

BGP談議

～add-pathとdiverse-pathのお話～

何か持っていると言われ続けてきました。今日何を持っているのか
確信しました…それは BGPバックアップパス

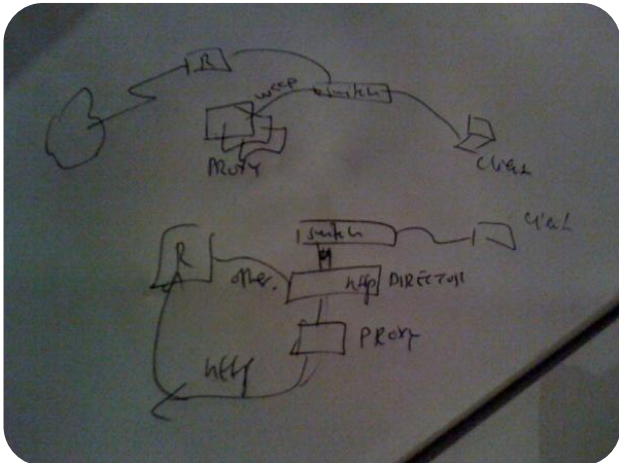
Shishio Tsuchiya

shtsuchi@cisco.com



そもそもBGPって

- 1989年にY.レクターとK.ラフィードがカフェでナプキンに書いたのが始まり
- そのナプキンを元に最初の接続実験実施[RFC1105](#)
- そのナプキンは額縁に入れてシスコに飾られてる
- ケチャップで、ものすごく汚れていた



そもそもBGPって

- [RFC1771](#)が1995年3月に発行
IBM:Y. Rekhter& Cisco Tony.Li
- [draft-ietf-idr-bgp4-01](#)～26まで改訂し、2006年1月に
[RFC4271](#)を発行した。

```
Network Working Group                                Y. Rekhter
Request for Comments: 1771                          T.J. Watson Research Center, IBM Corp.
Obsoletes: 1654                                       T. Li
Category: Standards Track                          cisco Systems
                                                    Editors
                                                    March 1995
```

A Border Gateway Protocol 4 (BGP-4)

どこで決めてるの？

- IETFでは主に2つ

IDR(Inter Domain Routing Group)とGROW(Global Routing Operations)

IDR	Working Group	GROW
Routing	Area	Operation & Management
BGP4のIPv4/IPv6でBGPv4の 利用のサポート 頑健性/拡張性の改善	目的	BGP4のIPv4/IPv6での利用の際 の運用的問題の考慮
45個(Obsoleteも含む) Standardが中心	RFC	6個 BCPとInfo.

BGPの特徴

- [RFC4271の9.2. Update-Send Process](#)では他のBGPスピーカーに伝えるルートは決定プロセスにより、選ばれたものを伝えなければならない事を定義している。
- 全く同じアドレスPrefix/NLRIのものの複数パスの広告
は許されていない。
- 以前アドバタイズされたルートと同じNLRIをもつルート
は暗黙のうちに取り換えられる。

バックアップパスがあればできる事

- ファーストコンバージェンス
バックアップパスがあれば重要障害時もローカルで切り替え可能
- マルチパスロードバランス
- 安定性と正確性の向上
MED Churn によるルートグラつき(oscillation) [RFC3345](#)

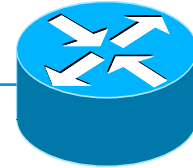
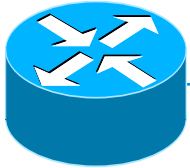
IETFでは現在add-path や diverse-pathというソリューションが議論され、標準化中

バックアップパスを伝える手法

- BGP add-path
- BGP Best External + Fullmesh
- BGP Diverse Path

add-pathって何だろう？

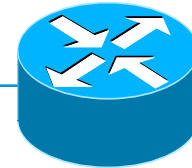
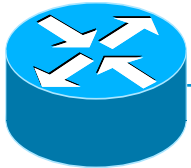
- 通常のBGP



10/8,AS_PATH:64500,64501,NH:203.0.113.10

10/8,AS_PATH:64510,64511,NH:203.0.113.10

- add-path



10/8,ID=1,AS_PATH:64500,64501,NH:203.0.113.10

10/8,ID=2,AS_PATH:64510,64511,NH:203.0.113.10

add-path詳細

Capability Exchange

+	-----	+
	Address Family Identifier (2 octets)	
+	-----	+
	Subsequent Address Family Identifier (1 octet)	
+	-----	+
	Send/Receive (1 octet)	
+	-----	+

- BGPピアにてCapability Exchangeを行う([RFC3992](#))
Capability Valueは69
- Send/Receiveでマルチパスを
Receive(=1),Send(=2),Both(=3)可能かを示す

add-path詳細

NLRI Encoding

NLRI Encoding

RFC4271, RFC4760

+-----+	
Length (1 octet)	
+-----+	
Prefix (variable)	
+-----+	

Extended NLRI Encoding

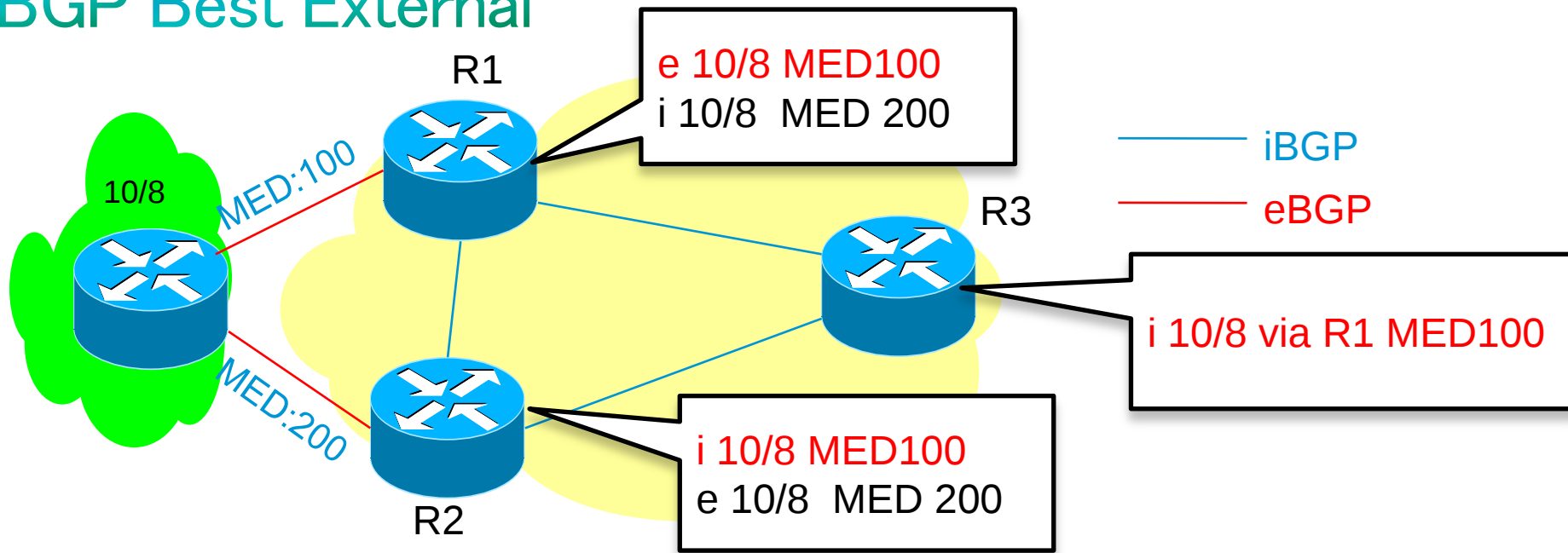
+-----+	
Path Identifier (4 octets)	
+-----+	
Length (1 octet)	
+-----+	
Prefix (variable)	
+-----+	

- UPDATEメッセージの中に4オクテットのPATH IDを含む
為に、Path Identifierフィールドをつける
- [RFC3107](#)でも同様にLengthの前部につける

バックアップパスを伝える手法

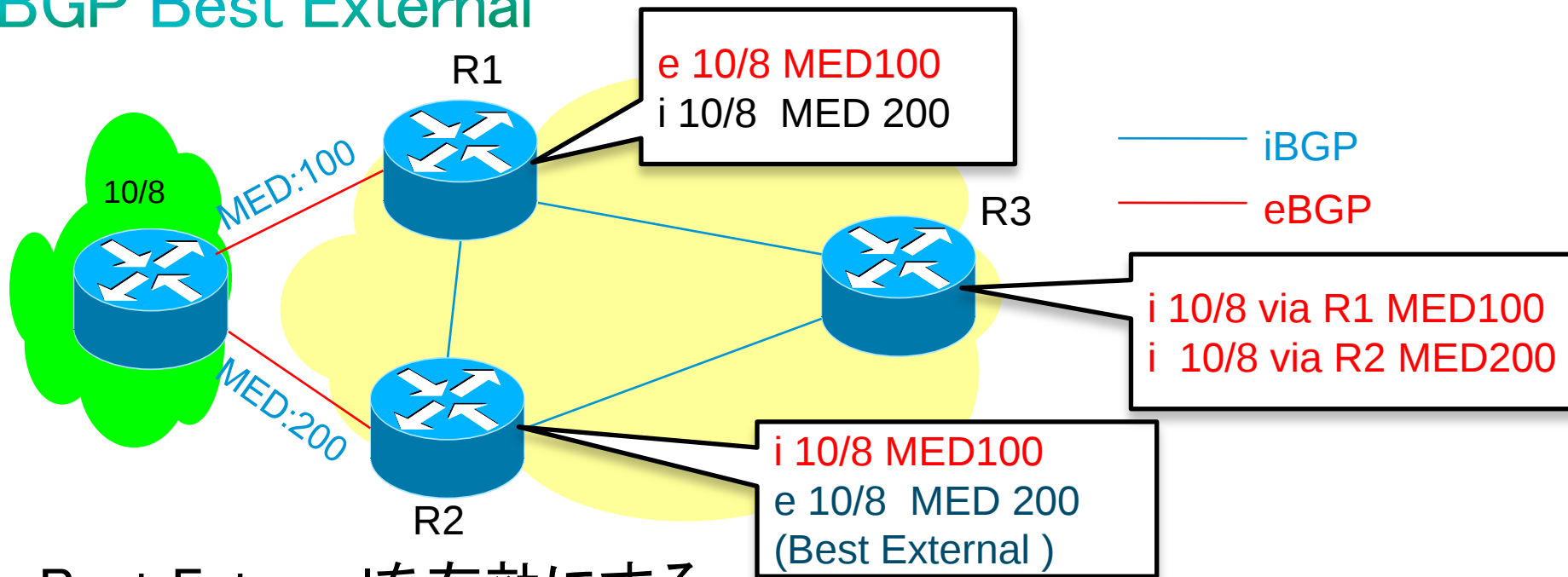
- BGP add-path
- BGP Best External + Fullmesh
- BGP Diverse Path

BGP Best External



- MEDの値により、R2ではいつもiBGPが選択される
- R3では代替えパスを持たない
- R1障害時にはWITHDRAWN/UPDATE、書き換えが行われる

BGP Best External

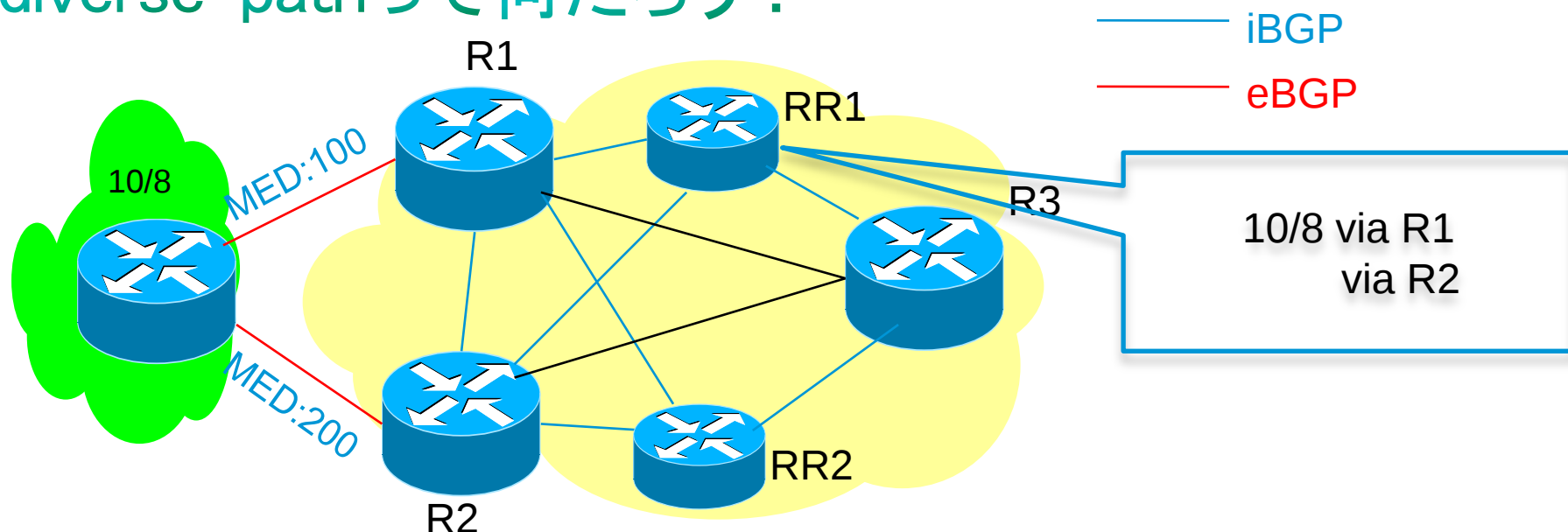


- Best Externalを有効にする
- R2では全体でベストパスを選出する(MEDにより選出)
- R2では外部経路の中のみでのベストパスを選出する(Best External)
- R3では代替パスを持つことが出来る

バックアップパスを伝える手法

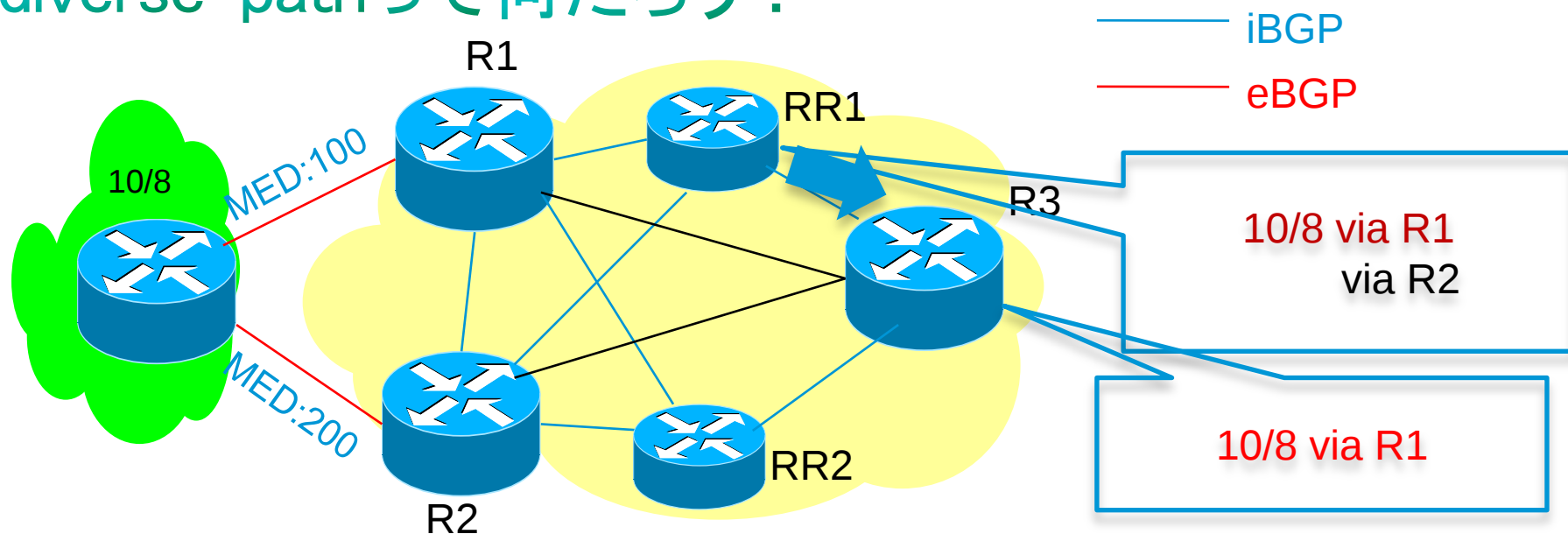
- BGP add-path
- BGP Best External + Fullmesh
- **BGP Diverse Path**

diverse-pathって何だろう？



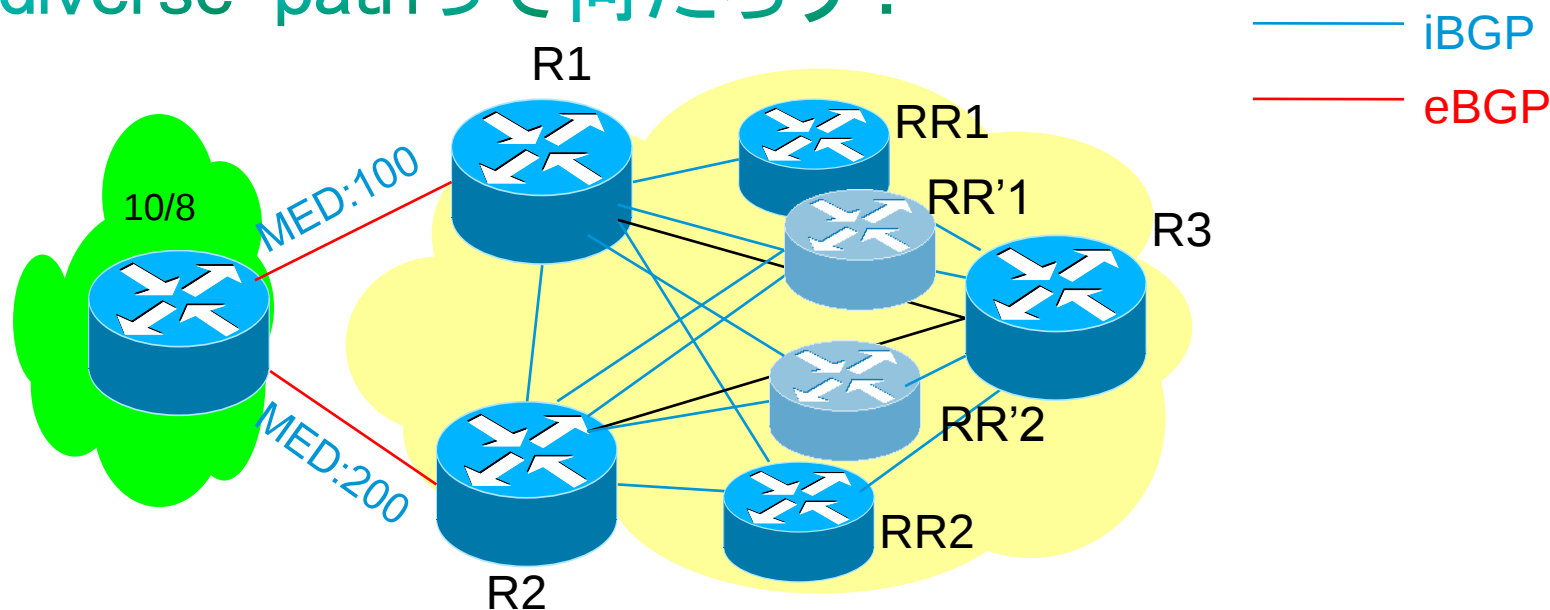
- R1はMED100,R2はMED200とする
- Best Externalを有効にする
- RR1,RR2ではR1およびR2経由のパスを持つ
- つまりRRではパスダイバーシティ(diverse-path)を持つ。

diverse-pathって何だろう？



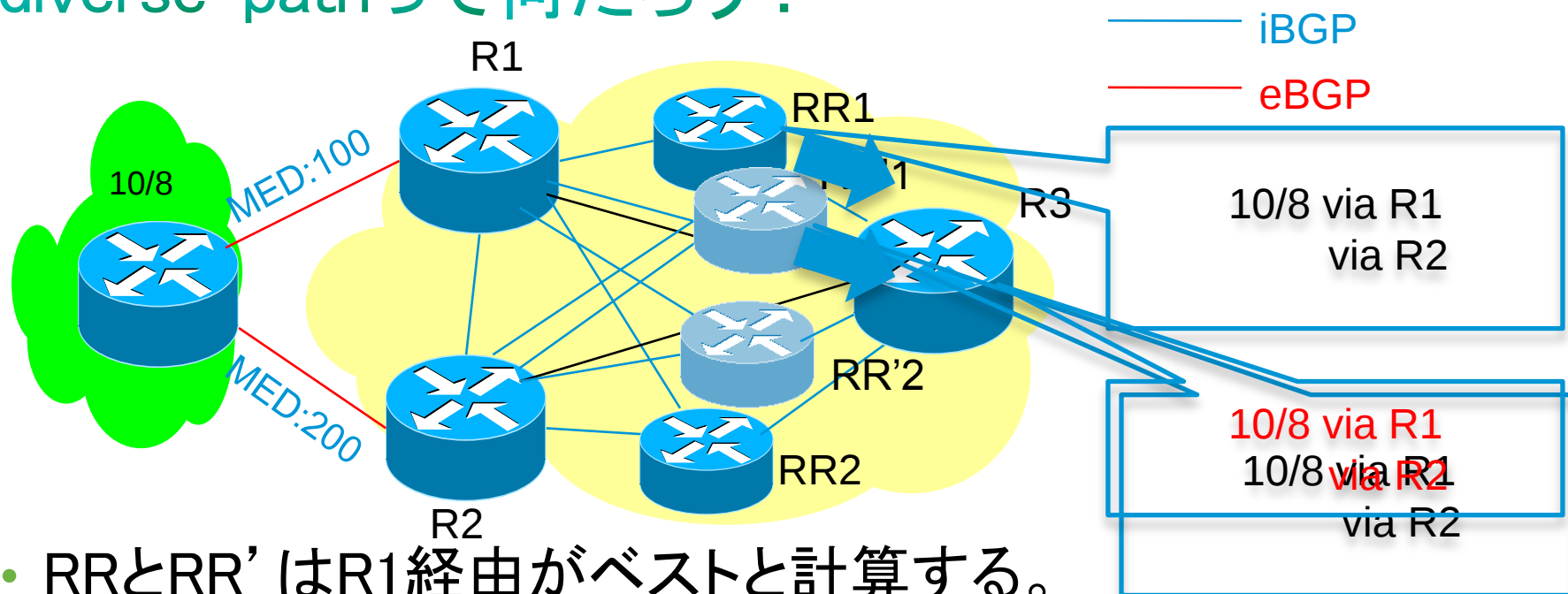
- RRではMEDの値よりR1経路がベストと判断する
- RRはベストパスをアドバタイズする
- R3はR2経由のパスを知ることが出来ない

diverse-pathって何だろう？



- diverse-pathを分配する為、新しいRRを追加
- 2nd Plane RR/Shadow RRという
- RRとShadow RRはiBGPピアを設定しない

diverse-pathって何だろう？

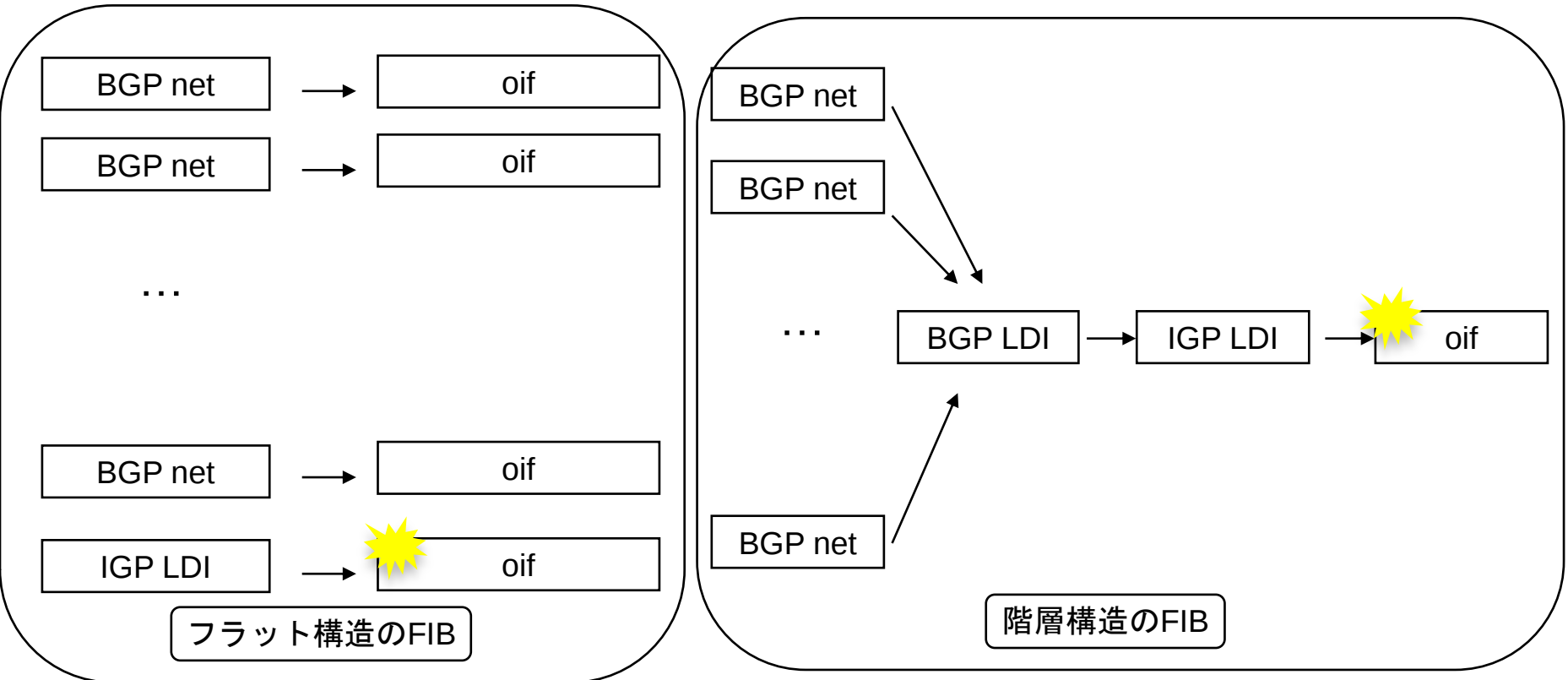


- RRとRR' はR1経路がベストと計算する。
- RR' ではベストパスを排除、残りのパスで再計算する
- RR' ではR2経由を見つけ、2nd Bestと認識する。
- RR' ではR2経由をベストパスとしてアナウンスする
- R3ではdiverse-pathを持つ

バックアップパスを伝える手法

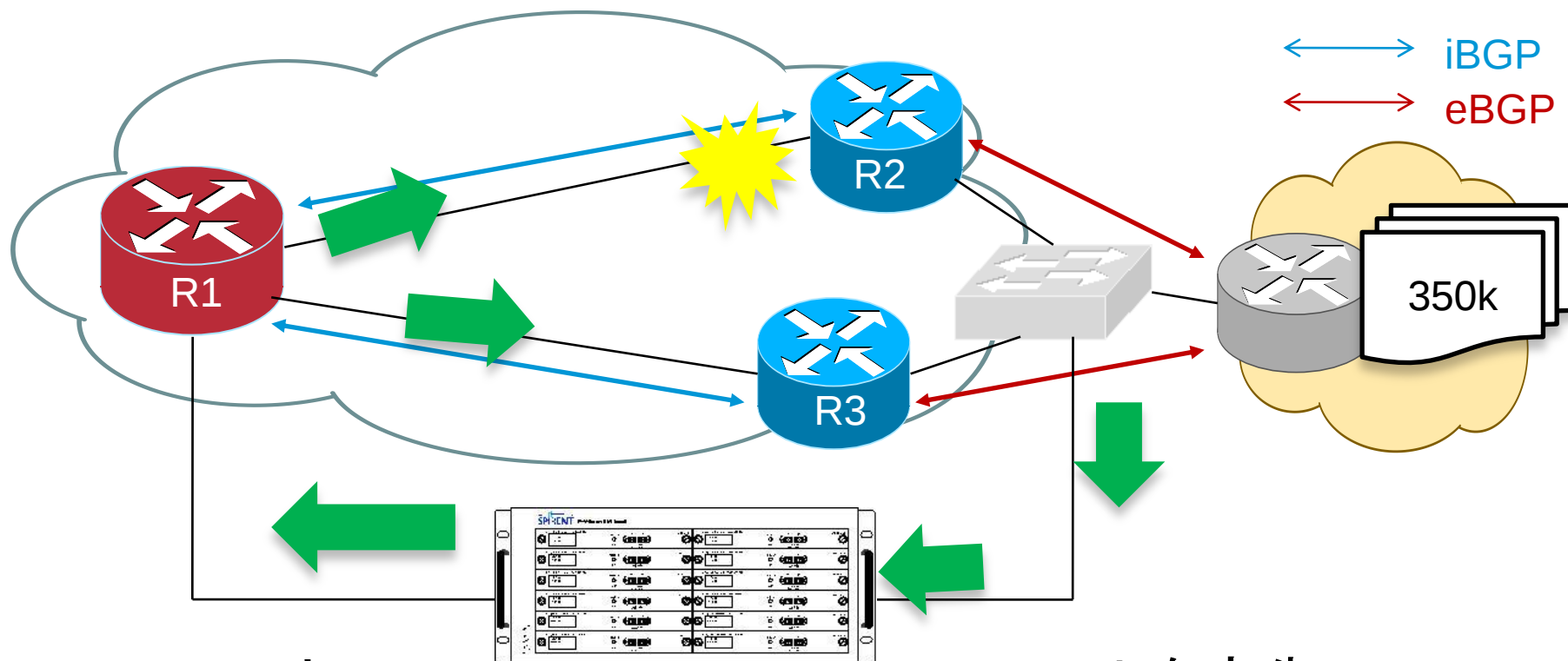
solution	add-path	BGP Best External +Full mesh	diverse-path
internet-draft	draft-ietf-idr-add-paths	draft-ietf-idr-best-external	draft-ietf-grow-diverse-bgp-path-dist
Intend to	Standard	Standard	Informational
Author	Cisco,Juniper	Cisco(was with Juniper)	Cisco,Verisign,KDDI
概要	BGPプロトコル拡張	ASBRの機能拡張	RRの機能拡張
導入手法	全てのルータがサポート (ASBR,RR,ABR)	ASBRがサポート iBGP フルメッシュが必要	RRがサポート ASBR/ABRでのShadow RR へのセッション追加

FIBの構造の違いーPrefix Independent Convergenceー



- フラット構造のFIBでは全てのコントロールプレーンはフラットに扱われ、BGP FIBはそれぞれoifを持つ
- 障害が起こると1prefix毎にoifの書き換えが行われる
- 階層構造では即座にIGP収束後、即座に書き替えが行われる

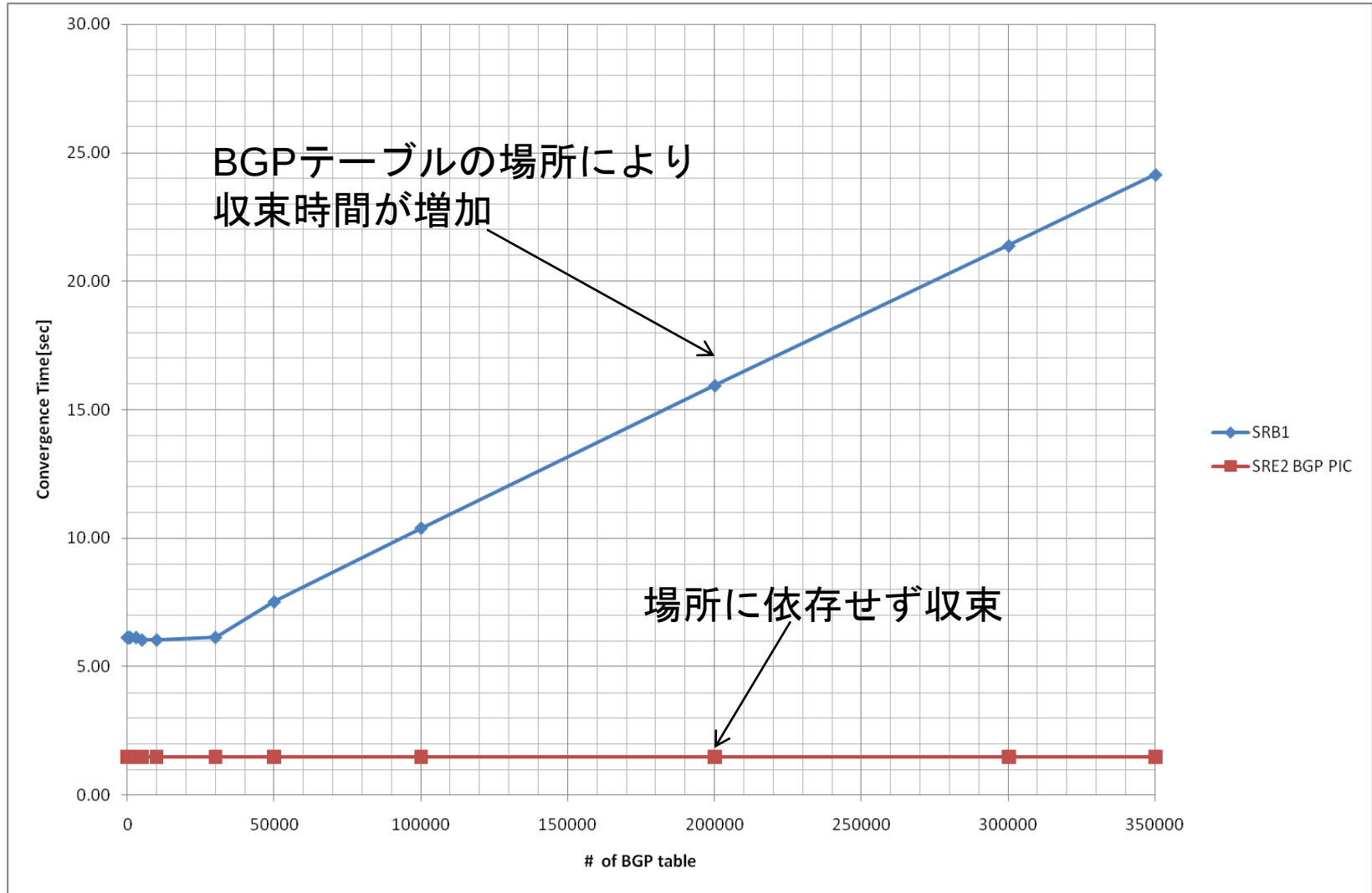
FIBの構造だけでどれだけ違うのか？



- R1-R2,R3はiBGP,エミュレータより350Kルートを広告
- R1-R2間のリンクに障害を発生させる
- 複数フローにて収束時間を測定
- 1,500,1K,3K,5K,10K,30K,50K,100K,200K,300K,350K番目

Convergence Time比較

フラット構造のFIB(SRB1) vs 階層構造のFIB(SRE2)



まとめ

- まだまだBGPは拡張し続けてる
- 高速コンバージェンスの為に代替パスを如何に伝えるかがポイント
- add-pathではプロトコルを拡張
- diverse-pathはRRの機能を拡張
- Full meshであればBest Externalも使用可能
- 代替パスがあればPICで高速切り替え可能

参考

IETF Internet Draft

- [Advertisement of Multiple Paths in BGP](#)
- [Distribution of diverse BGP paths](#)
- [Advertisement of the best external route in BGP](#)

Cisco Document

- [BGP ベストパス選択アルゴリズム](#)
- [BGP Best External](#)
- [BGP PIC Edge for IP and MPLS-VPN](#)

参考資料

NANOG48

- [Add Paths Overview](#)
- [To Add-Paths or Not Add-Paths](#)

RIPE61

- [BGP Add-Paths](#)
- [BGP Diverse Paths](#)

Thank you.

