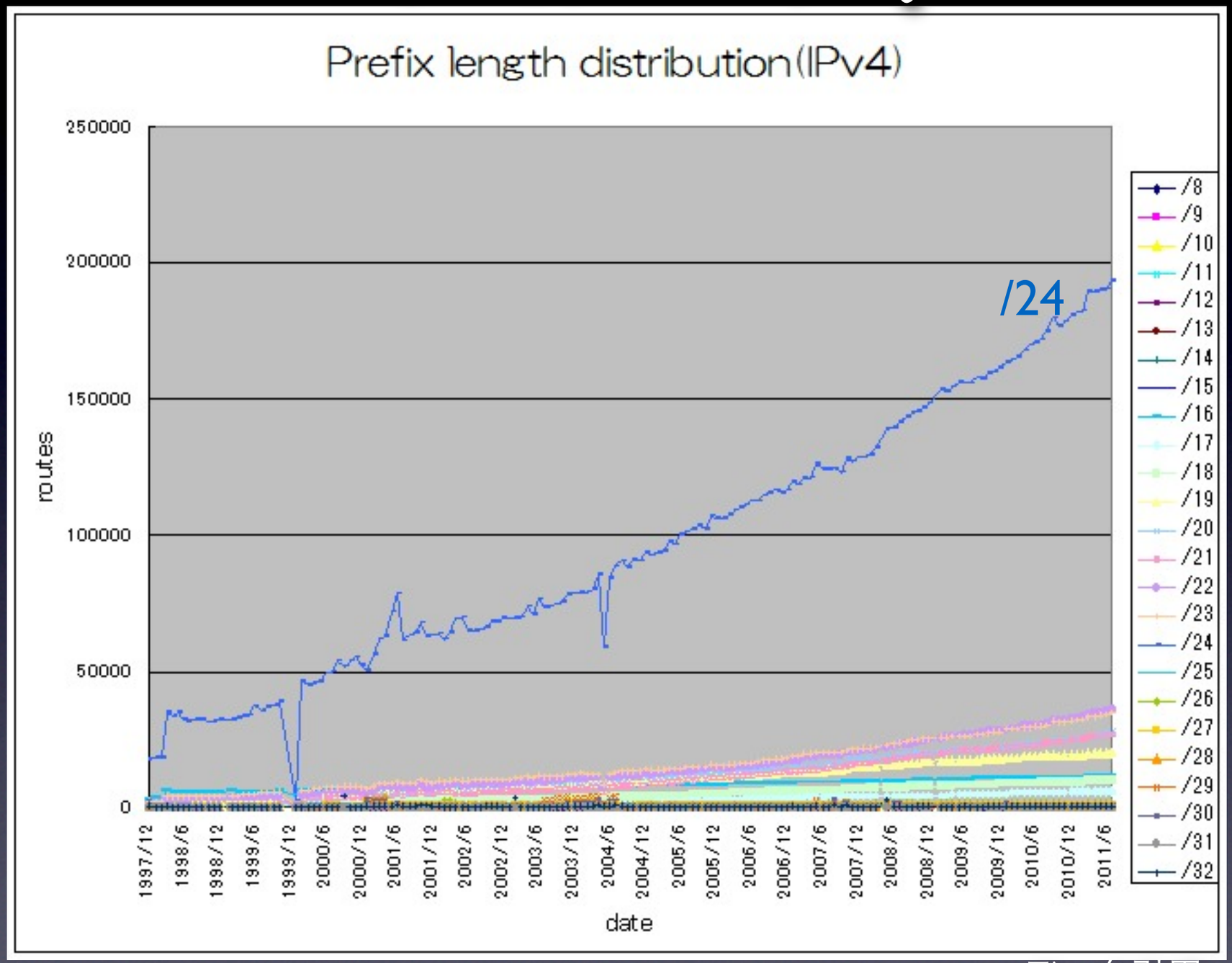
Comparison of growth ratio(IPv4/IPv6/moore curve)



昨今の経路数増も大きいが、伸び率が増している為、要注意。

IPv6はムーア曲線を超える成長

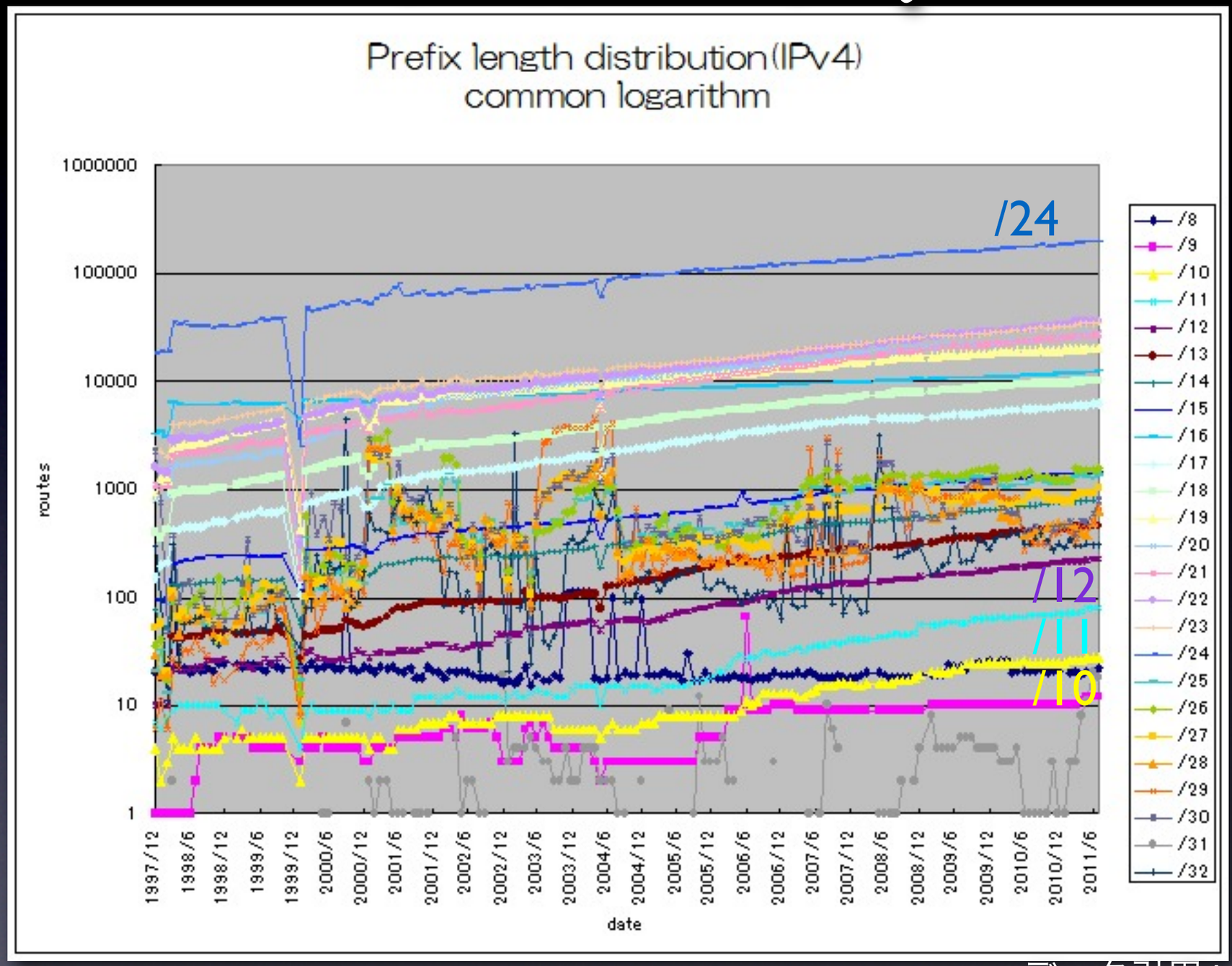
マスク長分布(IPv4)



データ引用 http://bgp.potaroo.net

/24が半分を占める

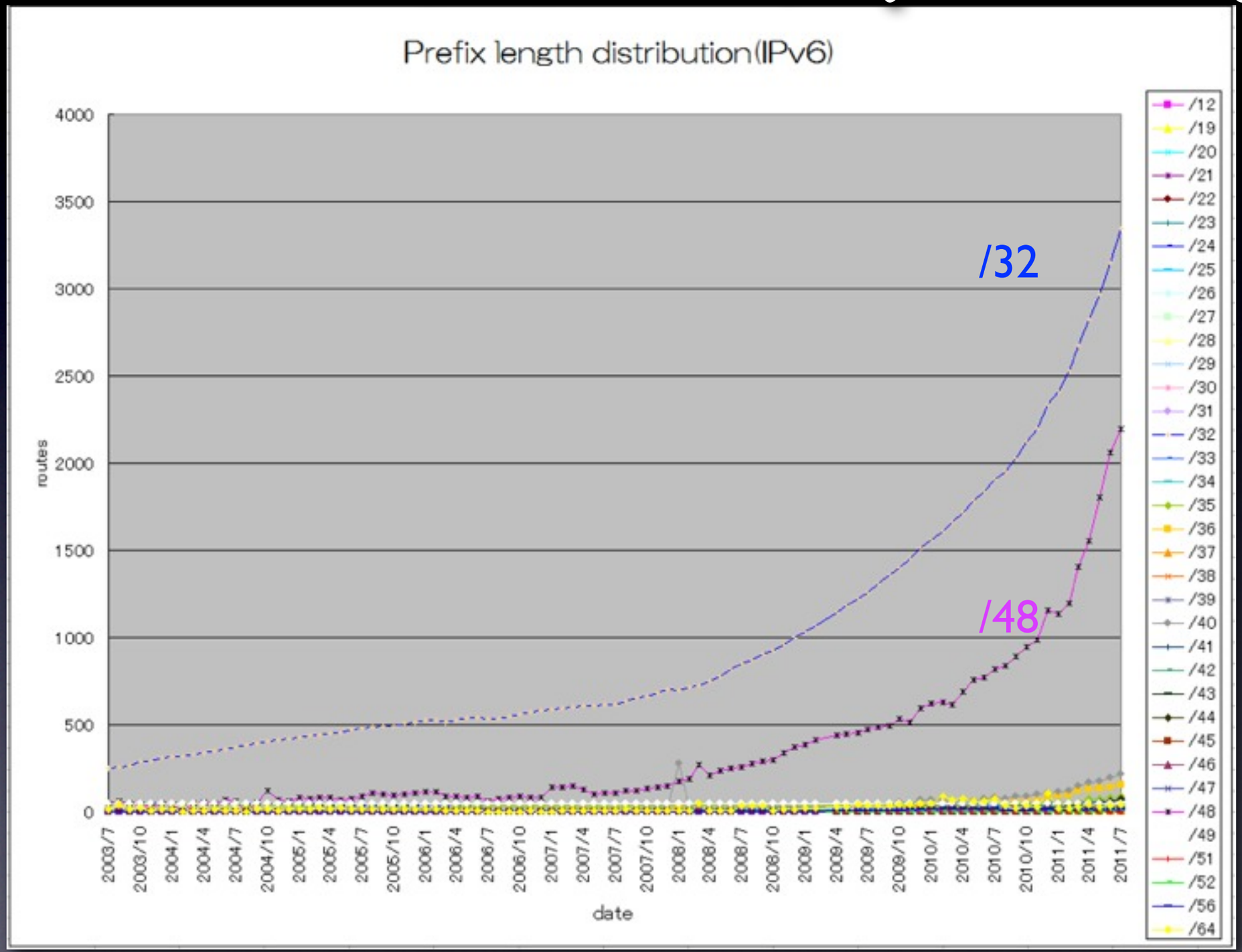
マスク長分布(IPv4)



データ引用 http://bgp.potaroo.net

/24以外にも/19～/22の経路の伸びが多い
/10、/11、/12の大きい空間も伸びている

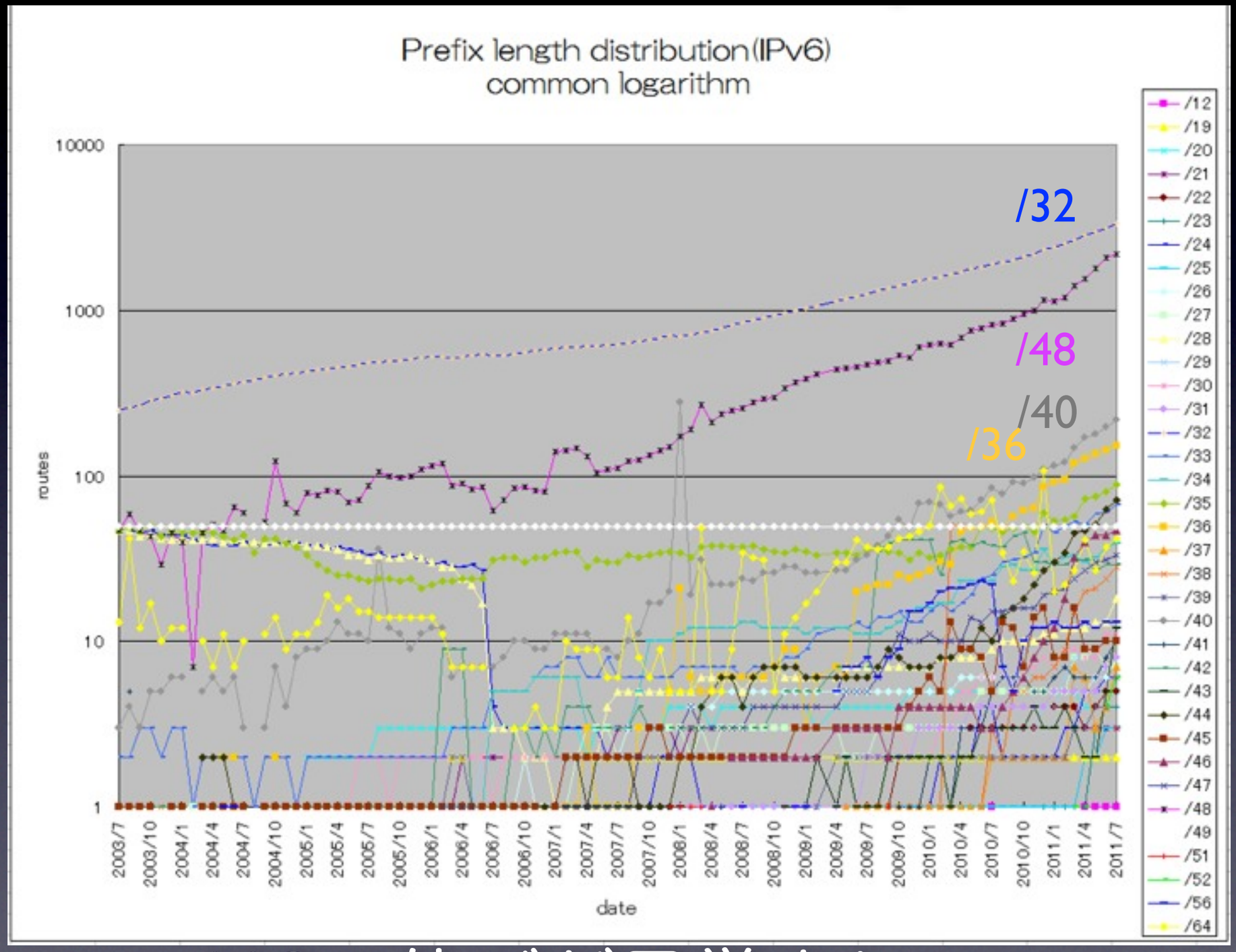
マスク長分布(IPv6)



データ引用 http://bgp.potaroo.net

/32が50%、/48が35%で占めている

マスク長分布(IPv6)



/48の伸びが目覚ましい データ引用 http://bgp.potaroo.net
/36,/40も伸びつつある

傾向まとめ

- IPv4は現状の伸びを維持
- IPv6は経路数こそ少ないものの伸び率に要注意

経路数増大

で

何が問題か？

経路が増えて困ること

1. 転送テーブル（FIB）の限界到達

→　限界に達すると、経路追加が出来なくなるなど

→　正常にパケット転送ができなくなる

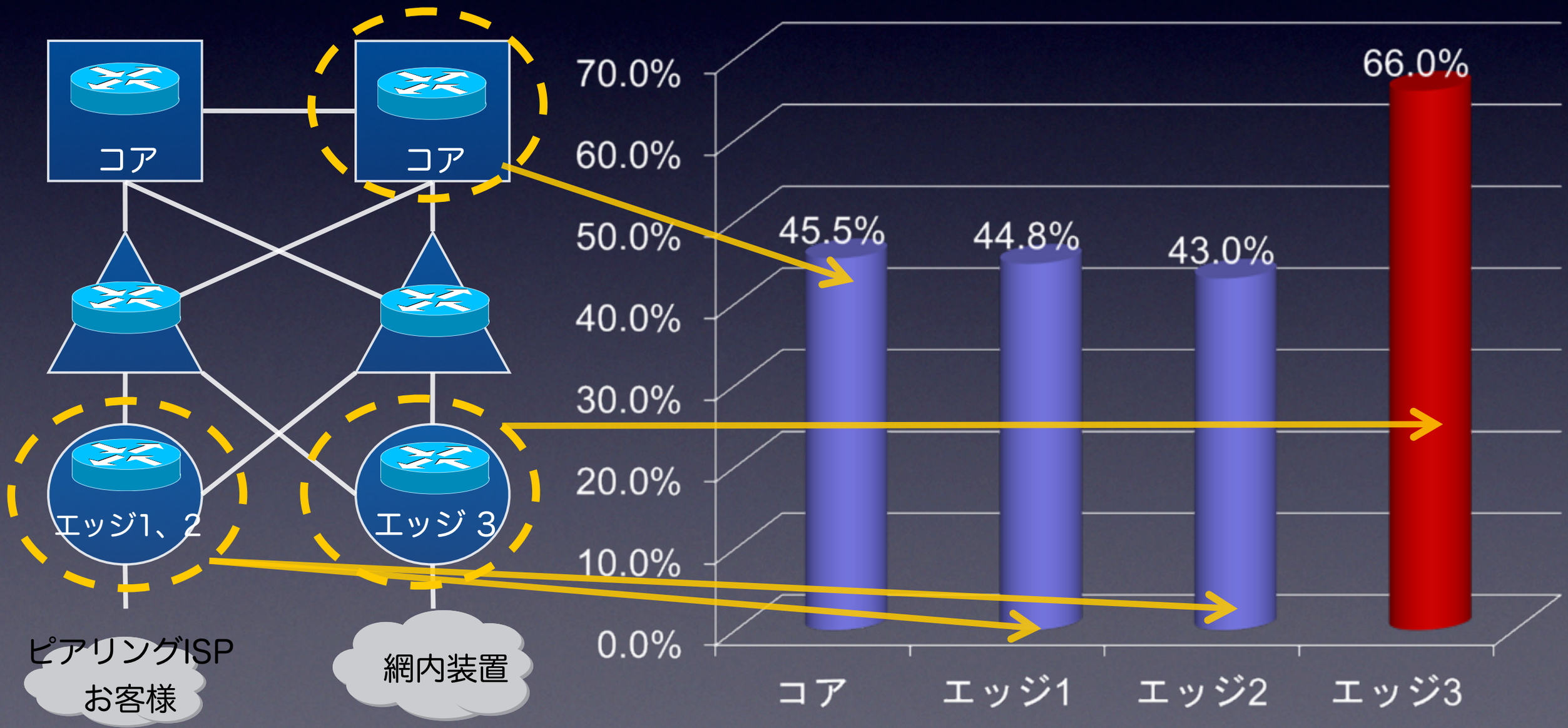
（通信障害が発生する）

2. 経路収束時間の長時間化

→　ベストパスの切り替わりに時間がかかる

FIB使用率の実際

- OCNのとあるフルルート保持ルータのFIB
  メモリ使用率



FIBの限界

- 最近のコアルータは100万経路くらい
- 各ルータのFIB容量（2009年当時）

最大収容経路数の例(IPv4のみ)

製品名	経路数	枯渇時期
Force10 EEシリーズ	26万経路	2008年中盤
ALAXALA AX2000Rシリーズ		
Force10 EFシリーズ	32万経路	2010年初頭？
Brocade 旧NI,BIシリーズ	40万経路	2011年終盤？
Brocade RX,MLXシリーズ	52万経路	2013年終盤？
Force10 EGシリーズ		
Brocade XMRシリーズ	100万経路	ずっと先？
ALAXALA AX7800Rシリーズ		

溢れたら上位機種へのリプレイス等が必要

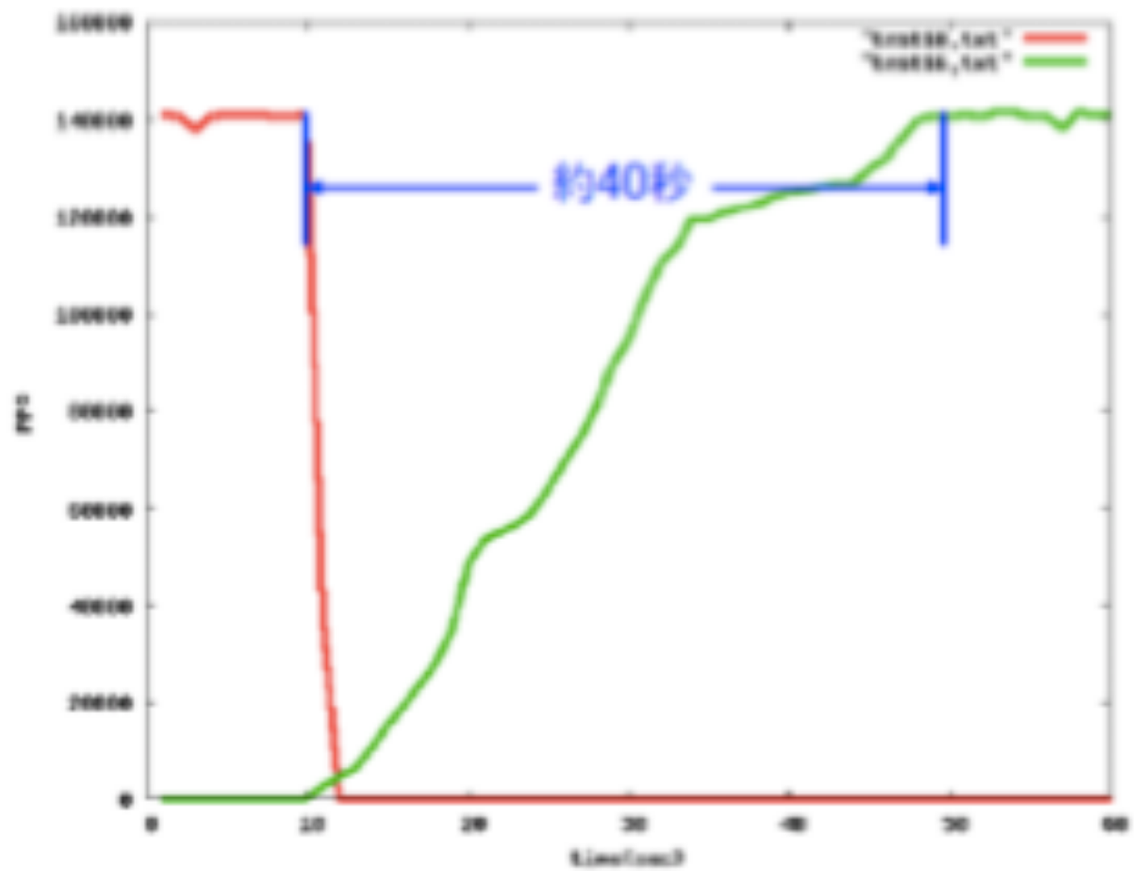
出典：IRS21　『Prefix増大とルータのリソース問題』

さくらインターネット研究所　大久保修一さん

コンバージェンスの実際

経路収束の遅延

- 実験結果



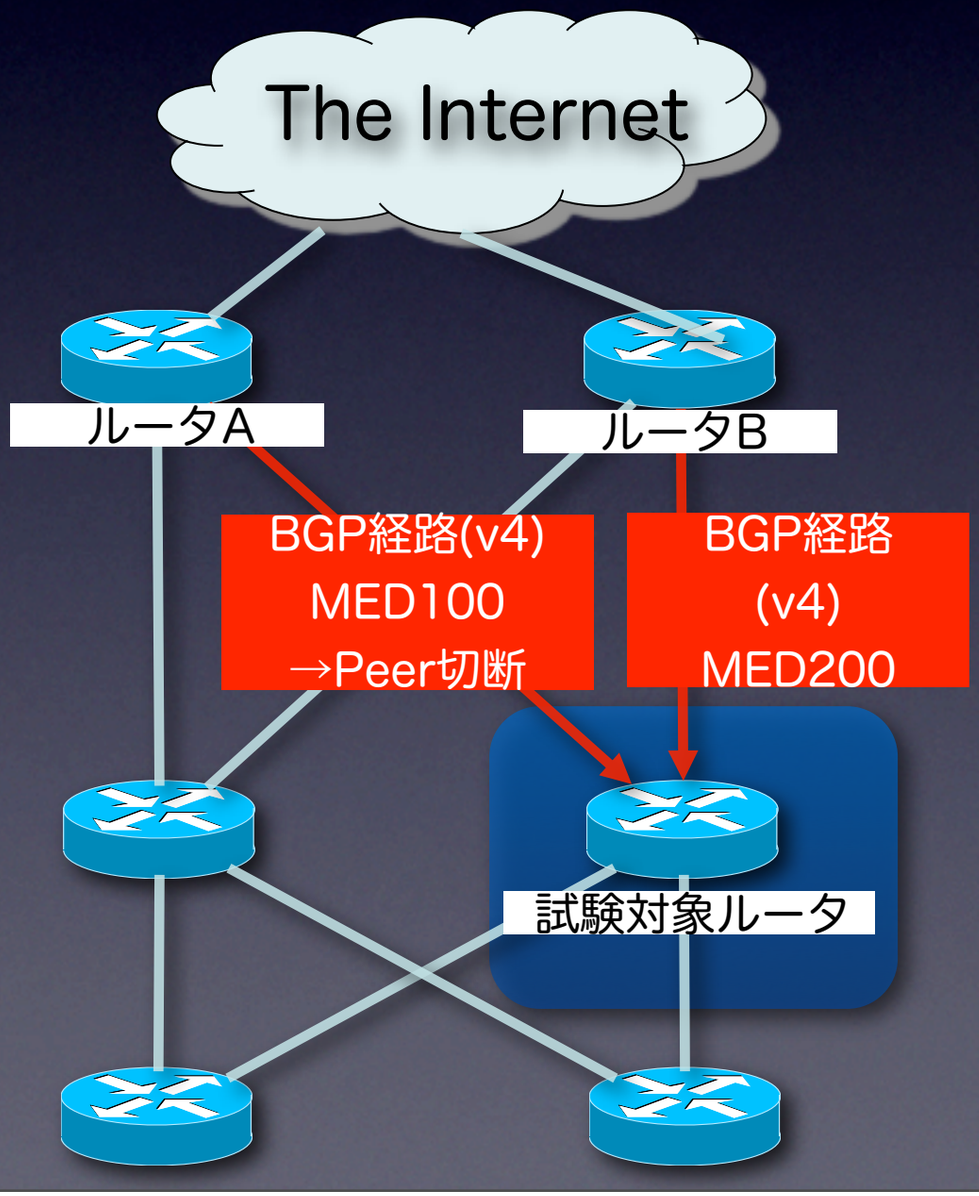
29万経路の切り替りに約40秒

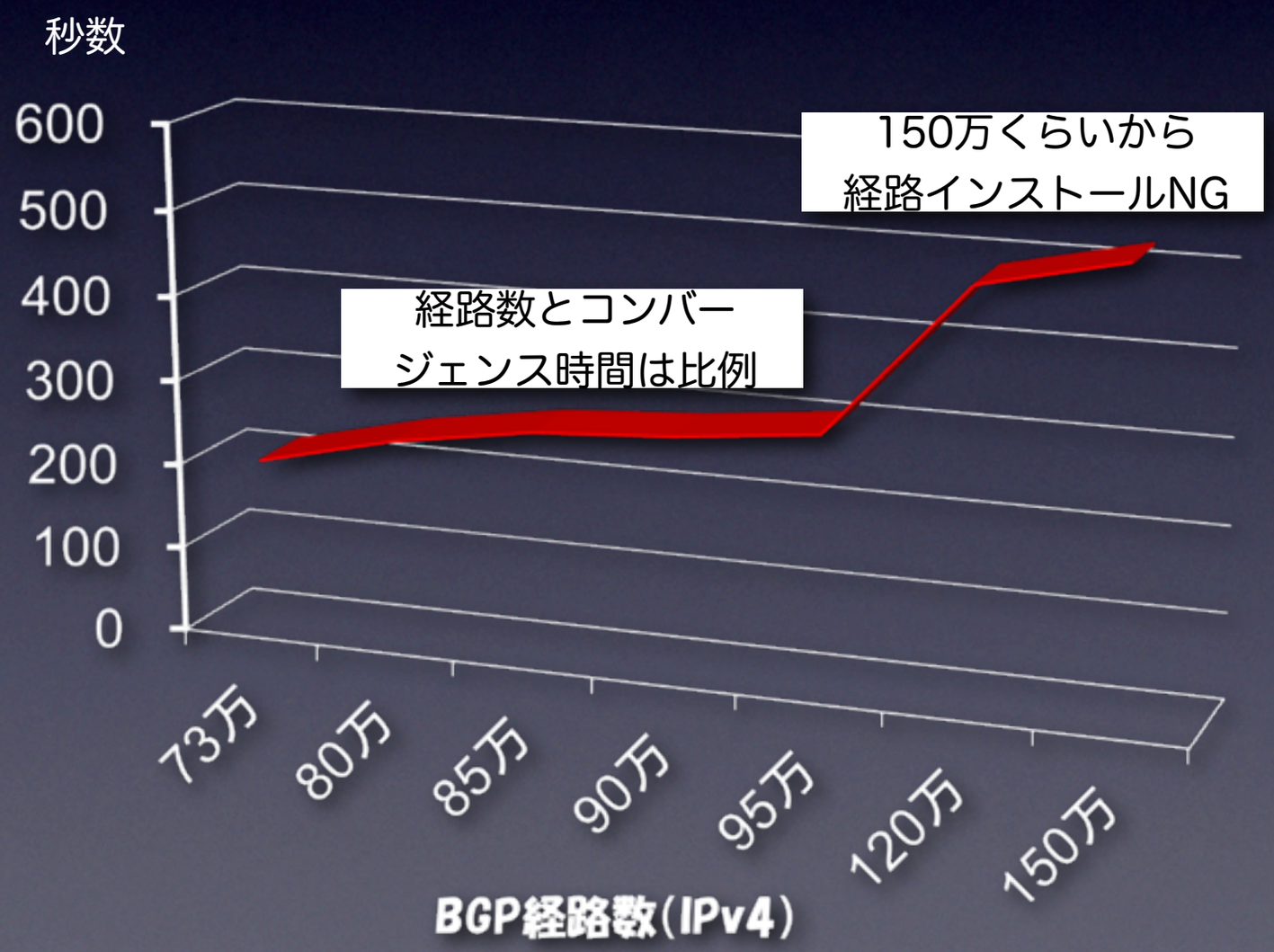
出典：IRS21　『Prefix増大とルータのリソース問題』

さくらインターネット研究所　大久保修一さん

コンバージェンスの実際

- OCNのとあるフルルート保持ルータのコンバージェンス時間





課題への対策

1. 転送テーブル（FIB）の限界到達

- 保持経路を削減する
- より高スペックのルータに更改する

2. 経路収束時間の長時間化

- 短縮技術を使う
- 保持経路を削減する
- より高スペックのルータに更改する

フルルートの必要性

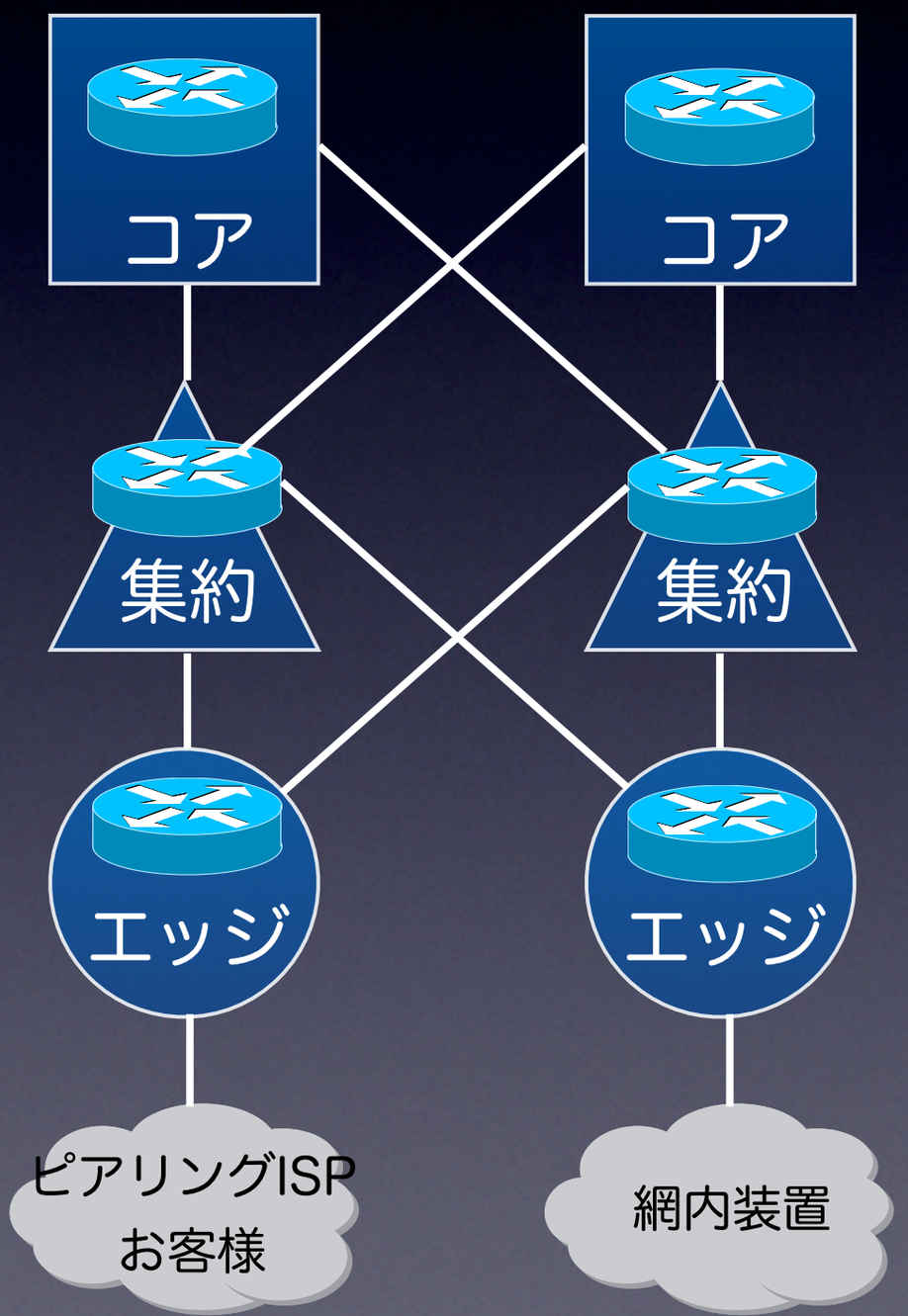
- デフォルトルートのみで良い構成もある
  - 配下にBGPフルルートのお客様が居ない、など
  - トラヒックはコアに集まりやすい
- フルルートを持ちたい場合
  - BGPフルルートのお客様
  - より最適なルーティング（closest exit）

で、どうしているか…

- フルルート数増加
- メモリ使用率上昇（グラフとにらめっこ☹）
- やりたくないけど経路削減（トラブルの元）
- EOL等を契機に高スペックルータへリプレイス＆再フルルート化（やった☻）
- OCNでは、メモリ容量が逼迫したルータは更改や経路削減をしています。

やりくり

通常時

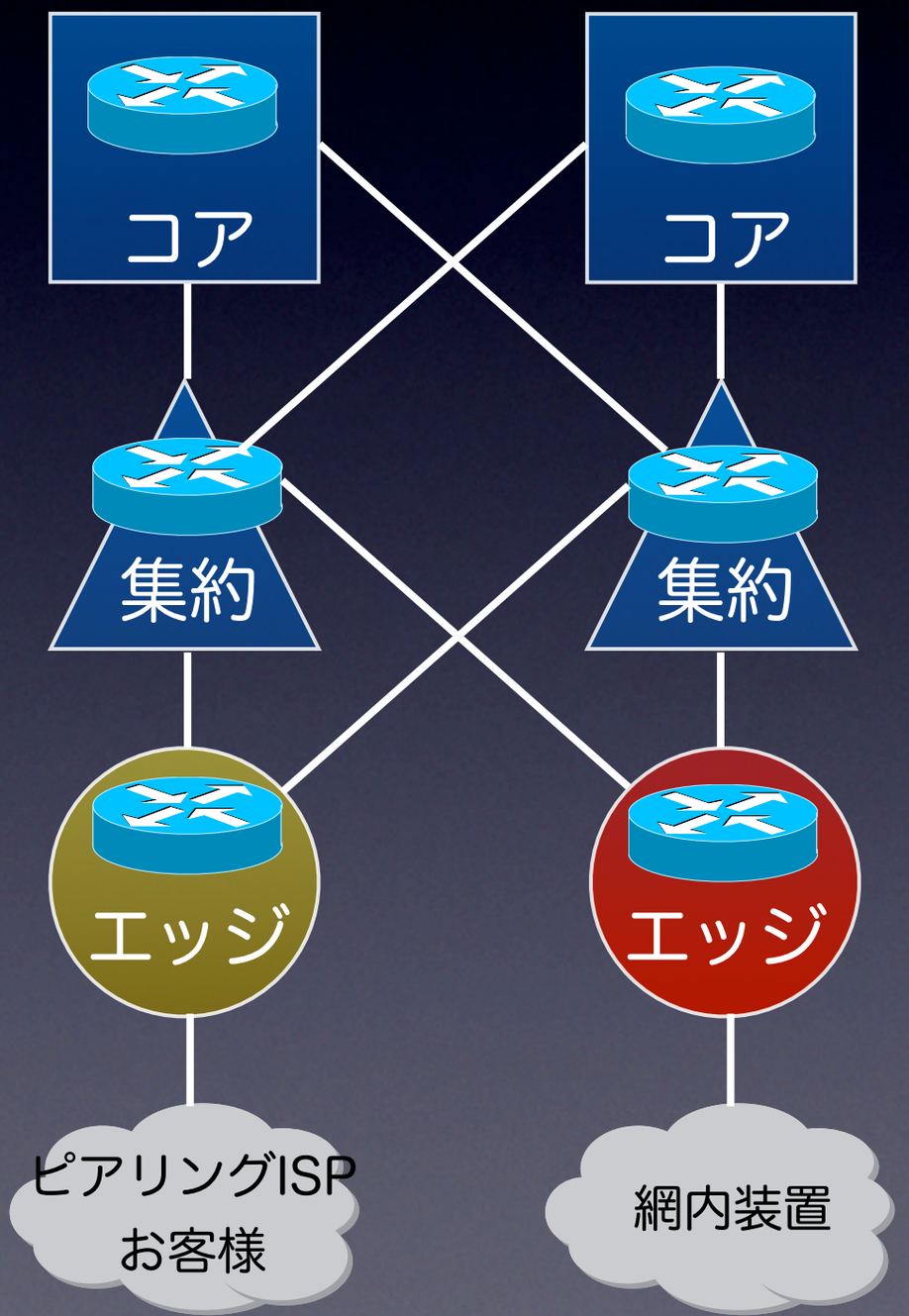


メモリ使用率

高
中
低

やりくり

経路増

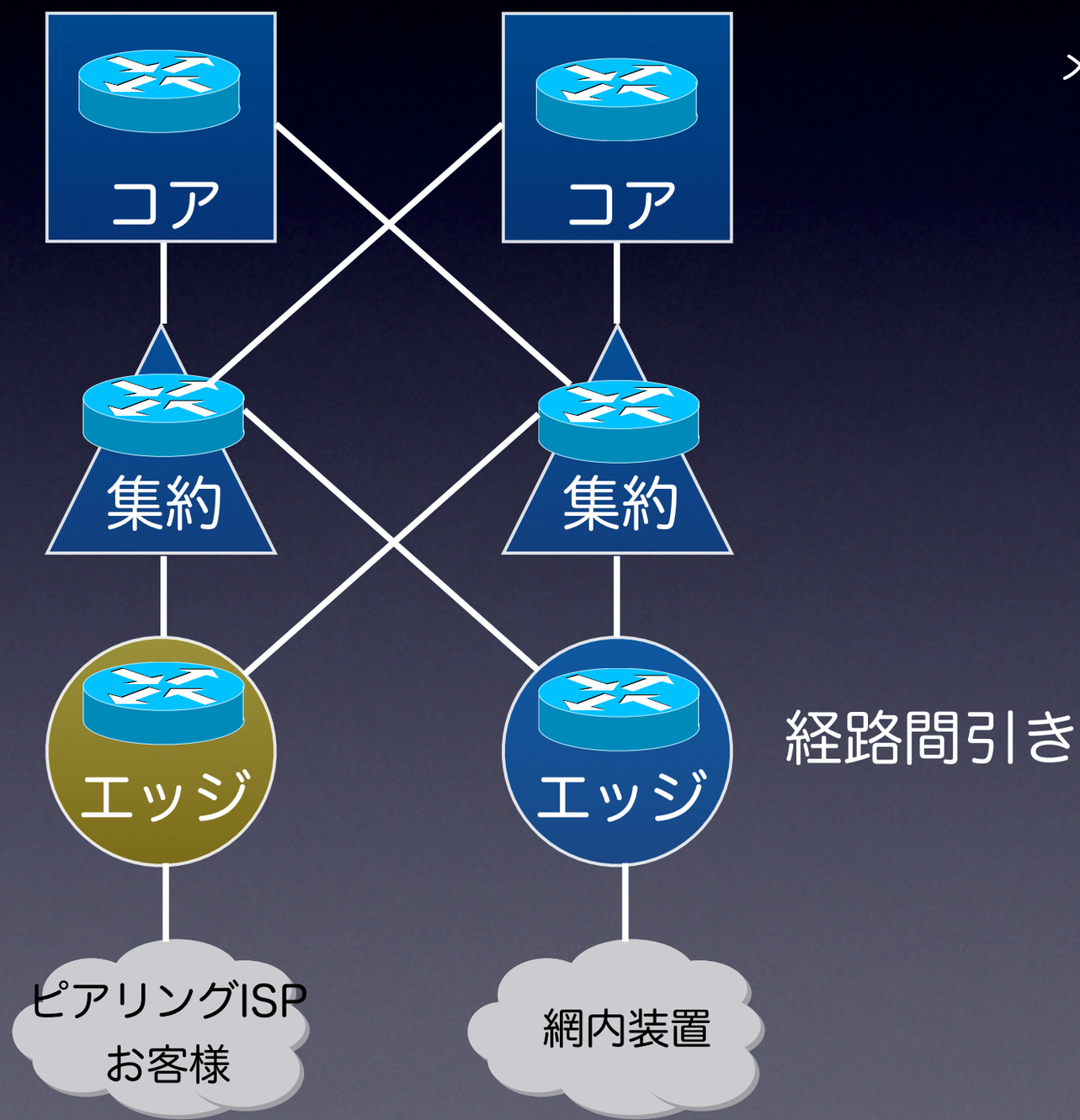


メモリ使用率

高
中
低

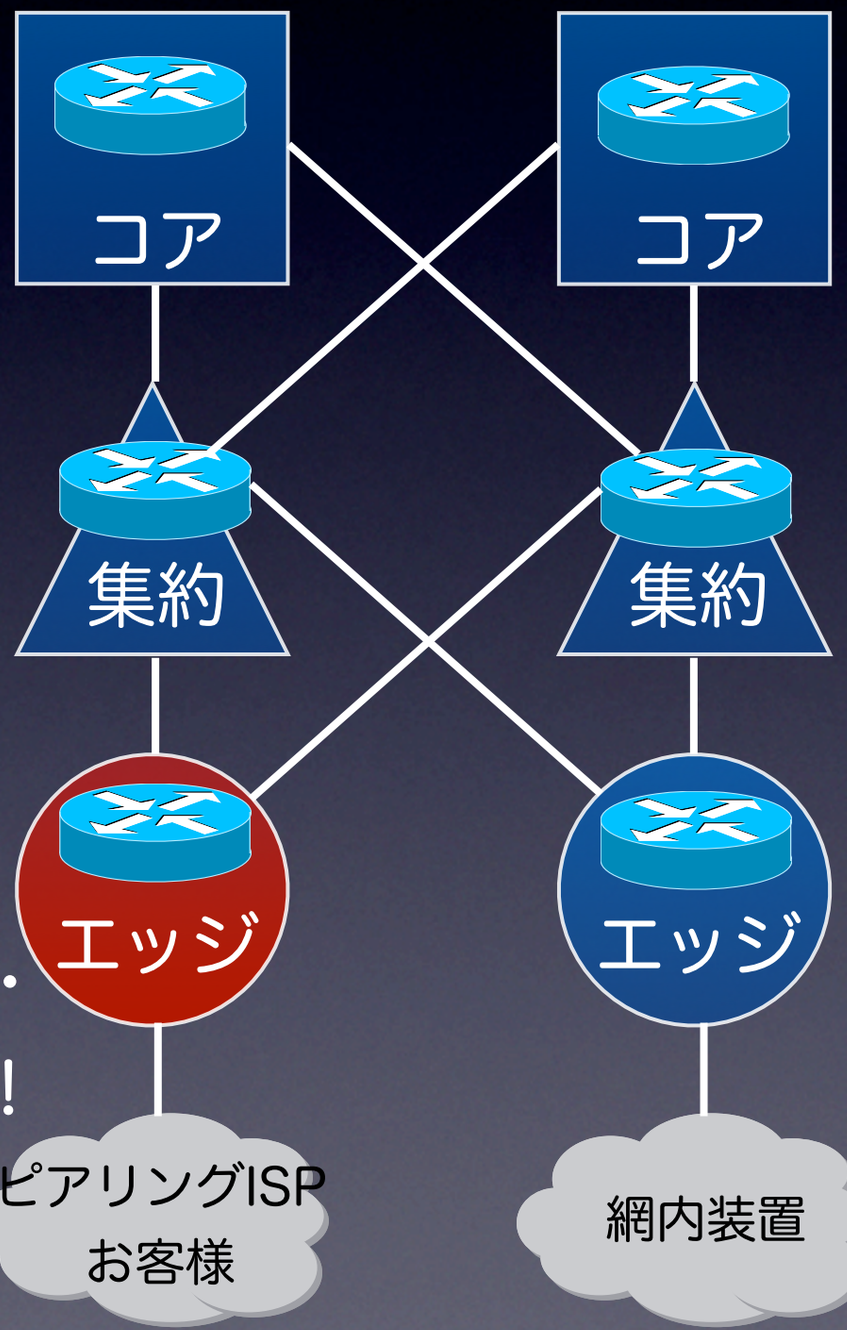
やりくり

経路間引き完了
とりあえず延命



やりくり

でも・・・



メモリ使用率

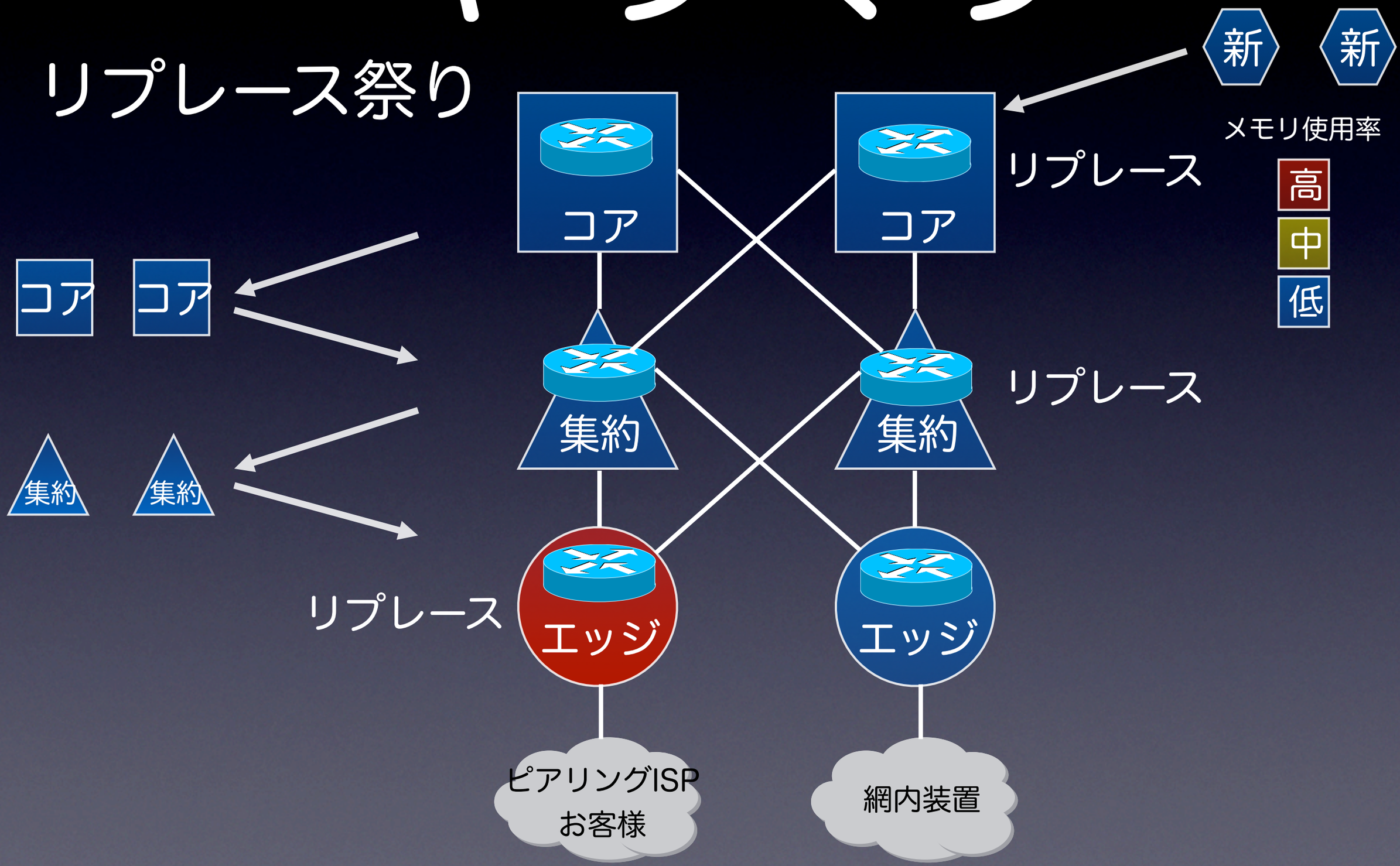
高
中
低

今度はフルルートが
必要なエッジルータが・・・
リプレースするしかない！！

経路間引き

やりくり

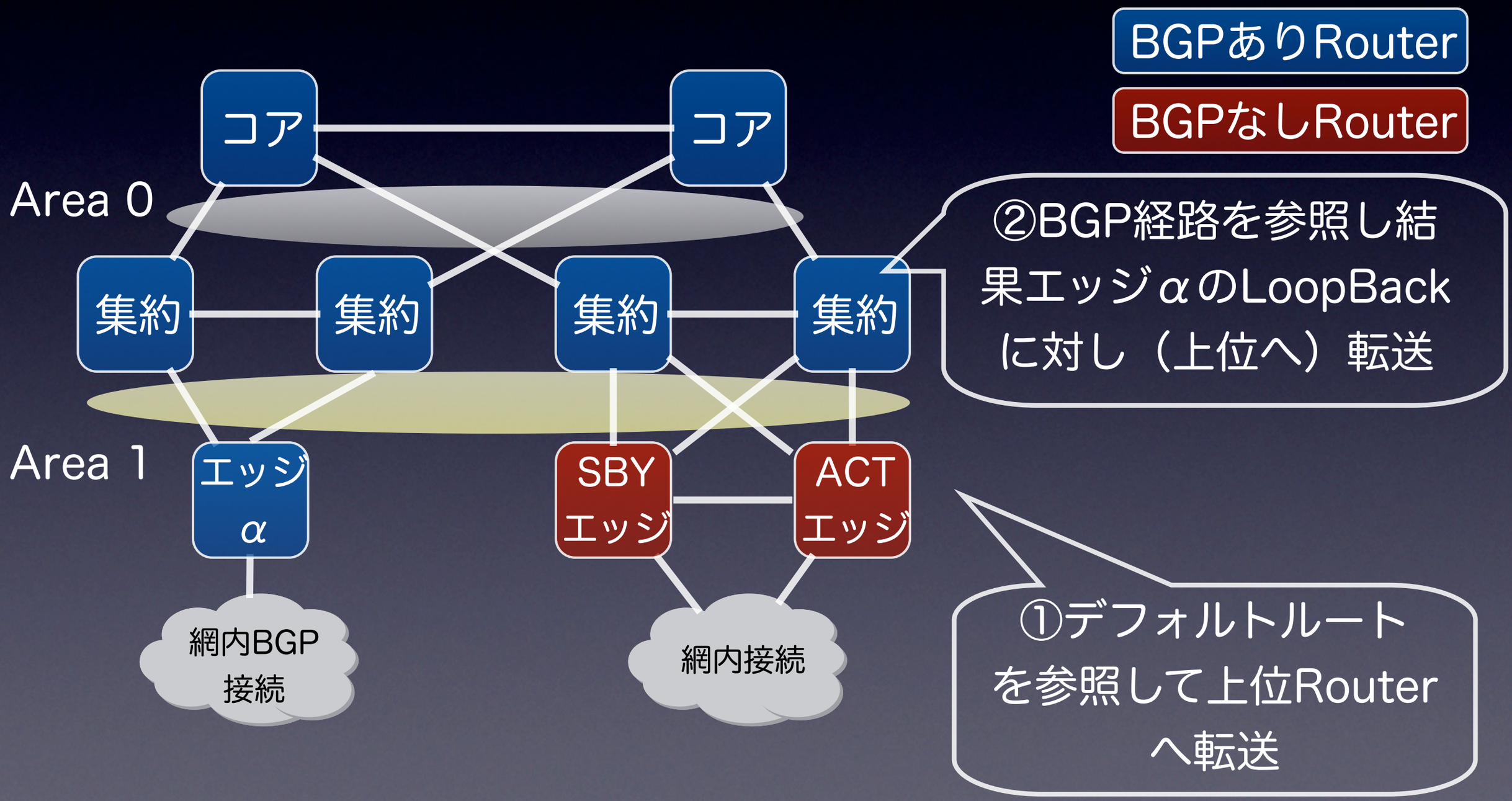
リプレース祭り



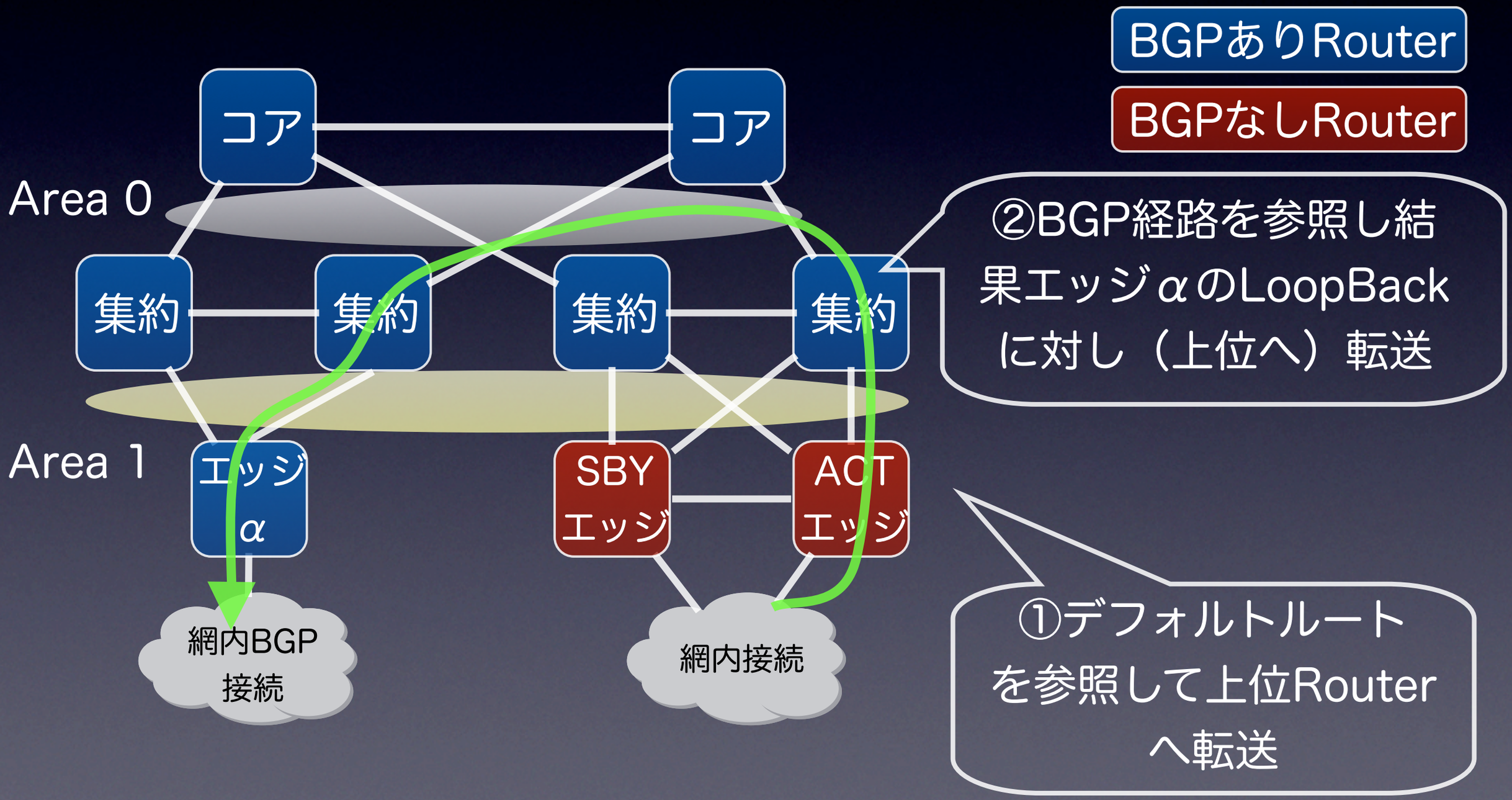
経路削減

- 経路削減区間でRoutingの不一致が発生する場合も
- 下手すると網内でピンポン発生（泣）
- しかも忘れた頃にトラブルの引き金となる

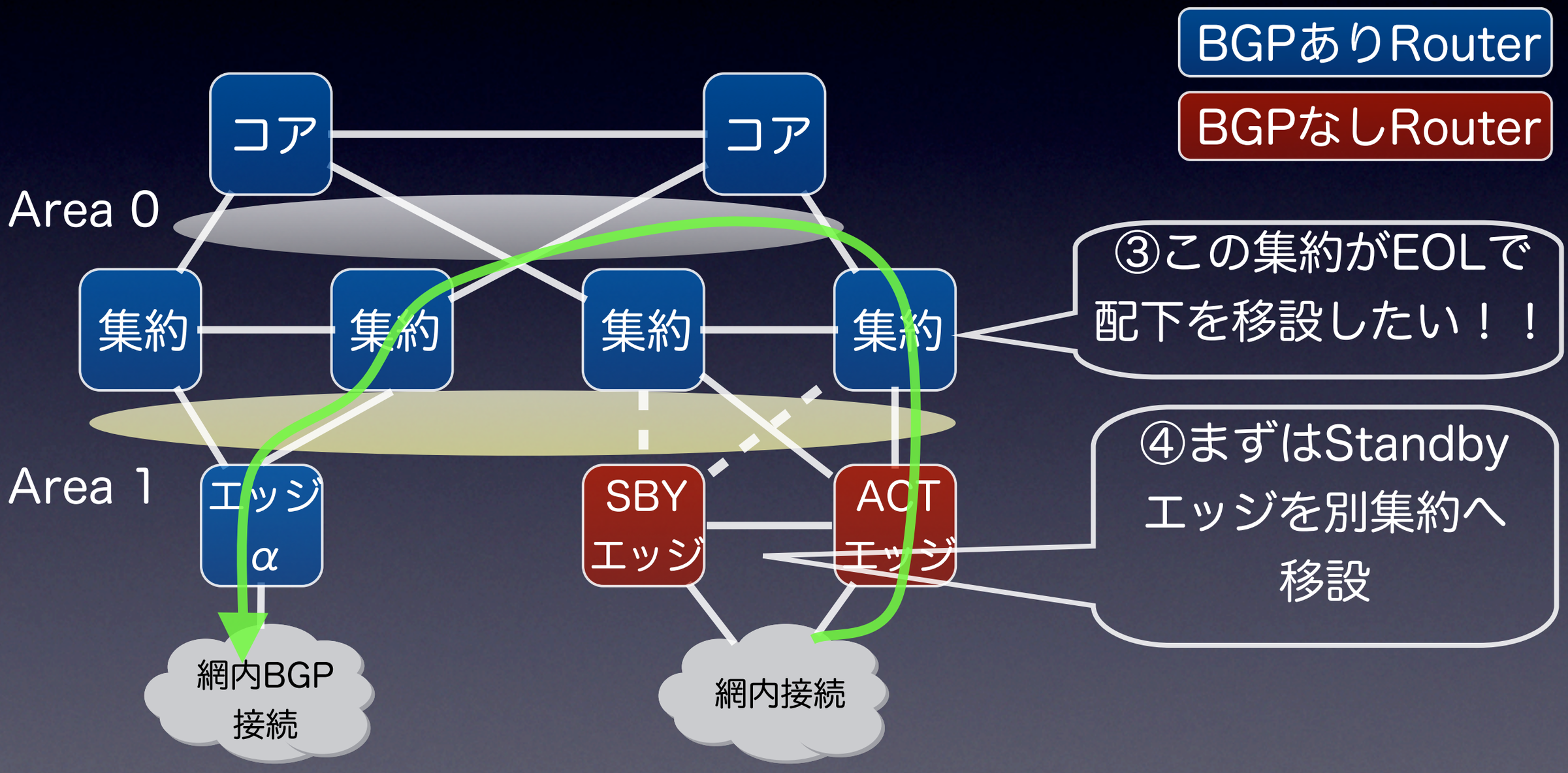
例えばこんな環境で・・・


BGPありRouter
BGPなしRouter
コア
コア
Area 0
集約
集約
集約
集約
②BGP経路を参照し結果エッジαのLoopBackに対し（上位へ）転送
Area 1
エッジα
SBYエッジ
ACTエッジ
網内BGP接続
網内接続
①デフォルトルートを参照して上位Routerへ転送

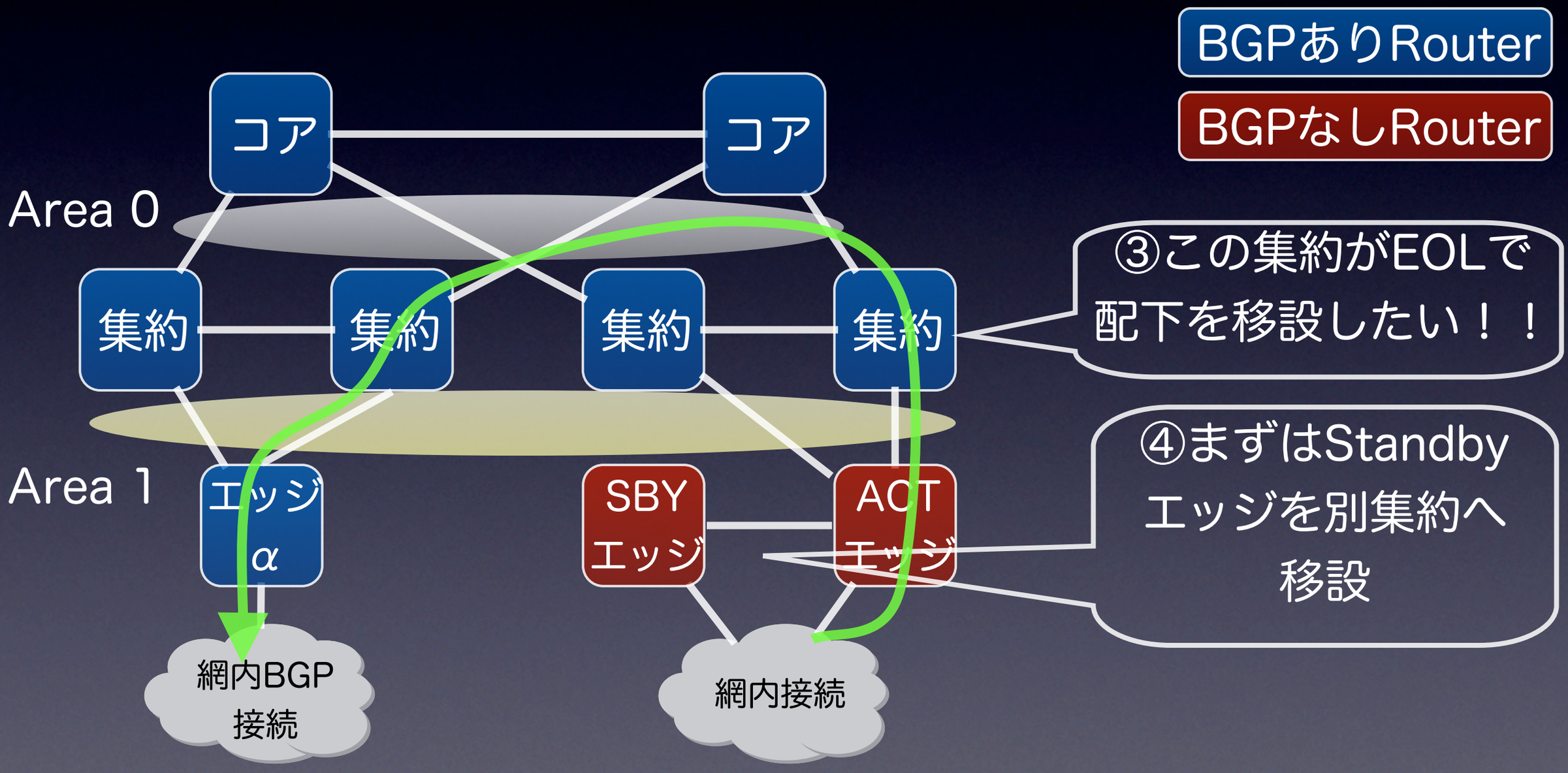
例えばこんな環境で・・・


BGPありRouter
BGPなしRouter
コア
コア
Area 0
集約
集約
集約
集約
Area 1
エッジα
SBYエッジ
ACTエッジ
網内BGP接続
網内接続
②BGP経路を参照し結果エッジαのLoopBackに対し（上位へ）転送
①デフォルトルートを参照して上位Routerへ転送

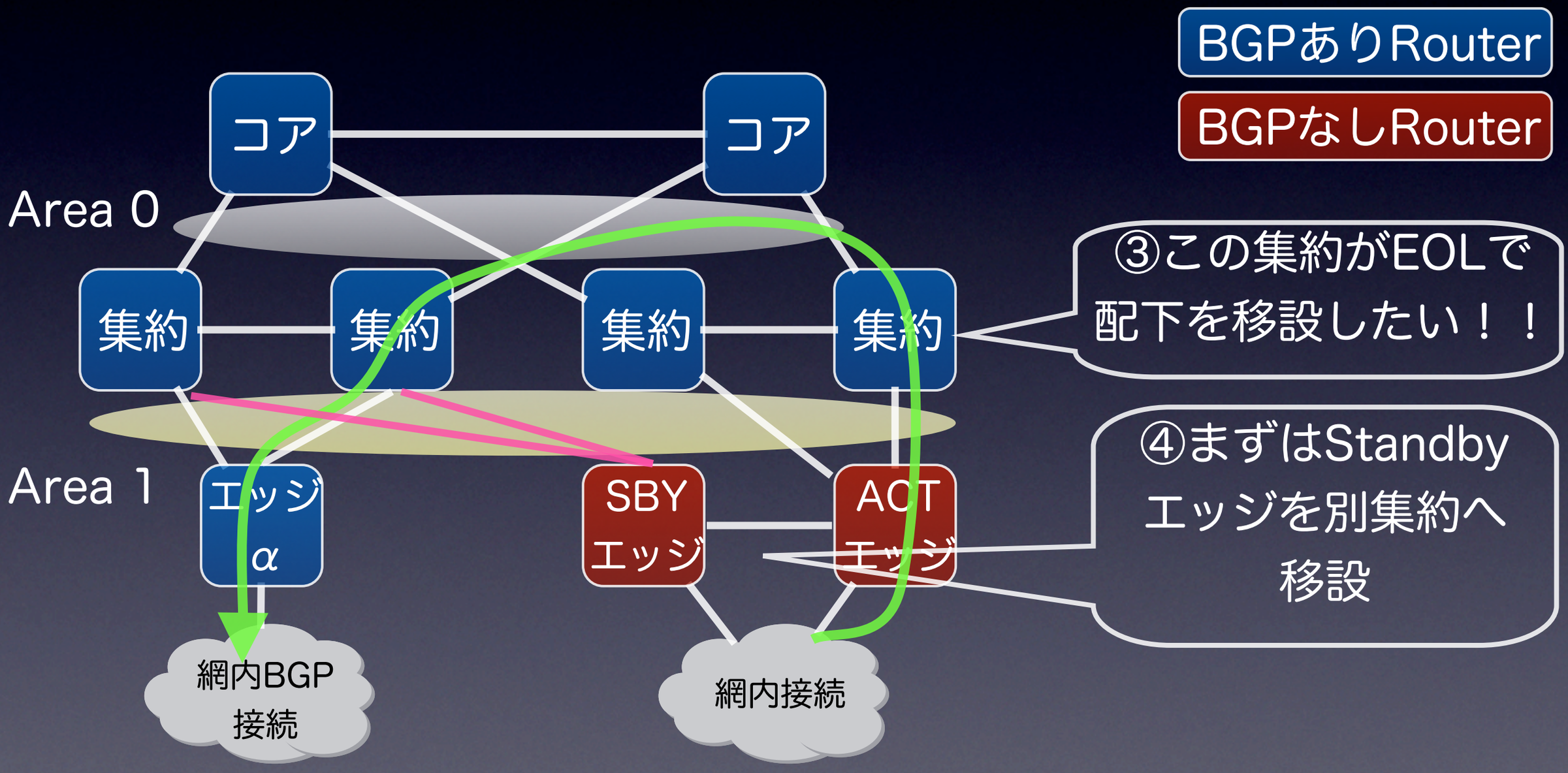
集約ルータ マイグレーション



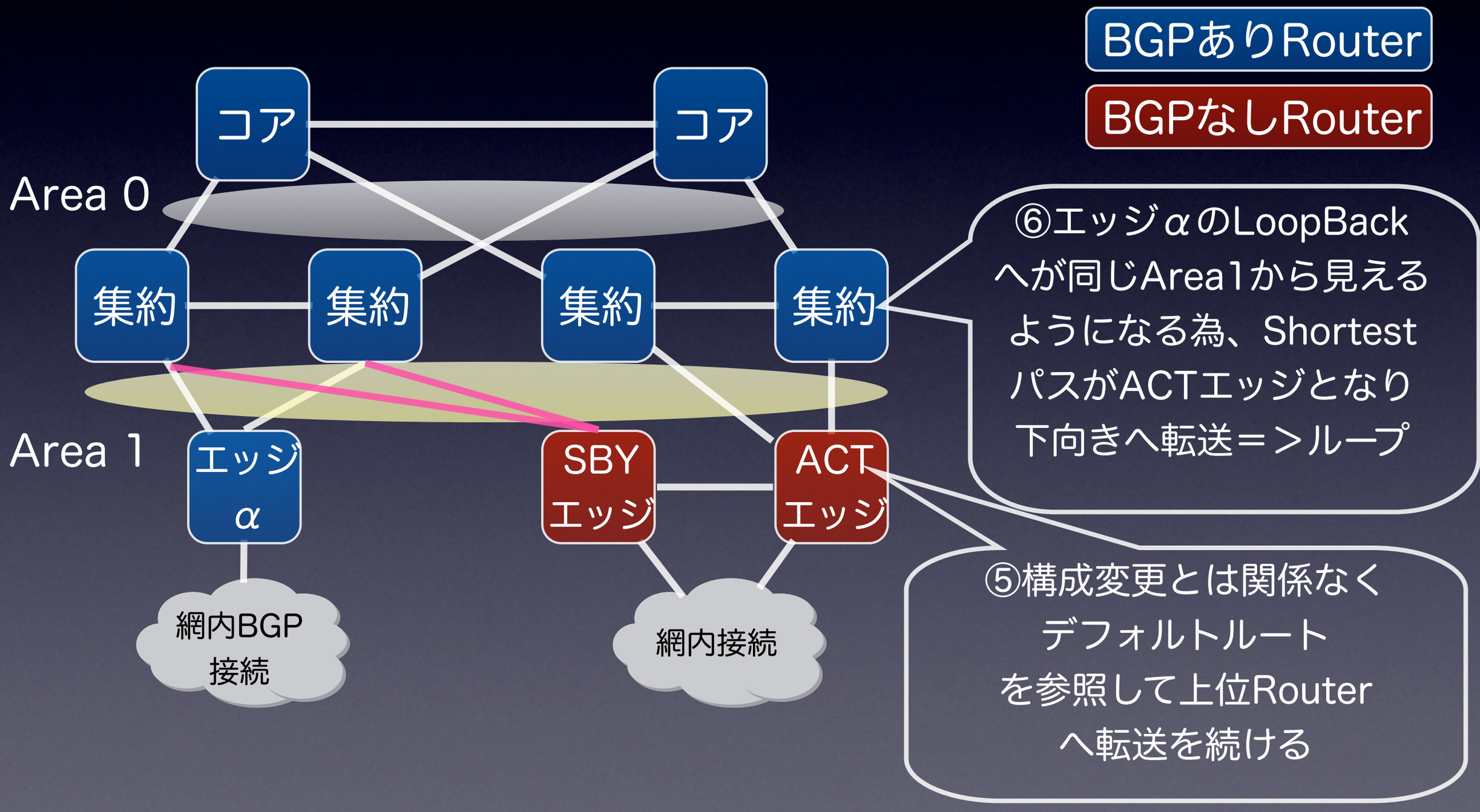
集約ルータ マイグレーション



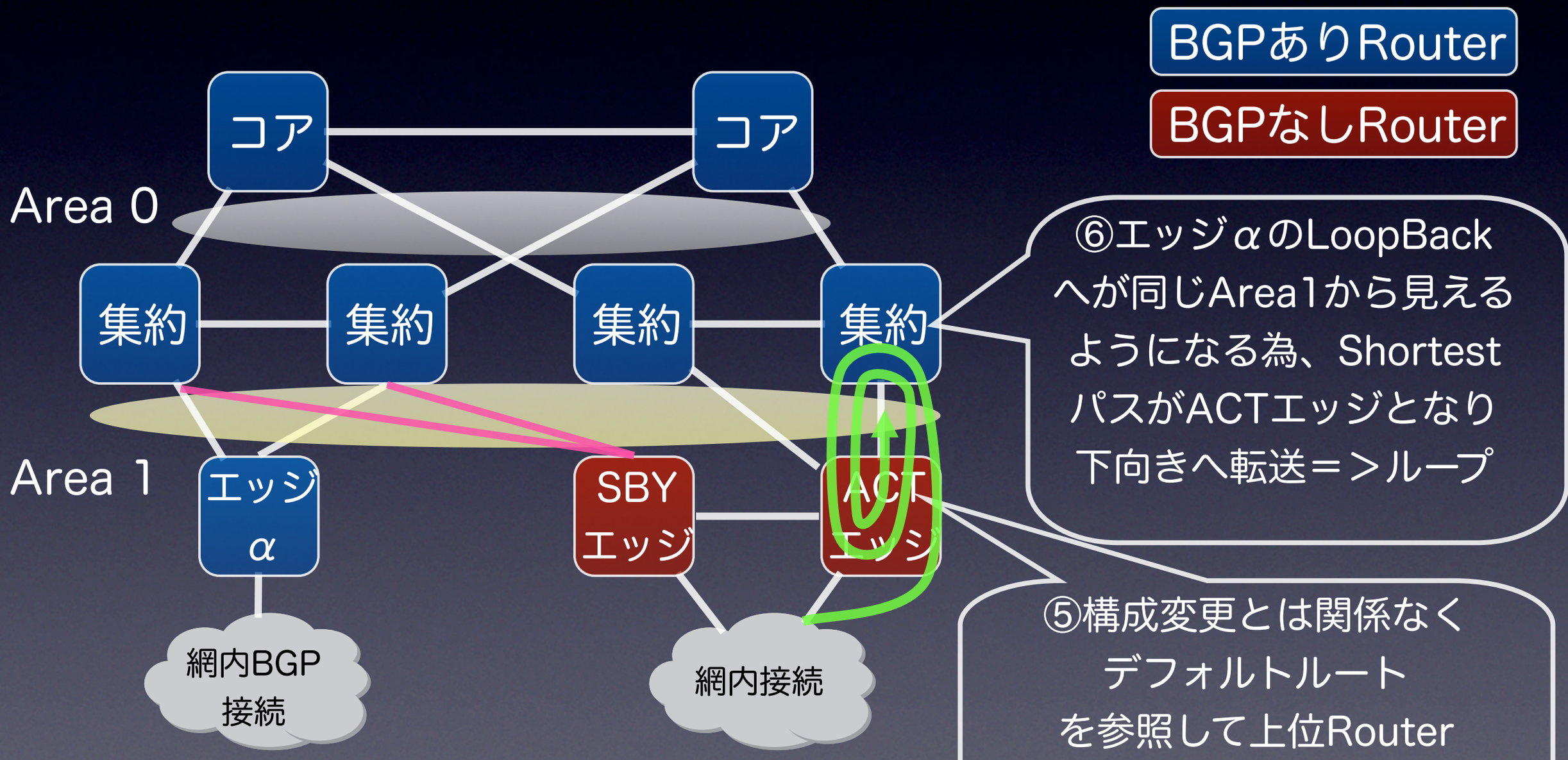
集約ルータ マイグレーション



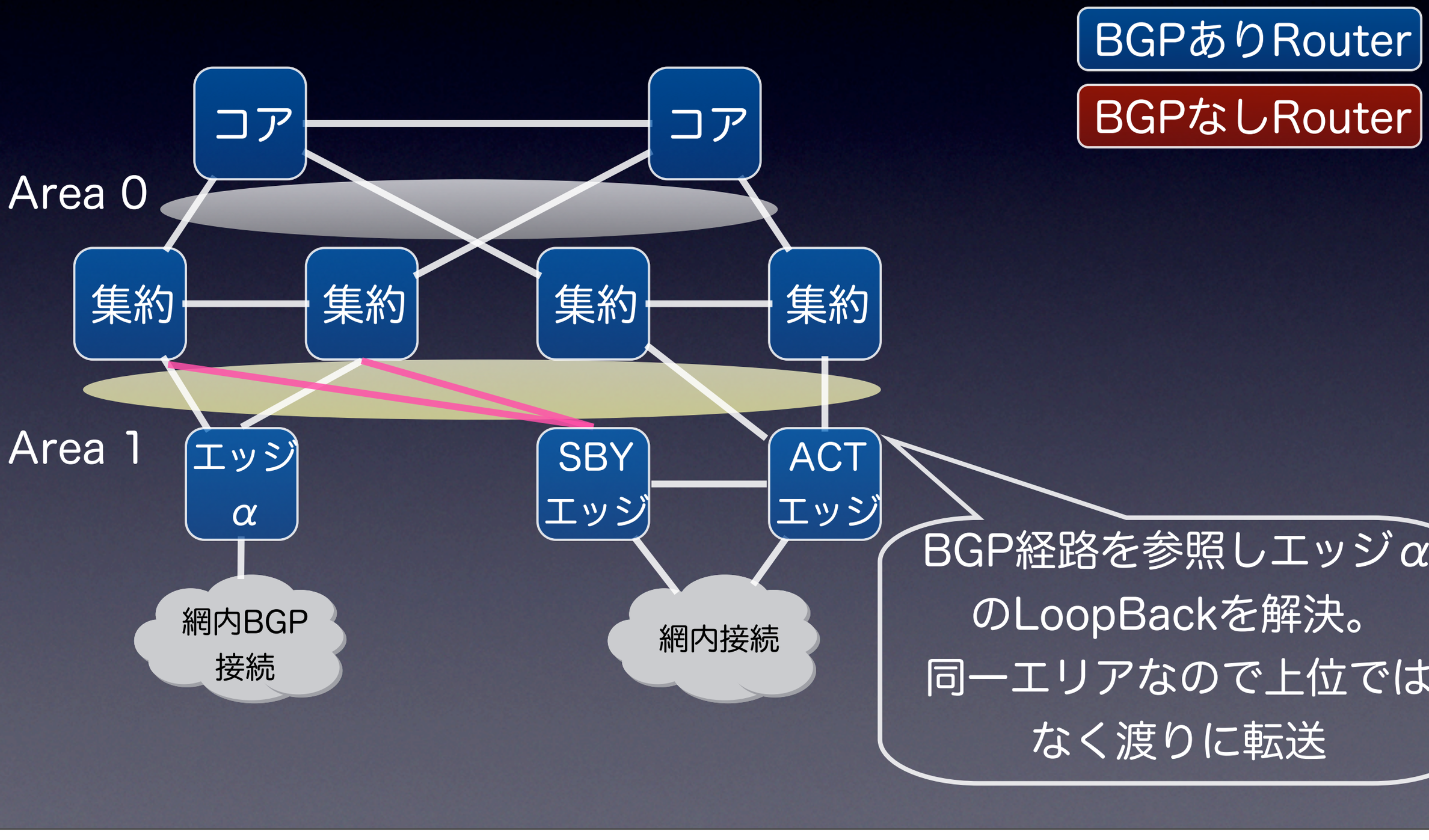
するとピンポンが発生！！


BGPありRouter
BGPなしRouter
コア
コア
Area 0
集約
集約
集約
集約
Area 1
エッジ
α
SBY
エッジ
ACT
エッジ
網内BGP
接続
網内接続
⑥エッジαのLoopBack
へが同じArea1から見える
ようになる為、Shortest
パスがACTエッジとなり
下向きへ転送＝＞ループ
⑤構成変更とは関係なく
デフォルトルート
を参照して上位Router
へ転送を続ける

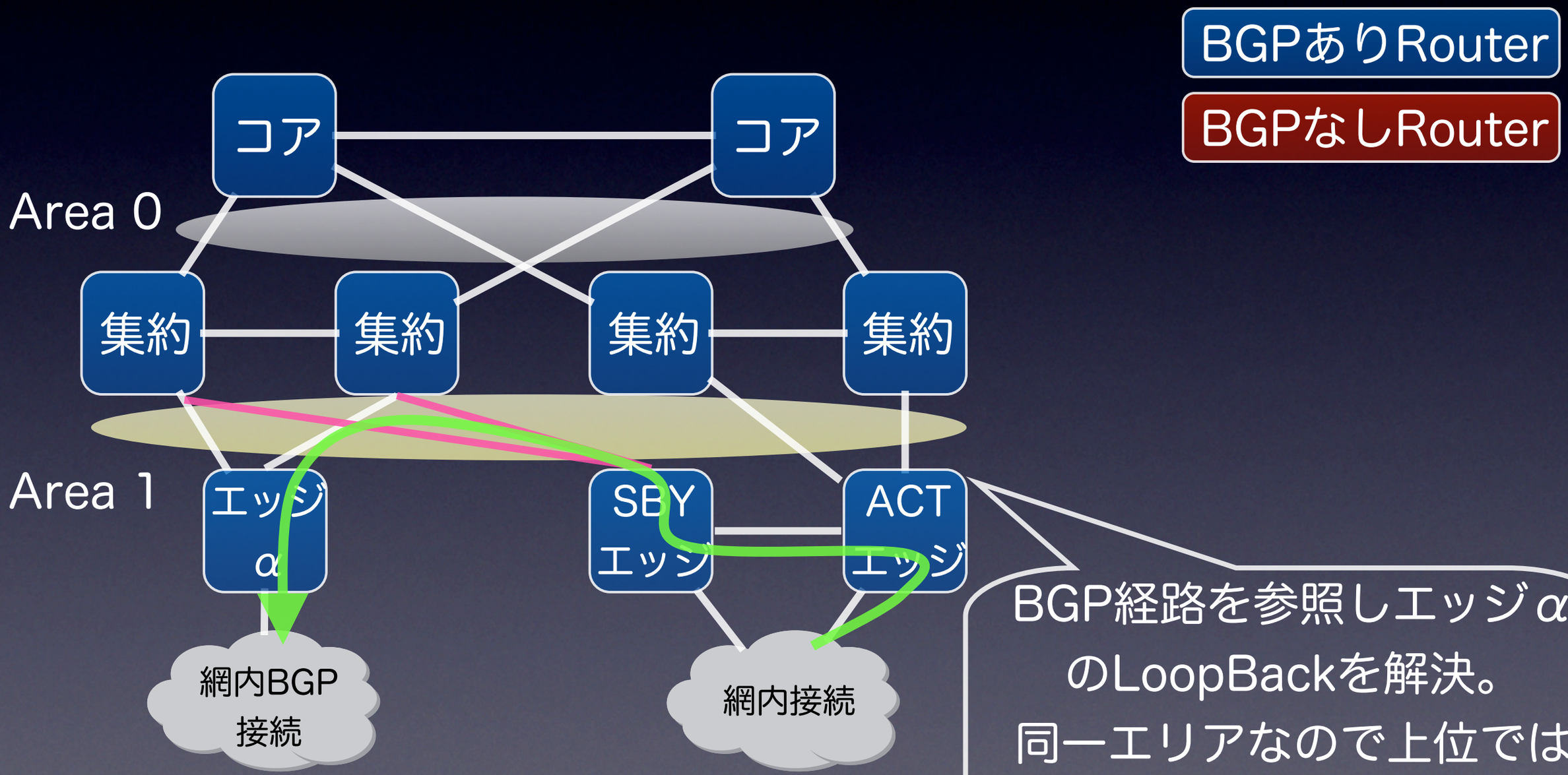
するとピンポンが発生！！


BGPありRouter
BGPなしRouter
コア
コア
Area 0
集約
集約
集約
集約
Area 1
エッジ
α
SBY
エッジ
ACT
エッジ
網内BGP
接続
網内接続
⑥エッジαのLoopBack
へが同じArea1から見える
ようになる為、Shortest
パスがACTエッジとなり
下向きへ転送＝＞ループ
⑤構成変更とは関係なく
デフォルトルート
を参照して上位Router
へ転送を続ける

もしBGP経路を持っていたら

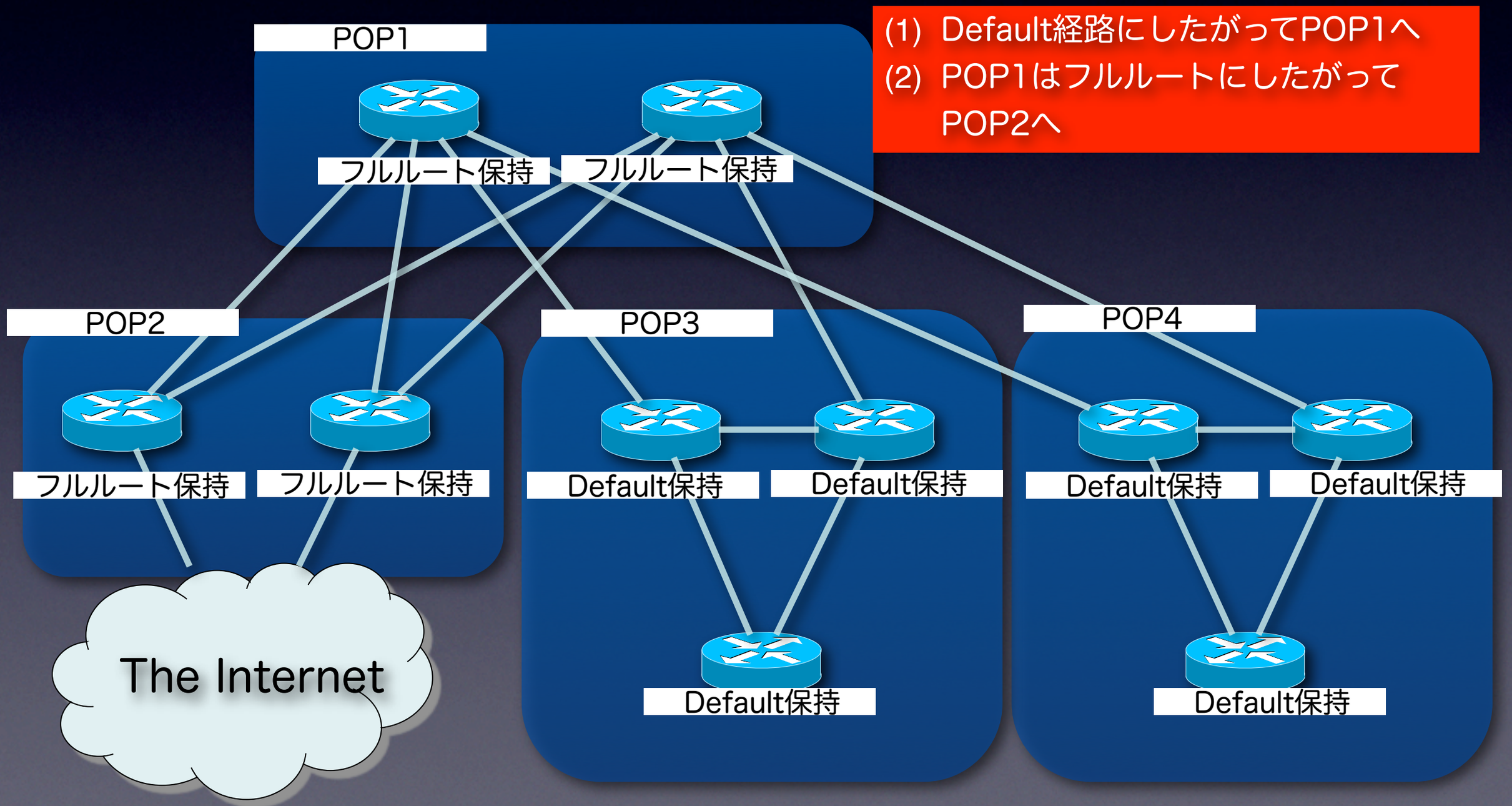


もしBGP経路を持っていたら



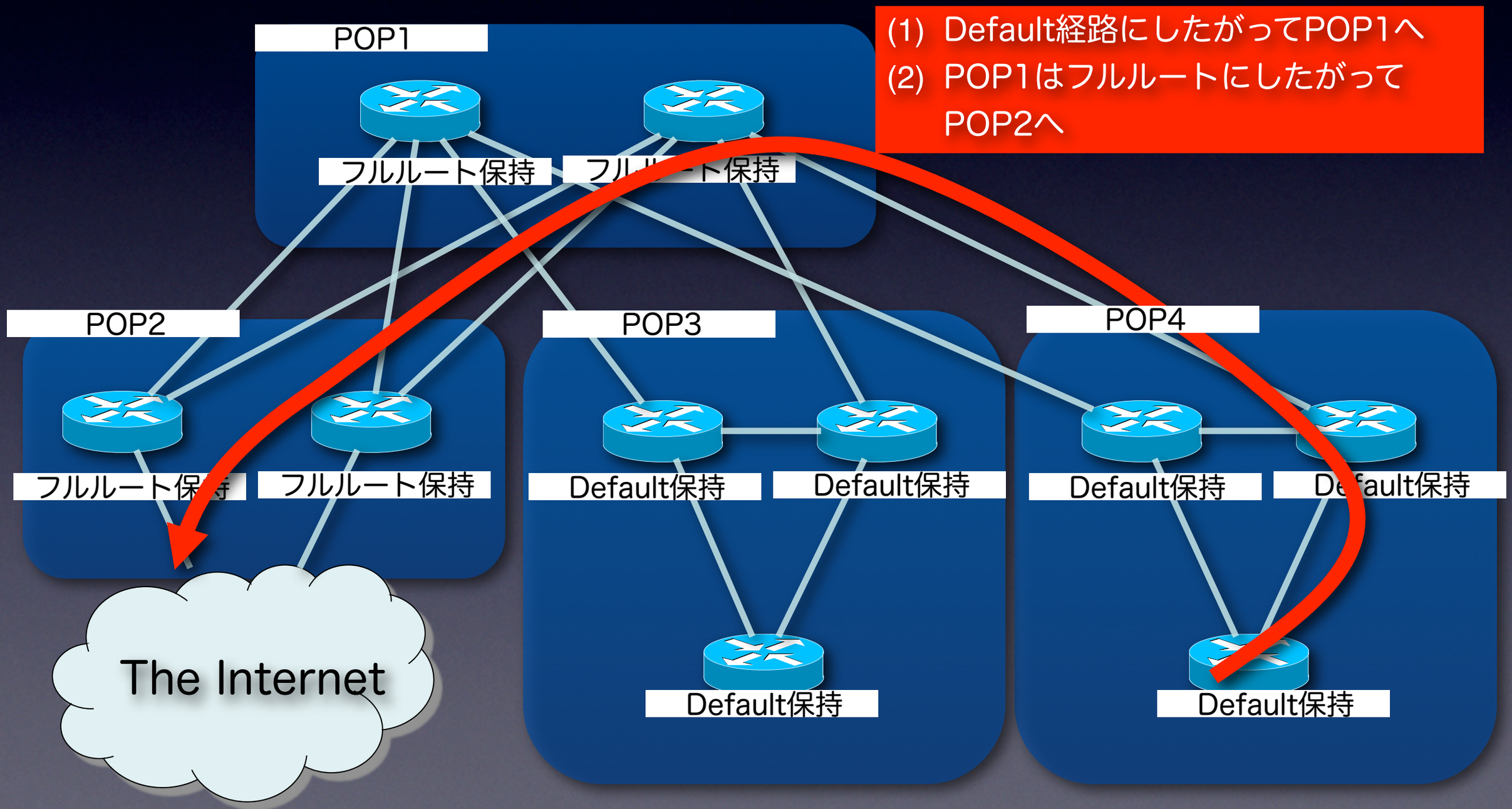
安易にやると...

- POP4→POP2（のインターネット）へ到達したい



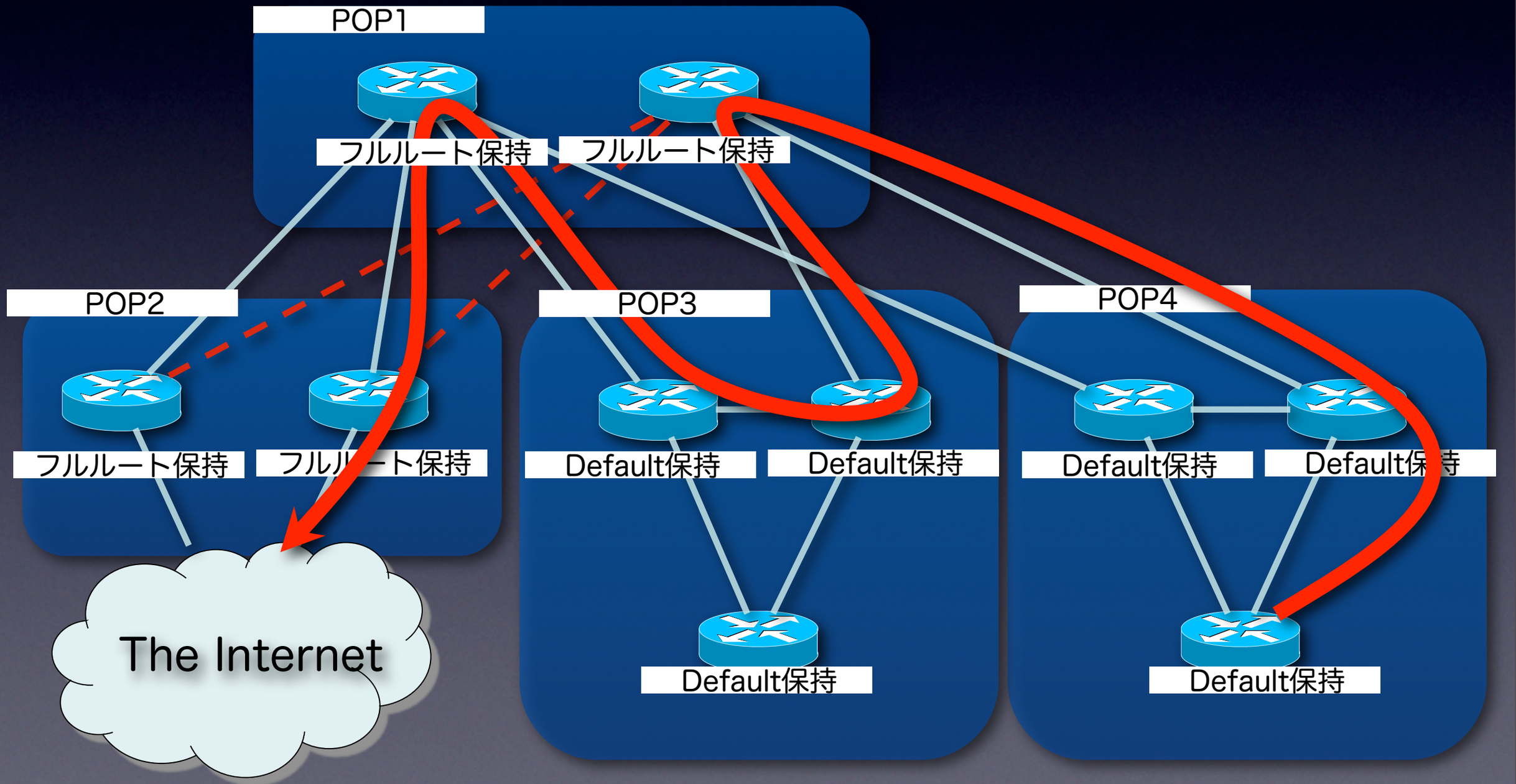
安易にやると...

- POP4→POP2（のインターネット）へ到達したい



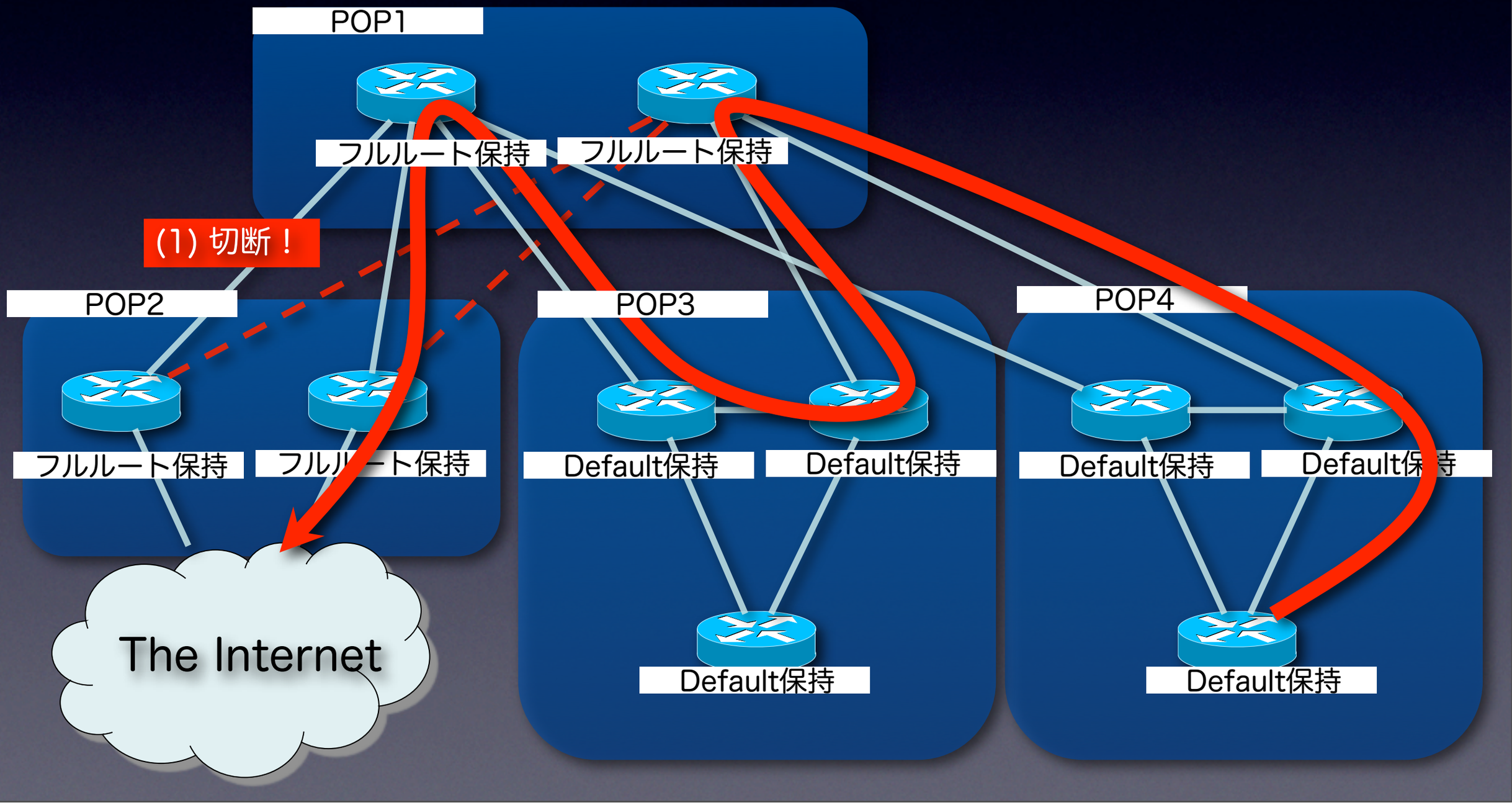
安易にやると…

- POP4→POP2（のインターネット）へ到達したい



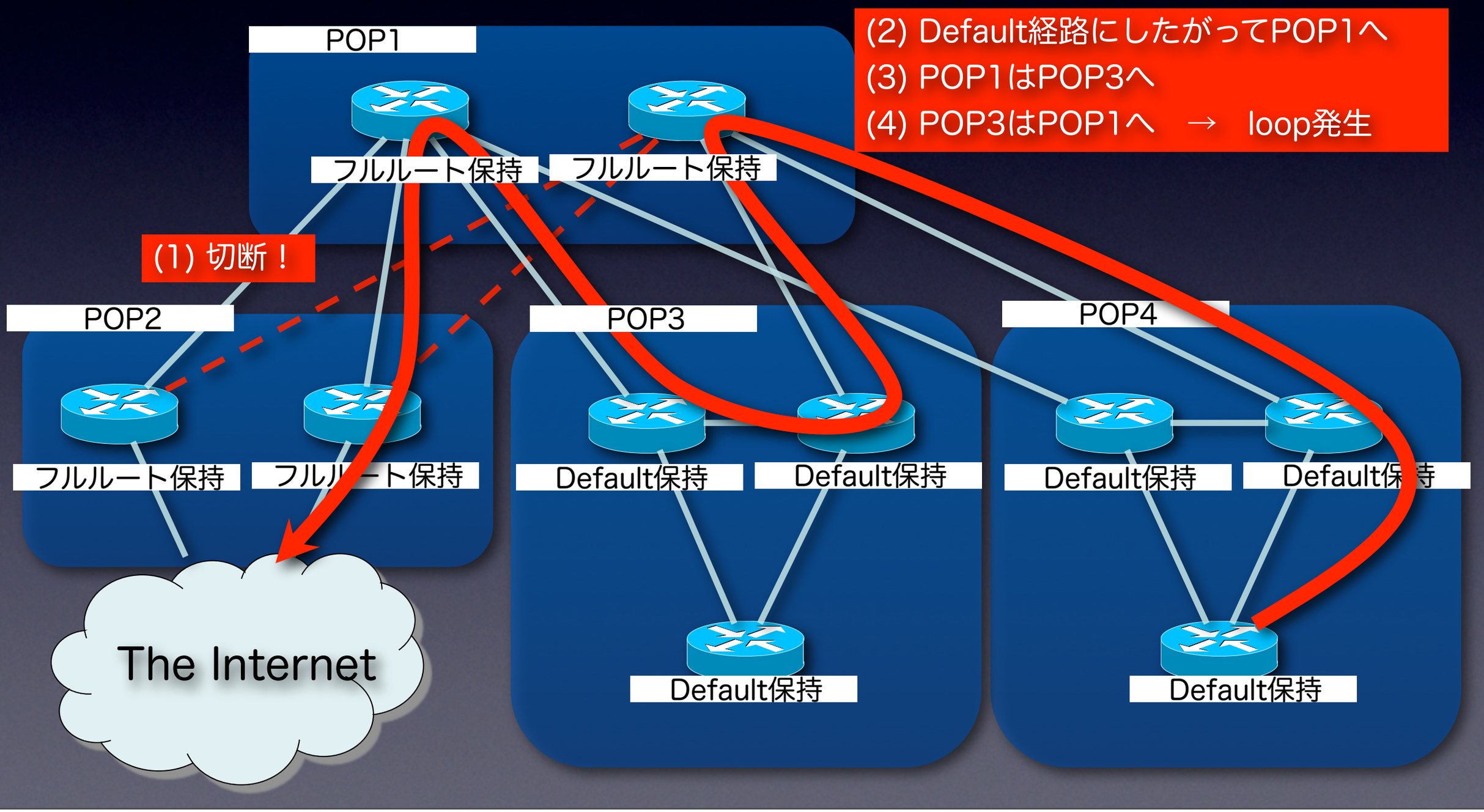
安易にやると…

- POP4→POP2（のインターネット）へ到達したい



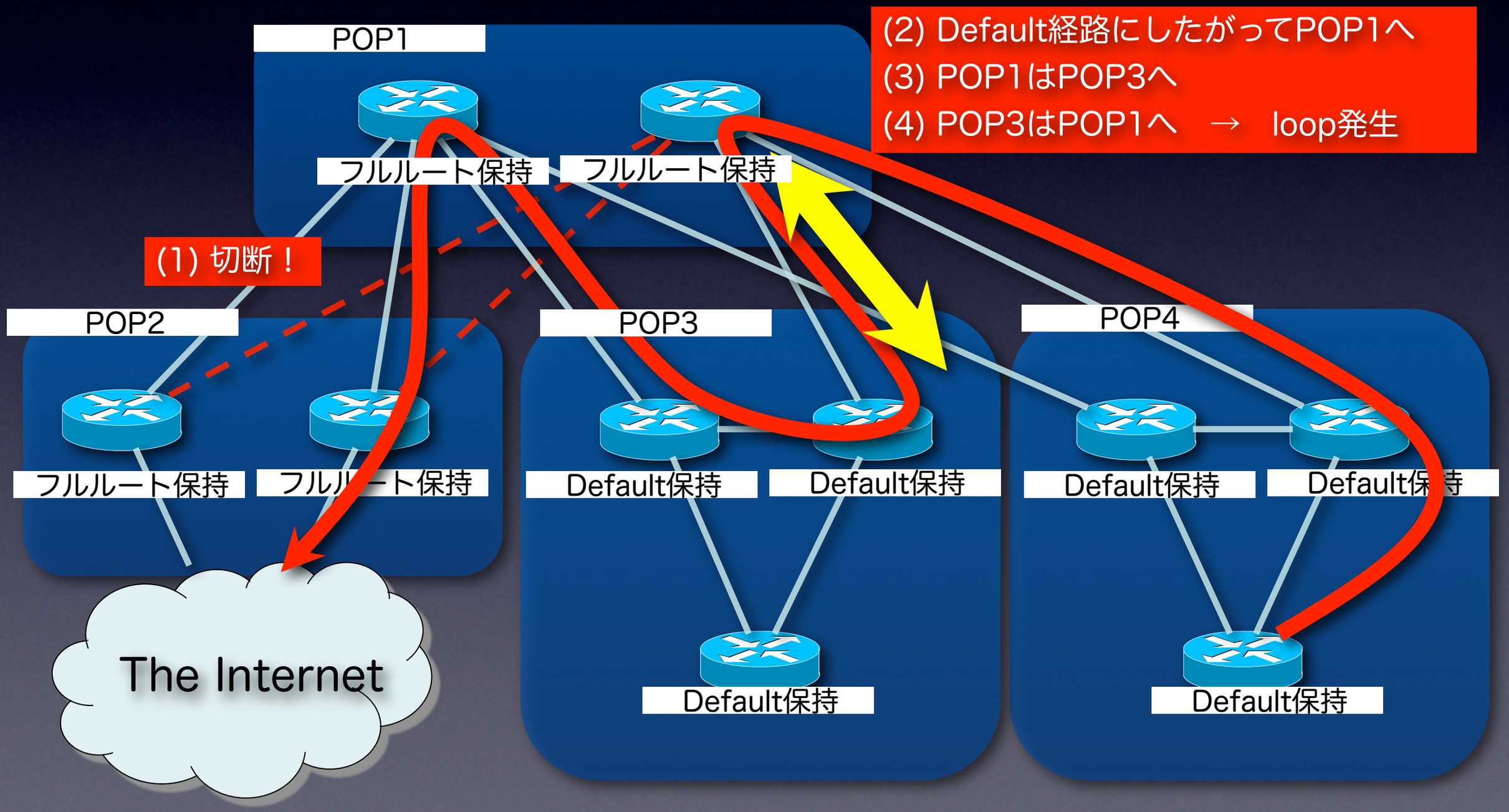
安易にやると…

- POP4→POP2（のインターネット）へ到達したい



安易にやると…

- POP4→POP2（のインターネット）へ到達したい



皆様どうしていますか？

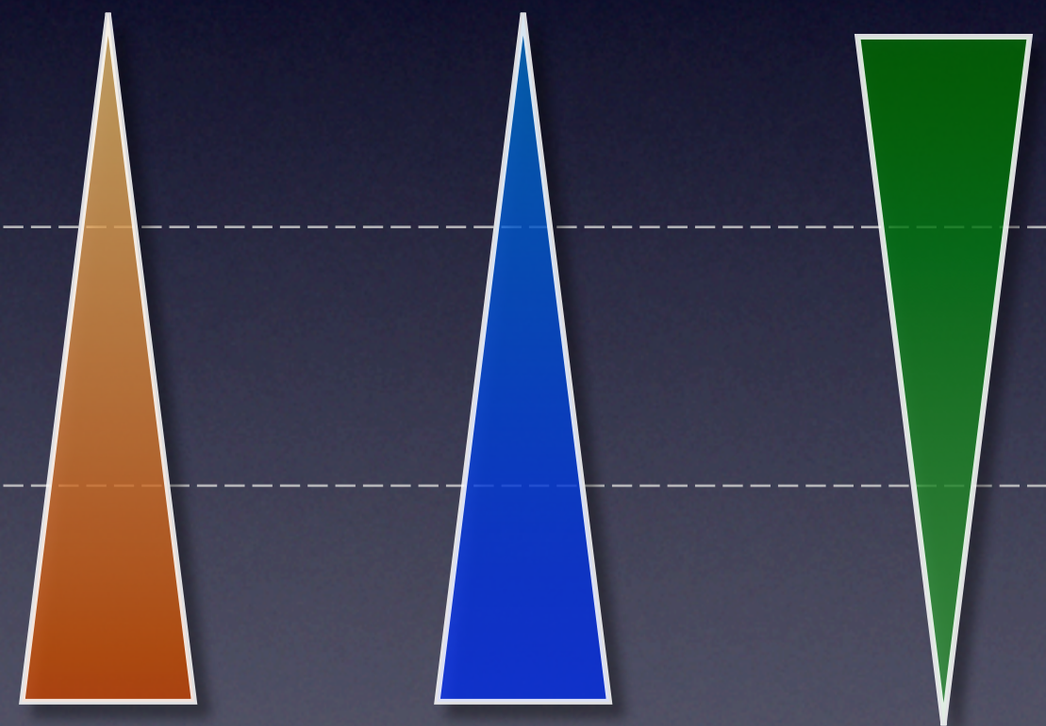
装置コスト 経路収束時間　運用性

- フルルート全部保持
- 限定的に保持
- 基本持たない

皆様どうしていますか？

装置コスト 経路収束時間　運用性

- フルルート全部保持
- 限定的に保持
- 基本持たない



皆様どうしていますか？

- フルルート全部保持
- 限定的に保持
- 基本持たない

質問タイム

異常経路

異常経路

- 過去の事例
- 検知
- 予防と対策

過去の事例

- 不正Attribute

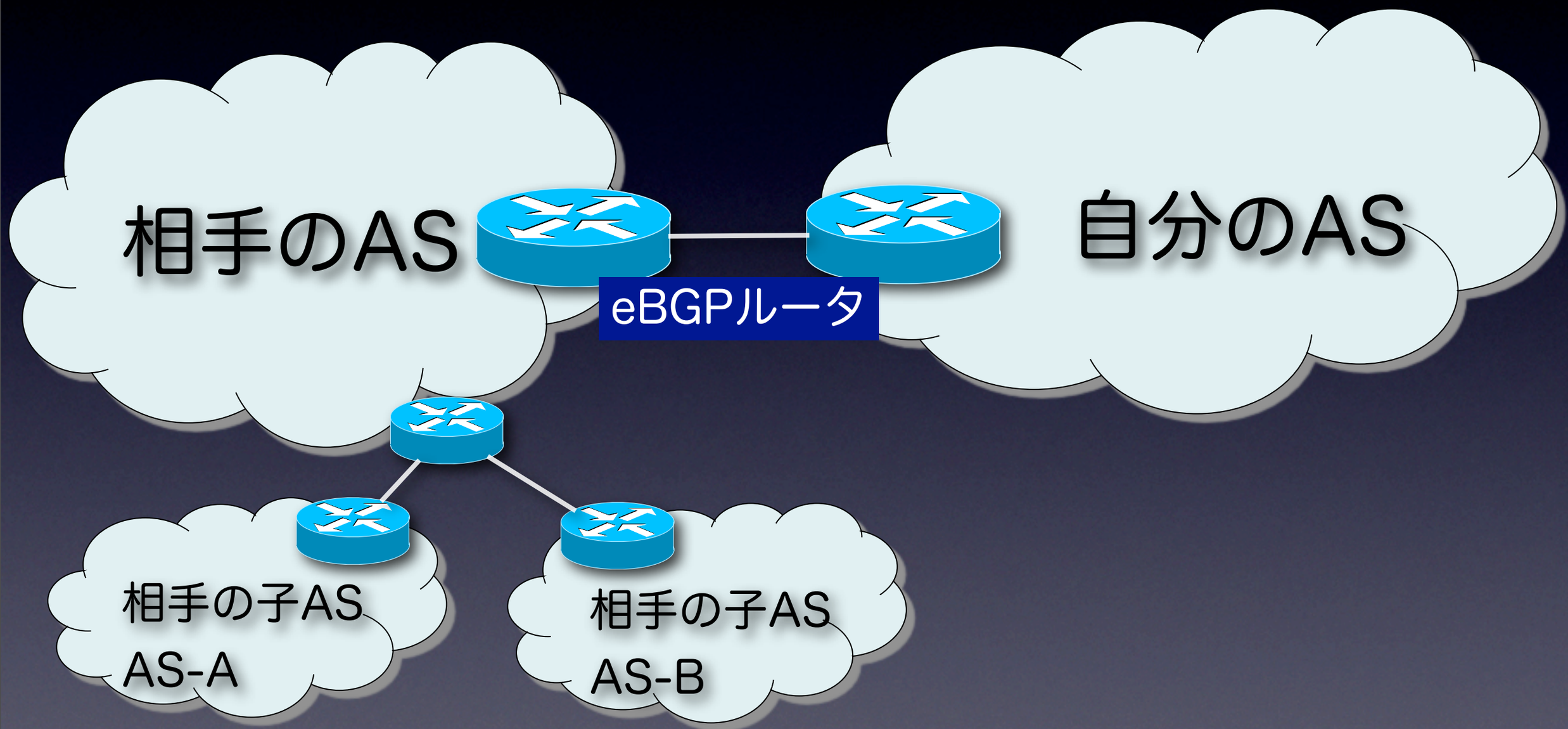
2009/1/26	“BGP Session Teardown due to AS_CONFED_SEQUENCE in AS4_PATH”
2009/2/17	“very long as paths”
2009/8/18	AS4PATH　AS0 “ (invalid or corrupt AS path) 3 bytes E01100”

検知

- 基本はBGP PeerDownが大量発生して大騒ぎ☹
- 収束させるためには経路の保存を日頃から行い解析できるようにしておく
- 検知してからどうするかは、ケースバイケース。

予防: 経路フィルタ

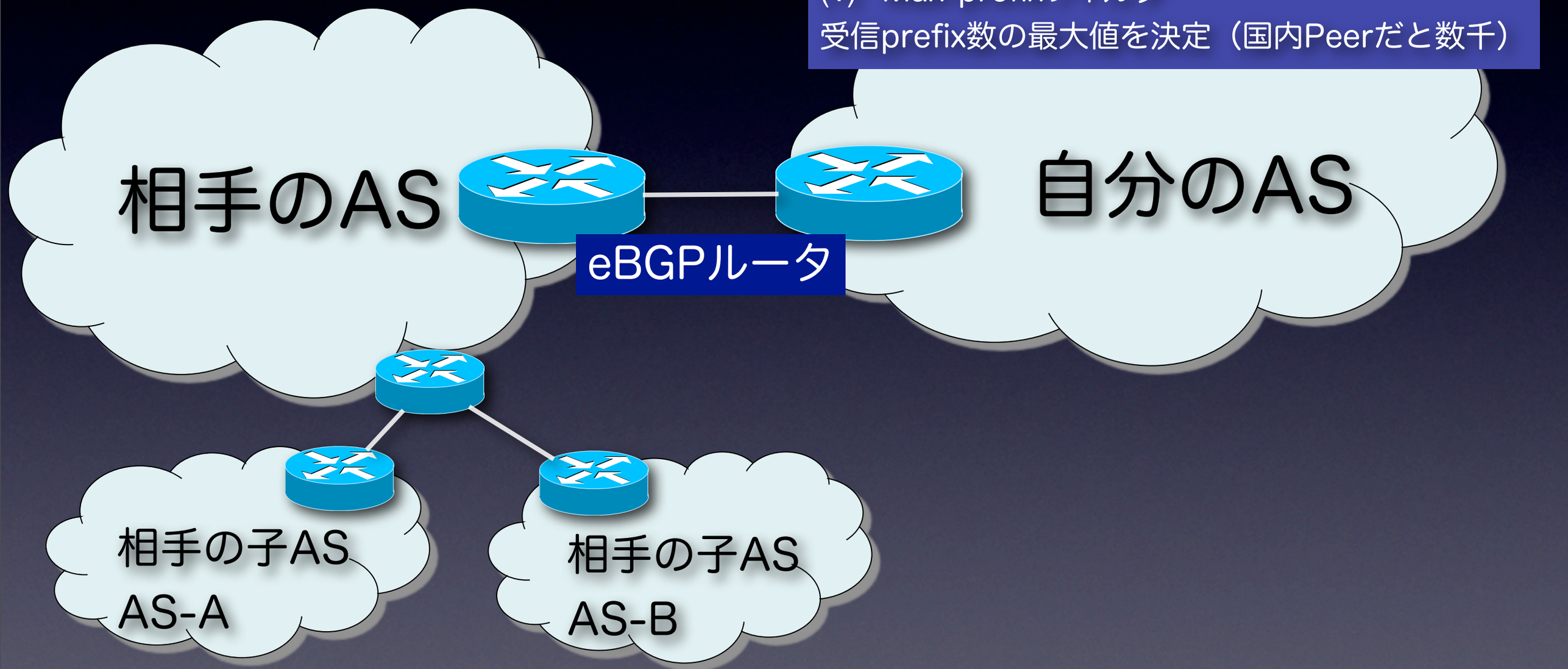
- 一般的な経路フィルタの方法



予防: 経路フィルタ

- 一般的な経路フィルタの方法

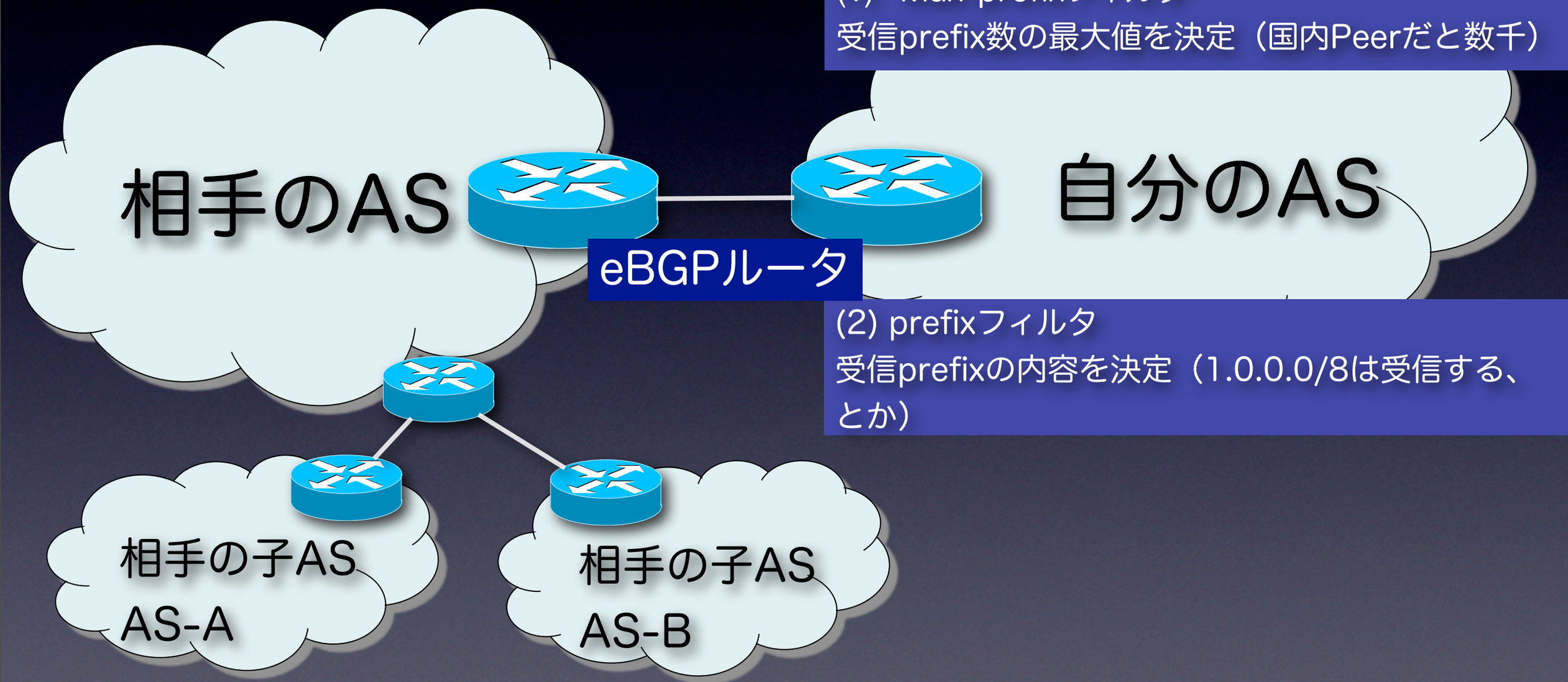
(1)　Max-prefixフィルタ
受信prefix数の最大値を決定（国内Peerだと数千）



予防: 経路フィルタ

- 一般的な経路フィルタの方法

(1)　Max-prefixフィルタ
受信prefix数の最大値を決定（国内Peerだと数千）

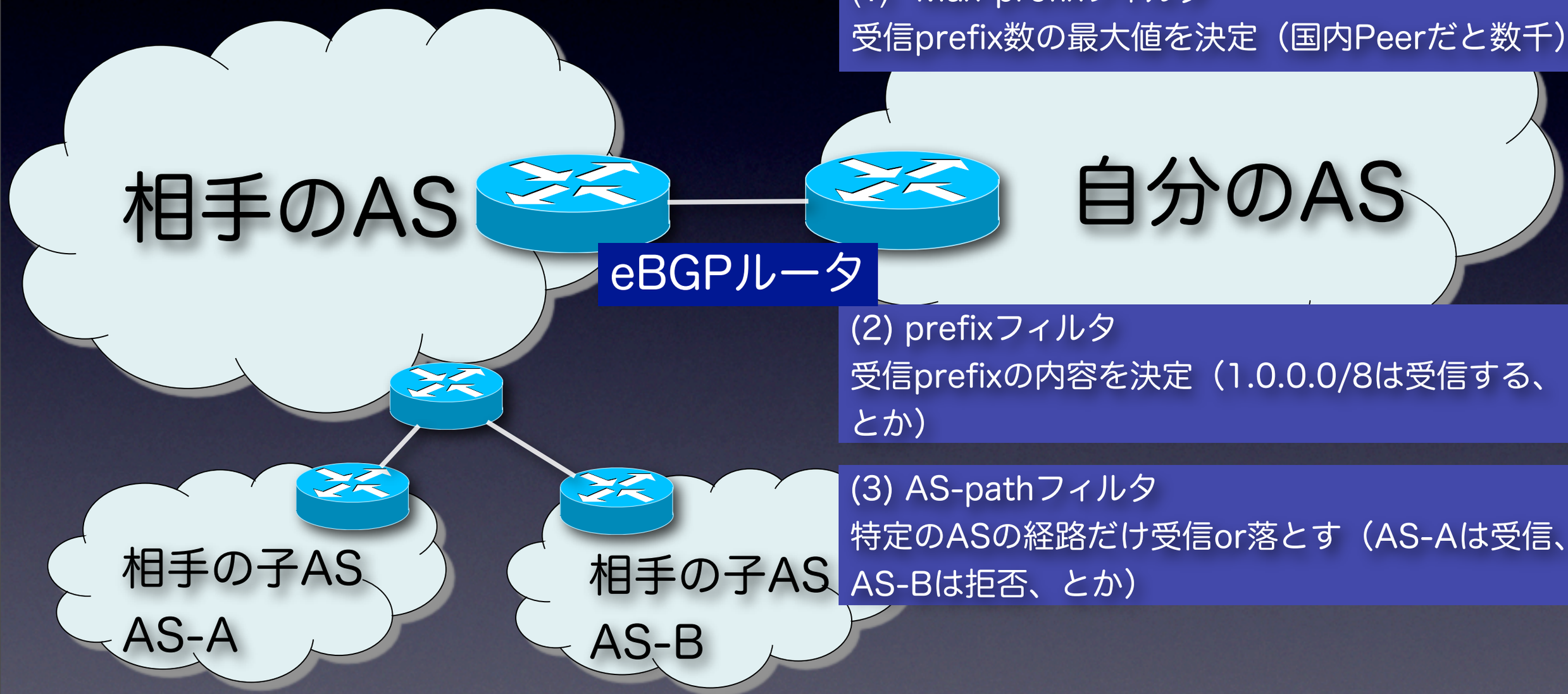


(2) prefixフィルタ
受信prefixの内容を決定（1.0.0.0/8は受信する、とか）

予防: 経路フィルタ

- 一般的な経路フィルタの方法

(1)　Max-prefixフィルタ
受信prefix数の最大値を決定（国内Peerだと数千）



(2) prefixフィルタ
受信prefixの内容を決定（1.0.0.0/8は受信する、とか）

(3) AS-pathフィルタ
特定のASの経路だけ受信or落とす（AS-Aは受信、AS-Bは拒否、とか）

予防: 経路フィルタ

- 一般的な経路フィルタの方法

(1)　Max-prefixフィルタ
受信prefix数の最大値を決定（国内Peerだと数千）

(2) prefixフィルタ
受信prefixの内容を決定（1.0.0.0/8は受信する、とか）

(3) AS-pathフィルタ
特定のASの経路だけ受信or落とす（AS-Aは受信、AS-Bは拒否、とか）

(4) Damping
受信prefixがFlapしたときには一定時間抑制する

予防: 経路フィルタ

- 一般的な経路フィルタの方法

(1)　Max-prefixフィルタ
受信prefix数の最大値を決定（国内Peerだと数千）

(2) prefixフィルタ
受信prefixの内容を決定（1.0.0.0/8は受信する、とか）

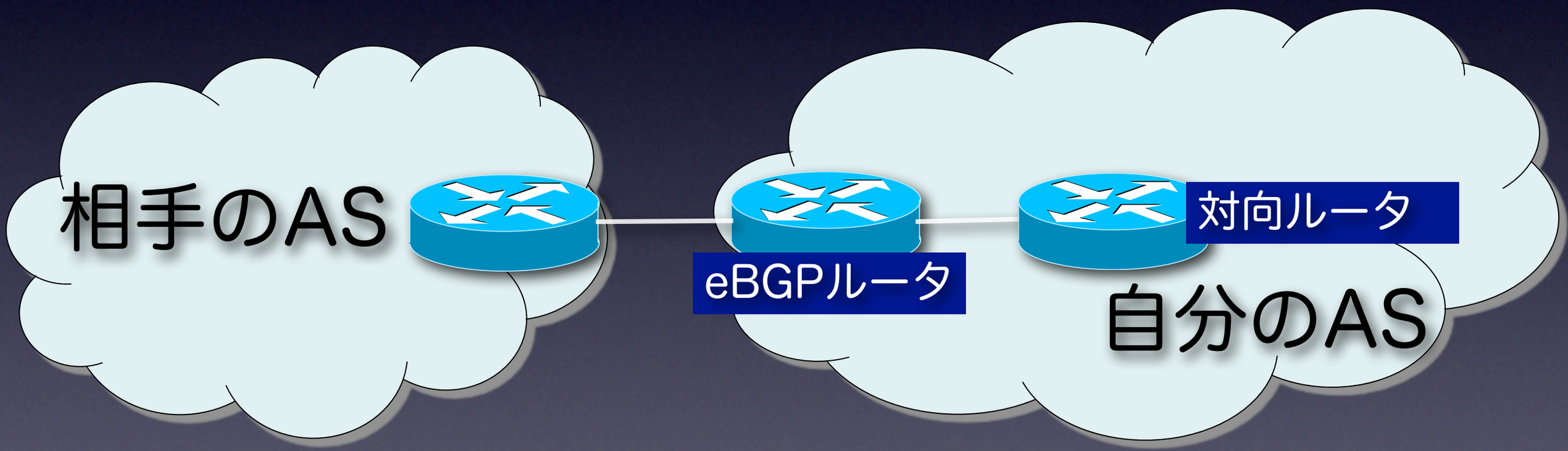
(3) AS-pathフィルタ
特定のASの経路だけ受信or落とす（AS-Aは受信、AS-Bは拒否、とか）

(4) Damping
受信prefixがFlapしたときには一定時間抑制する

- ただし、セキュリティホールになるような「異常経路」はBGP Attributeそのものが不正なことが多い

予防: 経路フィルタ

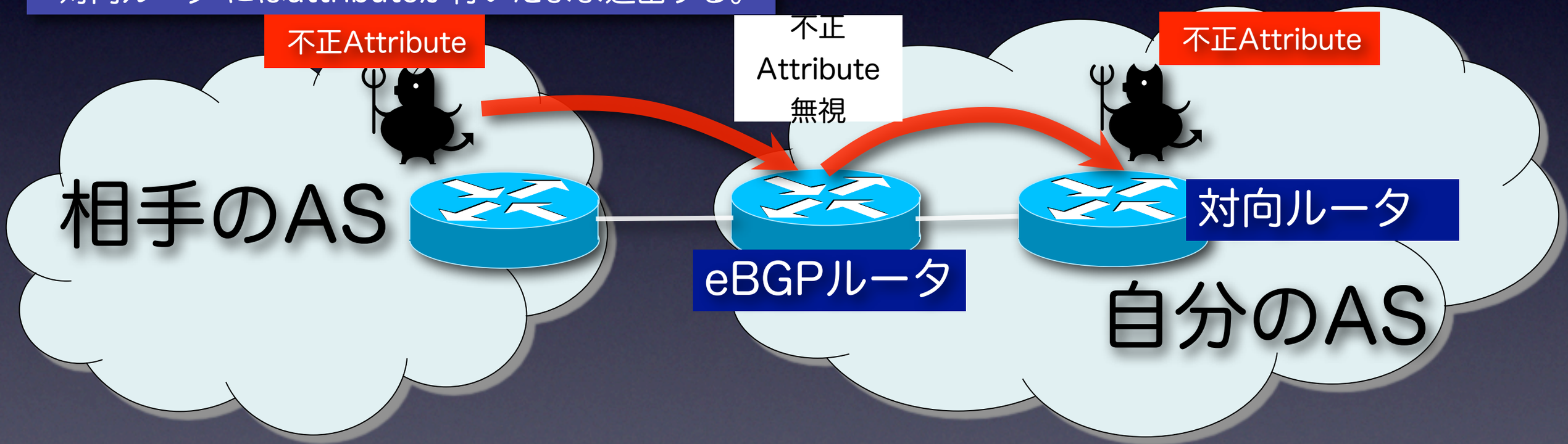
- BGP Attributeフィルタ @JUNOS
  - ignore/drop-path-attributes



予防: 経路フィルタ

- BGP Attributeフィルタ @JUNOS
  - ignore/drop-path-attributes

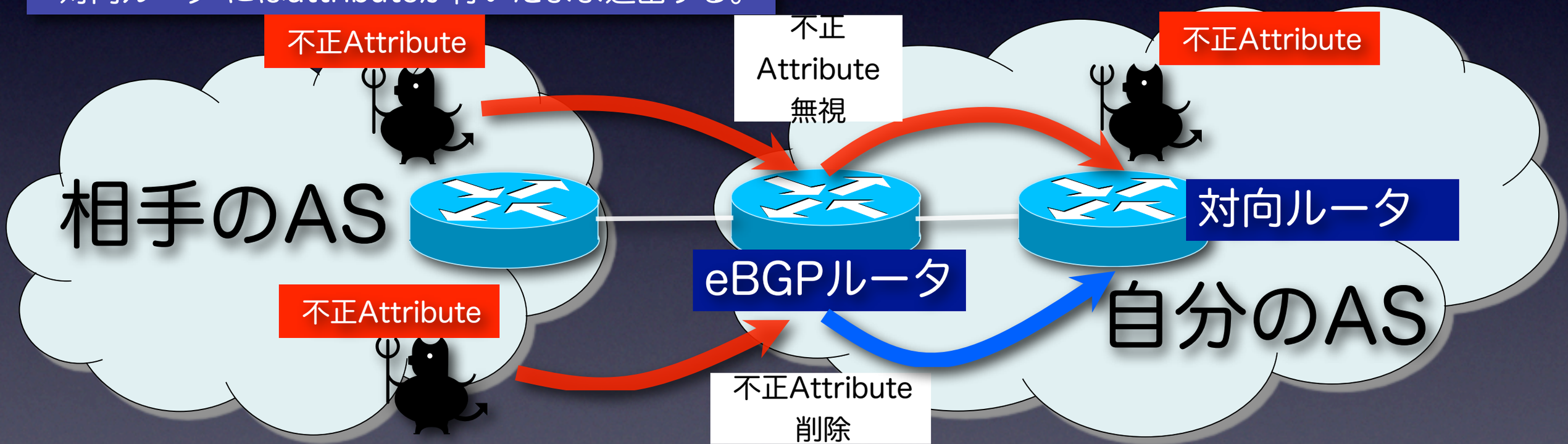
ignore-path-attributes
- attribute を除いた経路情報を自身の RIB に記録
- 対向ルータ にはattributeが付いたまま送出する。



予防: 経路フィルタ

- BGP Attributeフィルタ @JUNOS
  - ignore/drop-path-attributes

ignore-path-attributes
- attribute を除いた経路情報を自身の RIB に記録
- 対向ルータ にはattributeが付いたまま送出する。



drop-path-attributes
- attribute を除いた経路情報を自身の RIB に記録
- 対向ルータ にもattributeを除いた経路情報を送出する。

質問タイム

まとめ

抱える課題

- 経路数の増加
  - 最適な構成は利用形態により様々
  - たぶん今後も経路は増えていく
- 異常経路への対処
- みなさんのネットワークではどうされていますか？

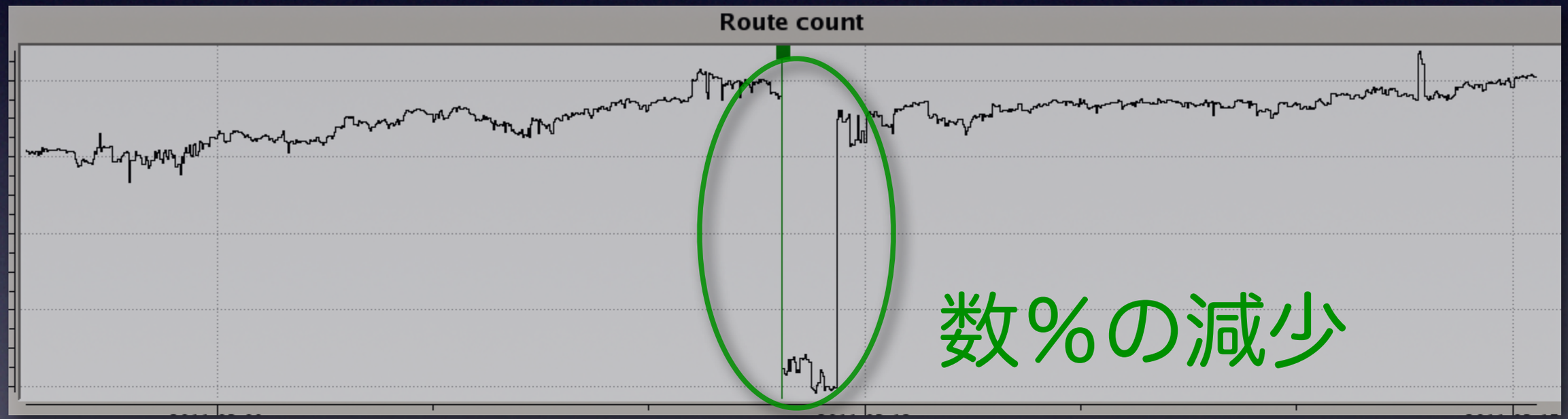
番外編

大震災による経路変動

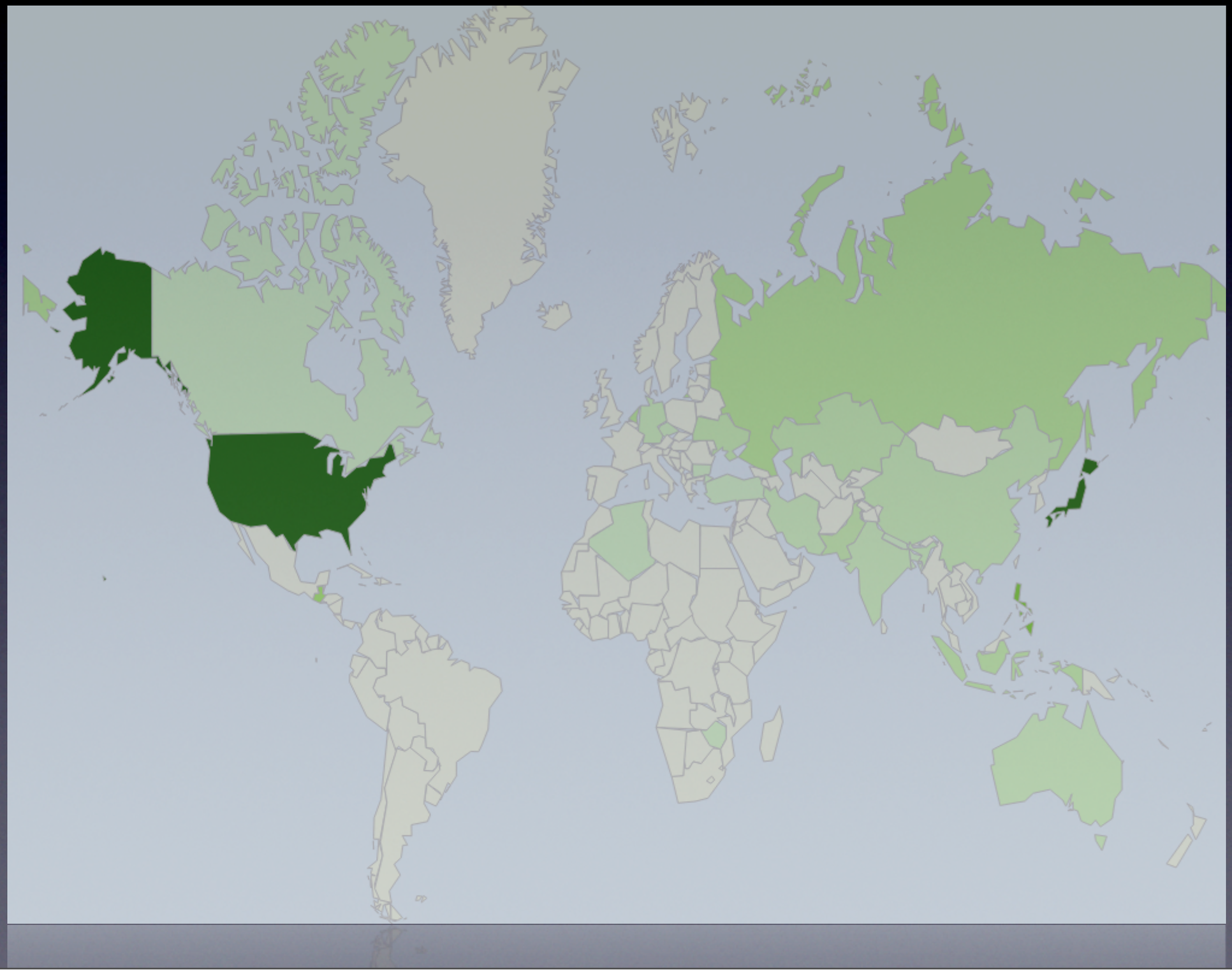
（番外編）
大震災による経路変動

- ①経路って変動したの？
- ②地震のあとって・・・

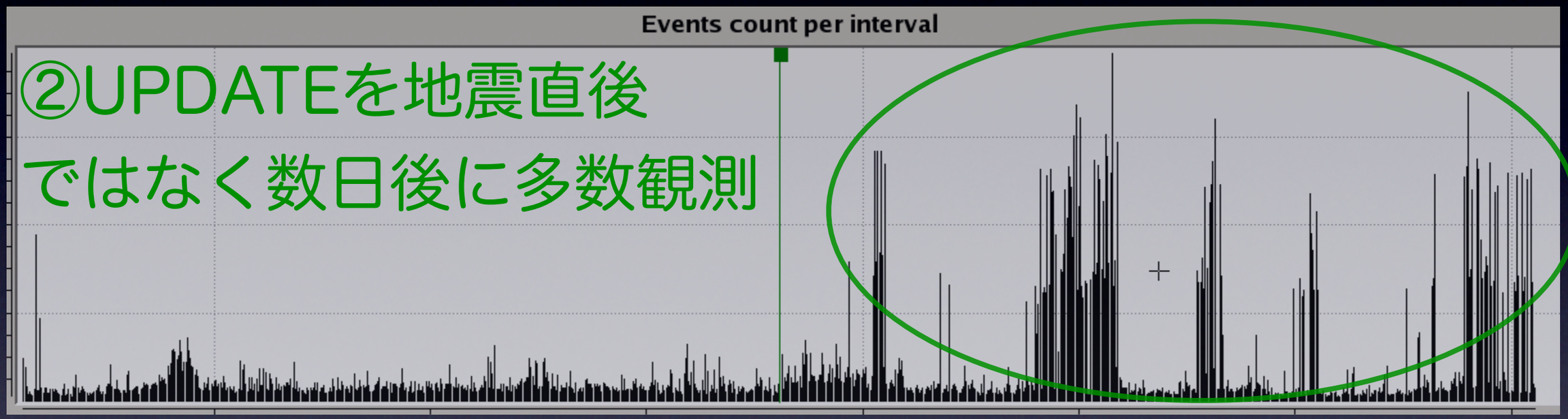
震災時の経路数



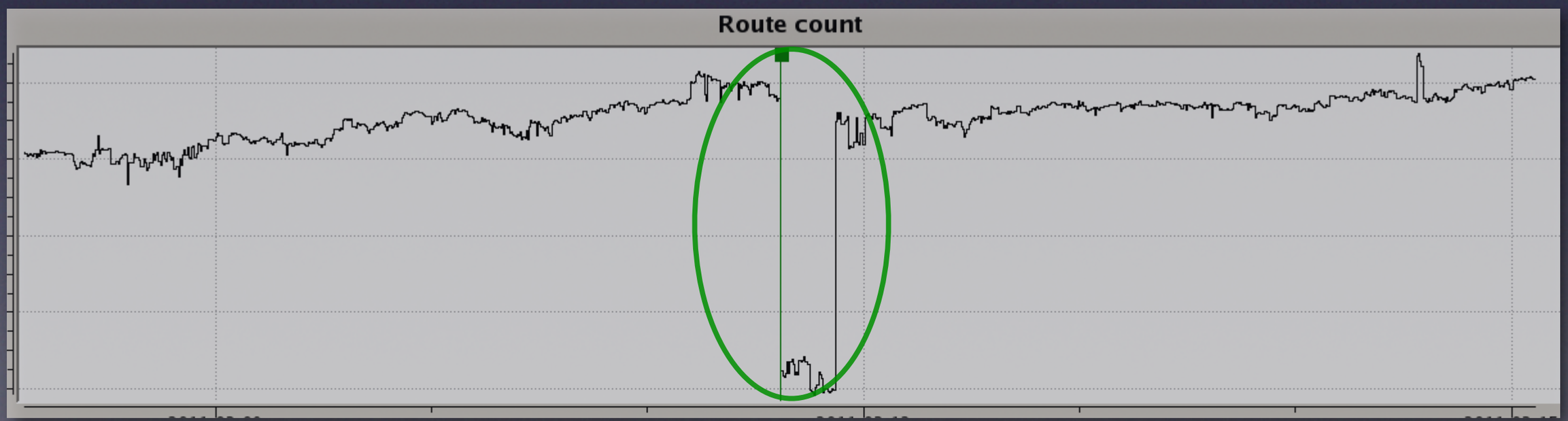
震災時の経路喪失数分布



震災時のUPDATE



②UPDATEを地震直後
ではなく数日後に多数観測



おわり

・ご清聴ありがとうございました

・続きは懇親会で！！