MEDって曲者？
～BGPベストパス選択の実際～

JANOG19 Meeting
NTTPCコミュニケーションズ
堀　優

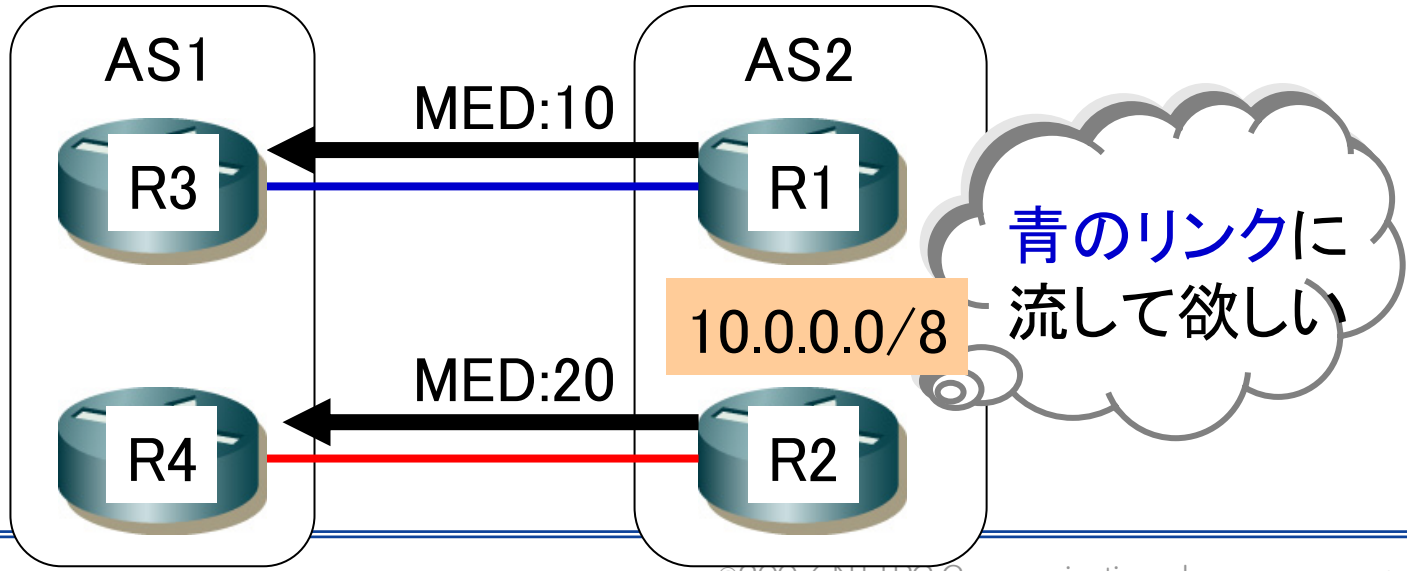
Agenda

1. MEDについて(おさらい)
2. BGPベストパス選択の動作について
3. 通常のアルゴリズムではループする
4. MEDの振舞を変更する2つの実装
5. MEDに関連したRFCs
6. RR/Confederation環境での注意点
7. まとめ

1-1. MEDについて(おさらい)
-議題の主役(曲者？)の紹介

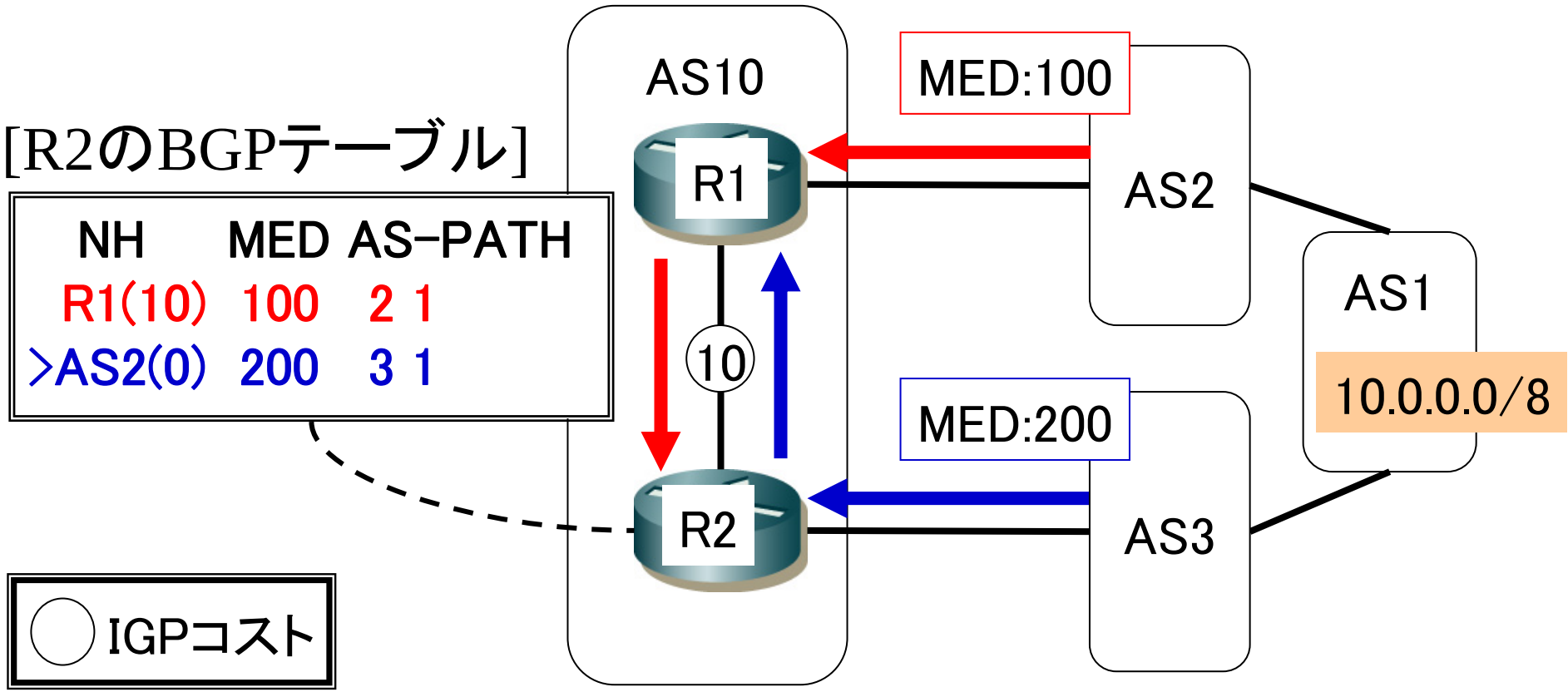
- そもそもMED(MULTI_EXIT_DISC)って何だっけ？
  - -BGPのパス属性の一つ
  - -Optional non-transitive属性
  - -値が小さい方が優先
  - -2つのAS間での接続が複数ある場合に、どのリンクを優先して欲しいかをピア先に通知

※AS2の意図通りトラフィックが流れるかは、AS1のポリシー次第



1-2. MEDについて(おさらい)
-隣接ASの差異により、ベストパス選択時の比較の動作が異なる

・ASごとに管理ポリシーは違うので、隣接ASが異なる場合には、MEDの比較は行われない



2-1. BGPベストパス選択について -経路情報の管理(C社の場合)

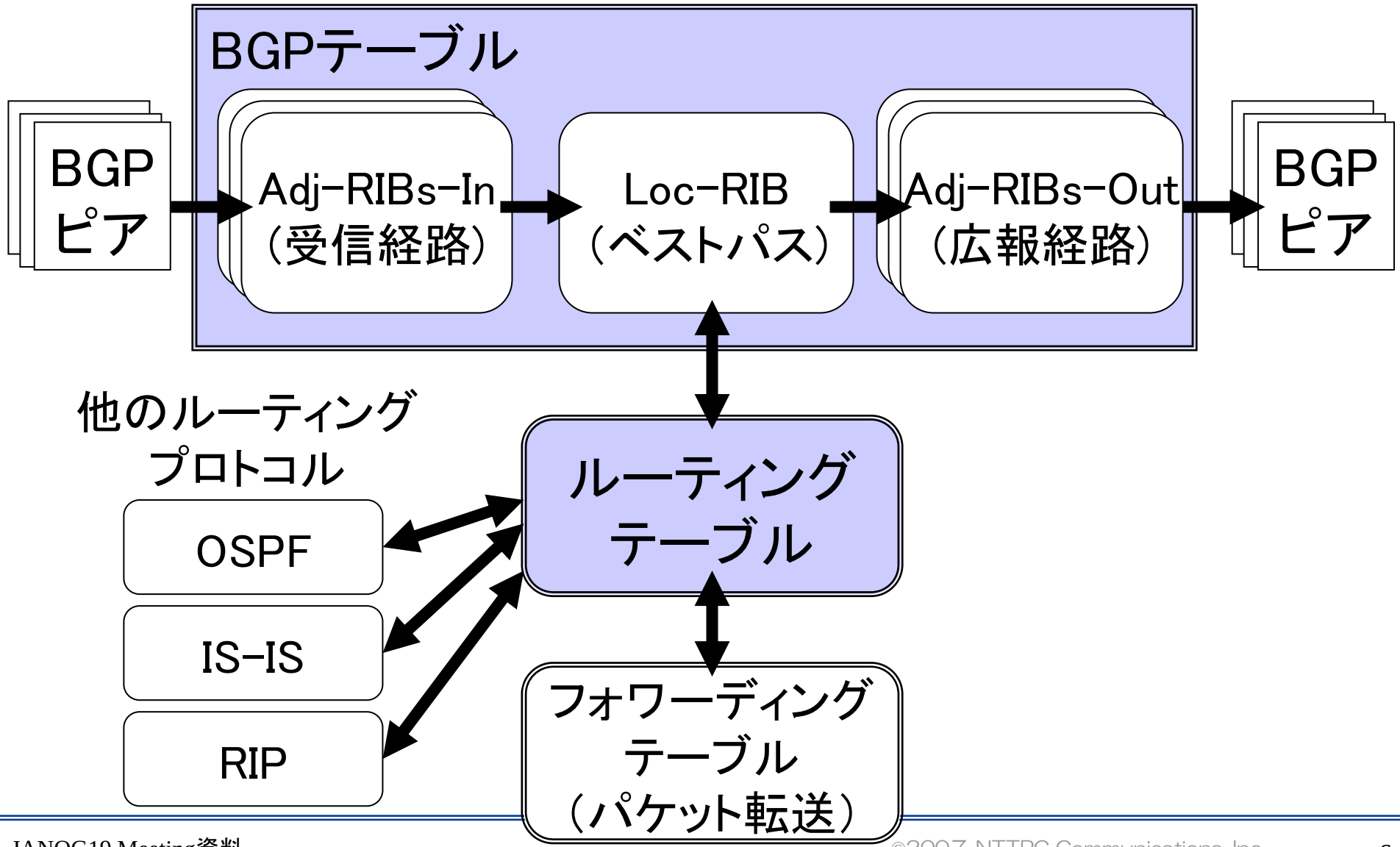
NTT PC COMMUNICATIONS

・2種類のテーブルで経路を管理

- -BGPテーブル
  - -BGPピアから受信した、全経路情報を管理
- -ルーティングテーブル
  - -BGPに限らず、IGPの経路情報も管理
  - -BGPの経路情報に関しては、ベストパスとして選択された経路情報のみを登録・管理する

※J社等、すべて1つのテーブルで管理する実装もある

2-2. BGPベストパス選択について -BGPテーブルとルーティングテーブルの関係



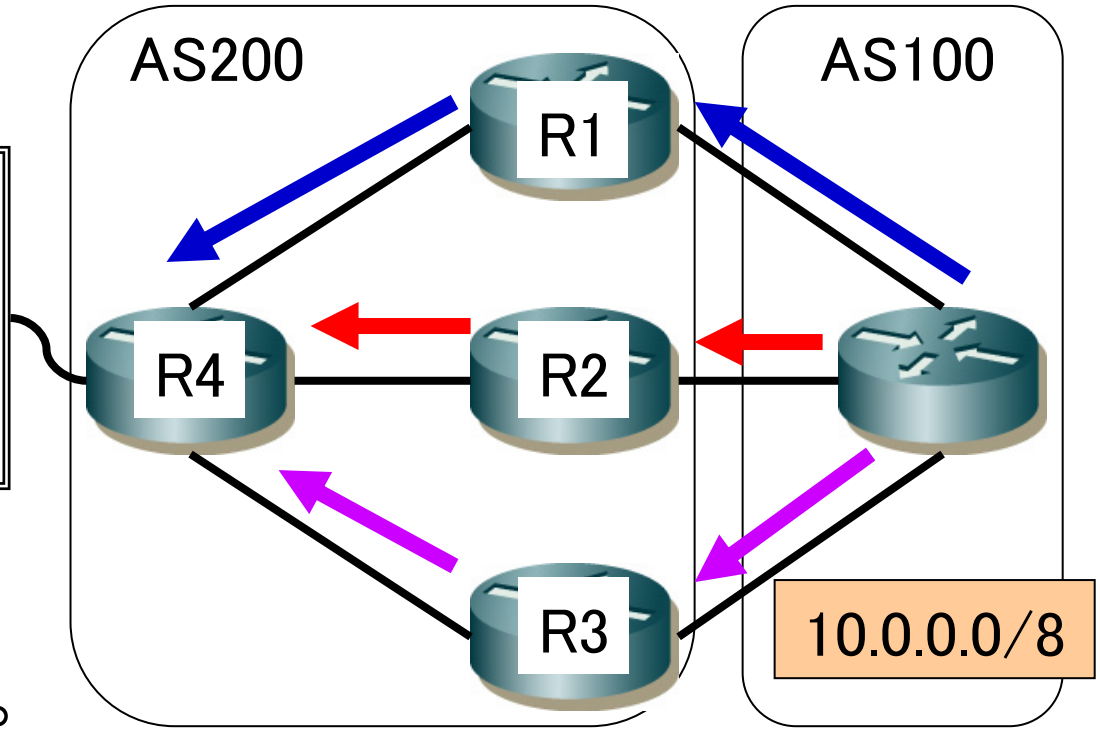
2-3. BGPベストパス選択について -経路情報の登録順序

NTT PC COMMUNICATIONS

・新しく受信した経路情報が、一番上に登録される
・BGPテーブルから削除されない限り、登録順序は変化しない

[R4のBGPテーブル]

Prefix	NH	AS-PATH
10.0.0.0/8	R3(5)	100
10.0.0.0/8	R2(2)	100
10.0.0.0/8	R1(10)	100



○Prefix:10.0.0.0/8は、
R1→R2→R3の順に受信。

2-4. BGPベストパス選択について -ベストパス選択プロセスの動作

NTT PC COMMUNICATIONS

- ベストパス選択は、優先度の高い属性を、一番上のエントリから順番に、タイブレーク方式で比較していく
- 一度ベストパスを決定した後、それ以降の比較は、“best”として認識している経路とのみ比較する

[R4のBGPテーブル]

	Prefix	NH	AS-PATH
	10.0.0.0/8	R3(5)	100
>	10.0.0.0/8	R2(2)	100
	10.0.0.0/8	R1(10)	100

1回目の比較(紫 vs 赤)

2回目の比較(赤[1回目の勝者] vs 青)

※R4は最終的に、赤(R2から)の経路を選択

2-5. BGPベストパス選択について -経路比較の優先順位(C社の場合)

NTT PC COMMUNICATIONS

優先順位	属性	内容
1	WEIGHT	Cisco固有
2	LOCAL_PREF	Local_Pref値の大きい経路を優先
3	LOCAL	Localで生成された経路を優先
4	AS-PATH	AS-PATH長の短い経路を優先
5	ORIGIN	IGP>EGP>INCOMPLETEの順で優先
6	MED	値が小さい経路を優先
7	Peer Type	eBGP>iBGP
8	IGP Metric	ネクストホップへのIGPコストが小さい経路を優先
9	Oldest Path	比較経路がeBGPの場合、古い経路(現状のベストパス)を優先
10	Router-ID	Router-IDが最も小さい経路を優先

※ただし、経路のNEXT_HOP属性への到達性があることが前提で、NEXT_HOPに到達できない経路は無視される。

2-6. BGPベストパス選択について
-評価対象の内容によって、比較動作が異なる

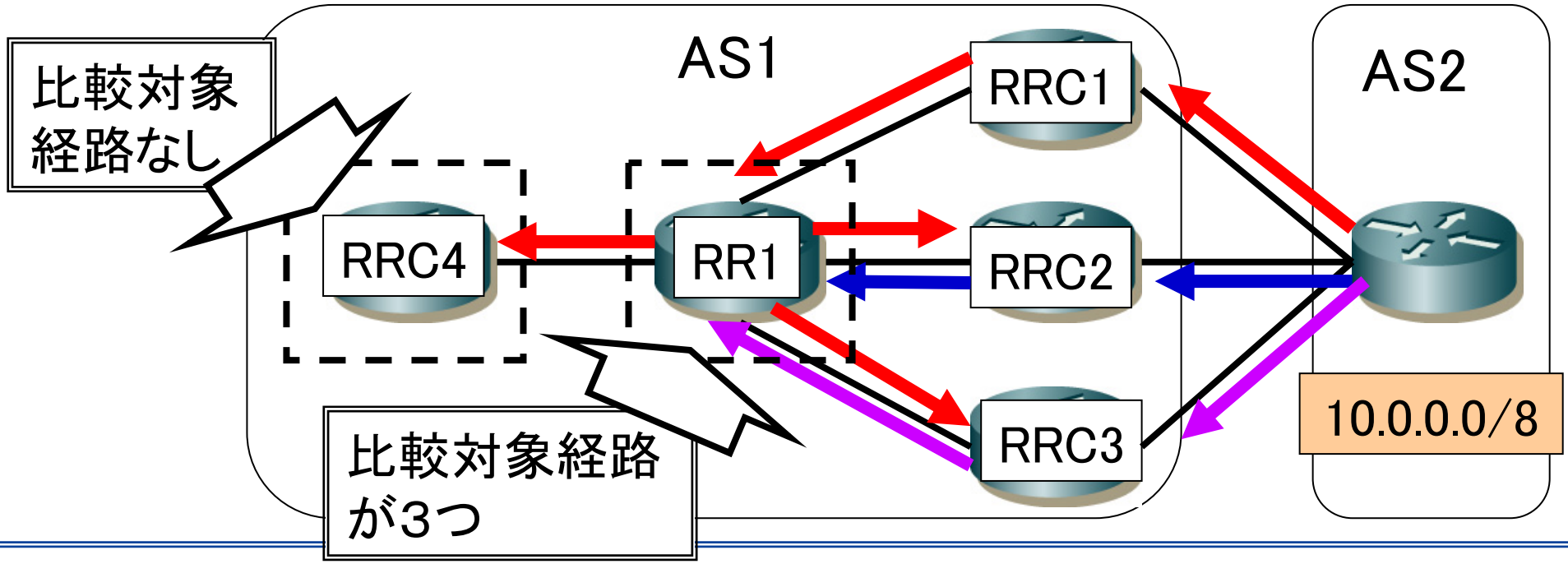
NTT PC COMMUNICATIONS

- 常に同じように比較するもの
  - -LOCAL_PREF
  - -AS-PATH
  - -ORIGIN
- 必ずしも、同じようには比較されない(出来ない)もの
  - -MED
    - -隣接ASが同一の場合は比較し、異なる場合は比較しない
  - -IGP Metric
    - -基本的に各BGPスピーカごとに変化

2-7. BGPベストパス選択について -BGPスピーカは、ベストパスのみ通知する

NTT PC COMMUNICATIONS

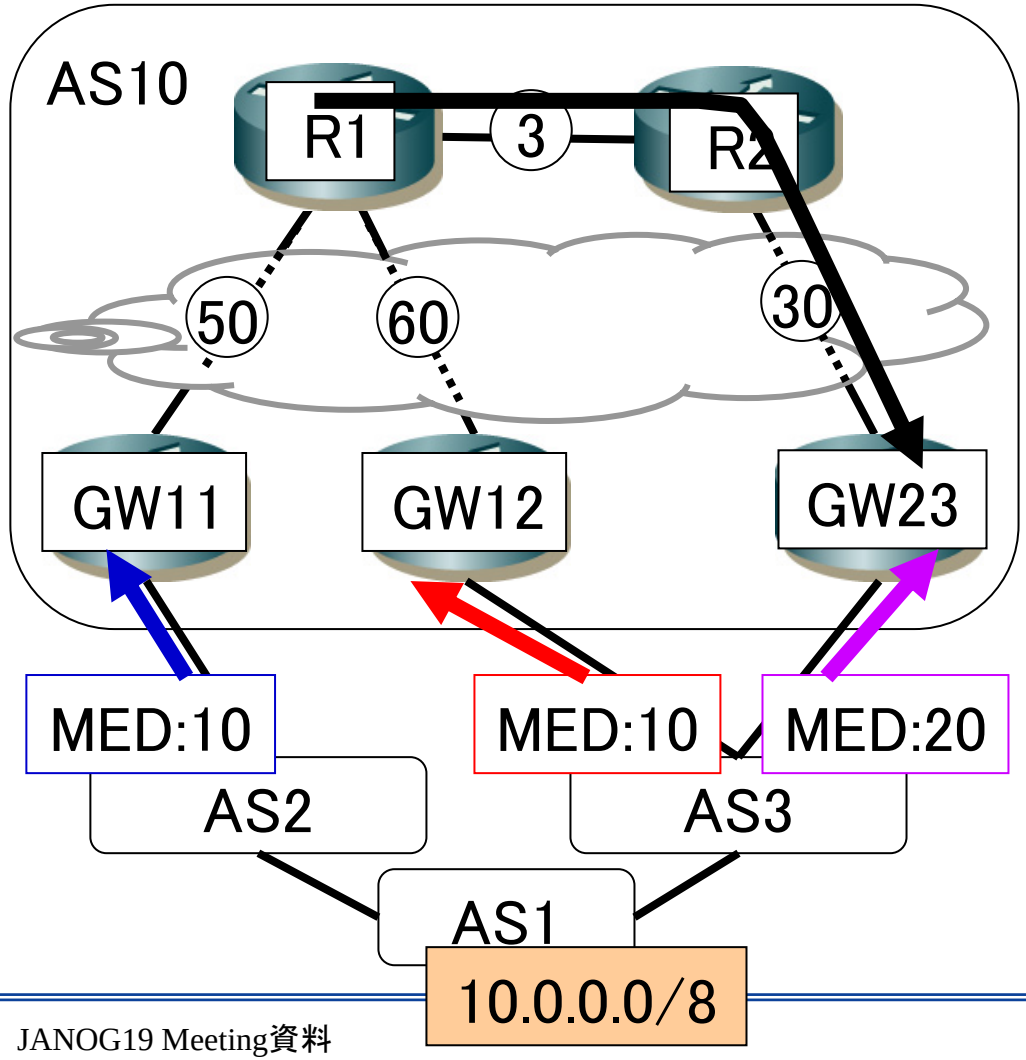
- 比較する経路情報は、それぞれのBGPスピーカごとに変わってくる
- ルートリフレクタや、コンフェデレーションを使用すると、比較する経路情報が少なくなりやすい



3-1. 通常のアルゴリズムではループする -BGPテーブルの登録順序次第でループは起きる

NTT PC COMMUNICATIONS

・始めは問題なかった。。。



[R1のBGPテーブル]

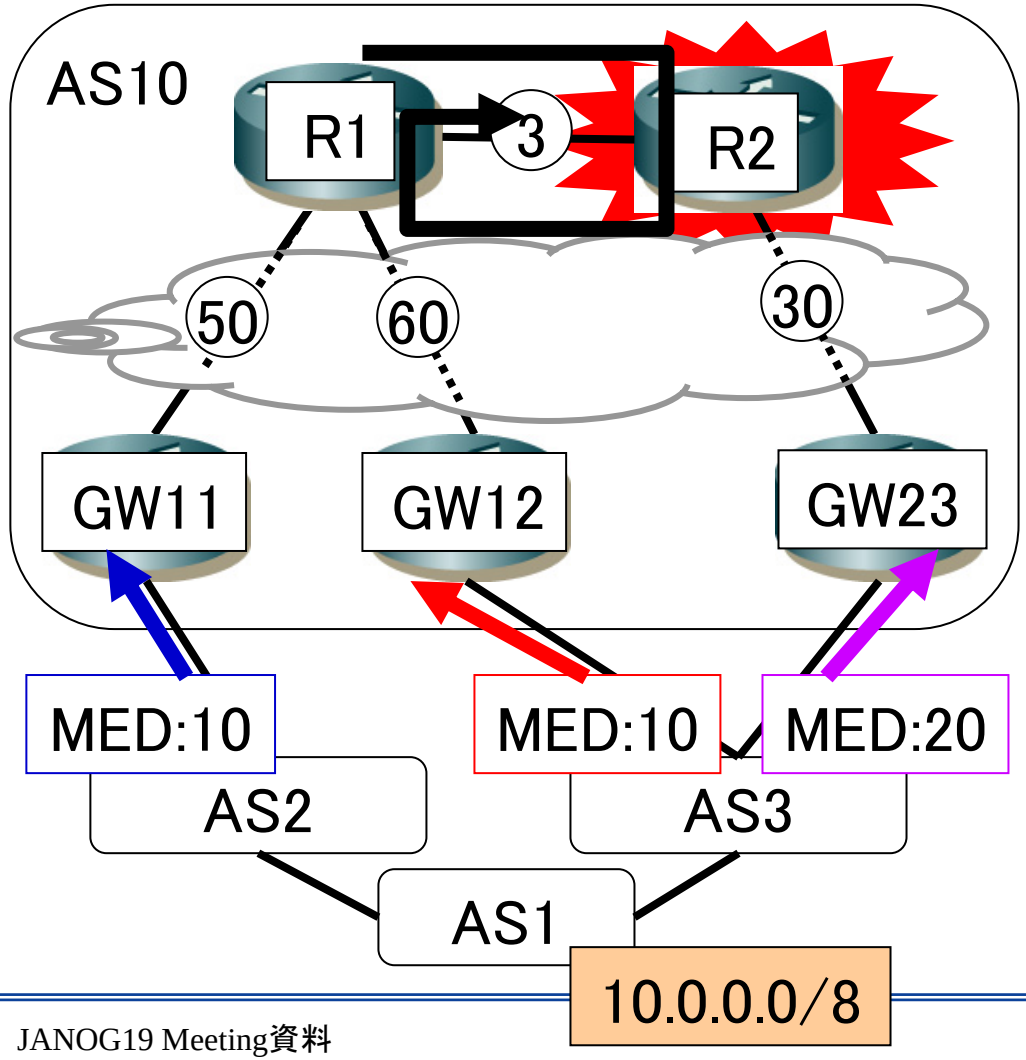
	NH	MED	AS-PATH
	GW11(50)	10	2 1
	GW12(60)	10	3 1
>	GW23(33)	20	3 1

[R2のBGPテーブル]

	NH	MED	AS-PATH
	GW11(53)	10	2 1
	GW12(63)	10	3 1
>	GW23(30)	20	3 1

3-2. 通常のアルゴリズムではループする -BGPテーブルの登録順序次第でループは起きる-続き

・R2がRebootし、BGPテーブルの登録順序が変わると・・・。



[R1のBGPテーブル]

	NH	MED	AS-PATH
	GW11(50)	10	2 1
	GW12(60)	10	3 1
>	GW23(33)	20	3 1

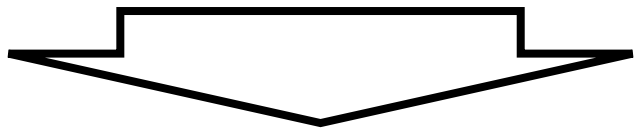
[R2のBGPテーブル]

	NH	MED	AS-PATH
	GW23(30)	20	3 1
	GW11(53)	10	2 1
>	GW12(63)	10	3 1

3-3. 通常のアルゴリズムではループする –何が問題でループが起きるか？

NTT PC COMMUNICATIONS

- 以下の2つの動作が、同時に働くためにBGPテーブルの登録順序次第で、パケットループを引き起こすことがある
  - –BGPテーブル上に、複数の経路情報が存在した場合でも、比較の対象はベストパスの経路とのみ
  - –隣接ASの差異によって、MED値の扱い方が変化



MEDの存在により、タイブレークルールが完全には行われなくなっている

4-1. MEDの振舞を変更する2つの実装
-always-compared-medと、deterministic-med

NTT PC COMMUNICATIONS

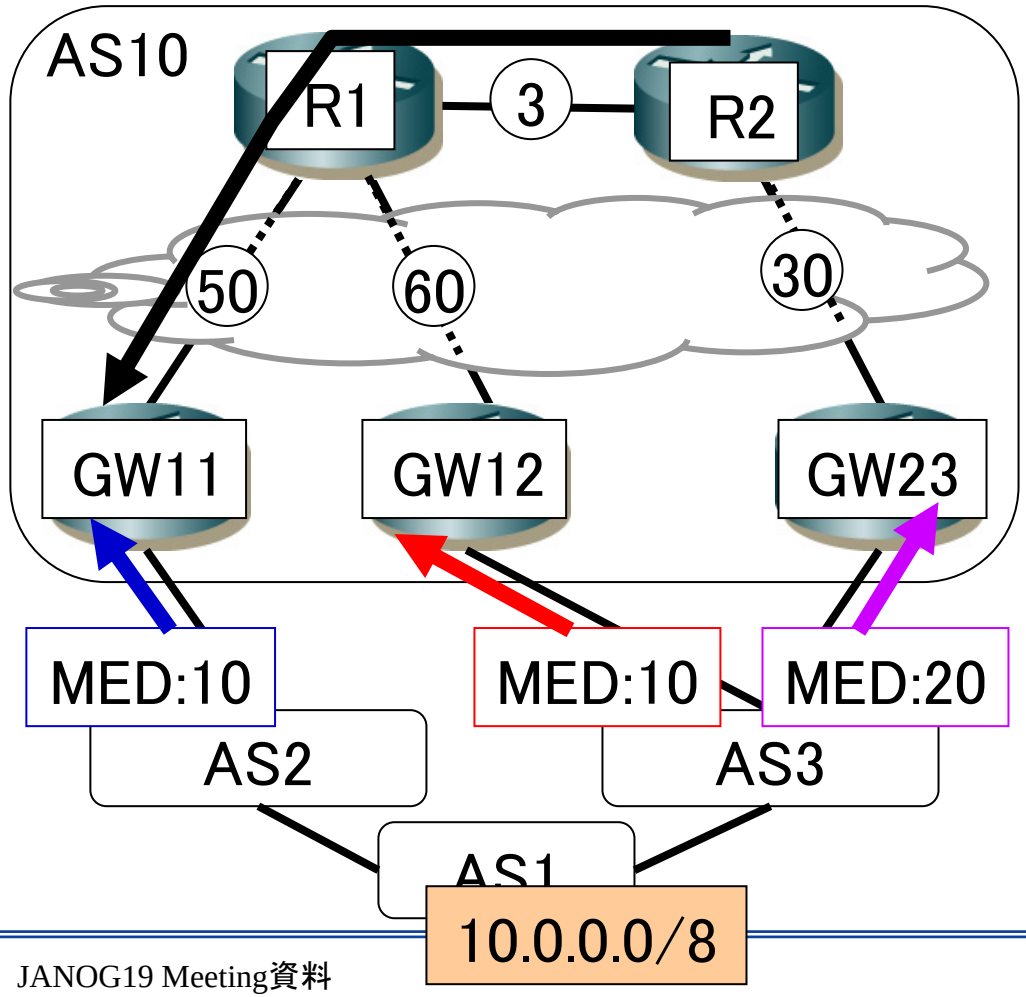
- “bgp always-compare-med”とは？
  - -隣接ASの差異に関係なく、MED値を常に比較

- ”bgp deterministic-med”とは？
  - -隣接ASごとに比較対象の経路をグループ化し、同一AS内でのベストとなる経路を選択した後、違うASグループでベストとなる経路と比較する
  - -隣接ASごとの経路のグループ化は、すべて内部処理で行われる

※ベストパス選択アルゴリズムが変わるため、AS内のすべてのBGPスピーカに、同一の設定が必要になる

4-2. ループしないようにするためには？
-MEDの振舞を変更する#1(bgp always-compare-med)

- “bgp always-compare-med”を有効にすると・・・
  - -常にMEDを比較する以外に変更なし



[R1のBGPテーブル]

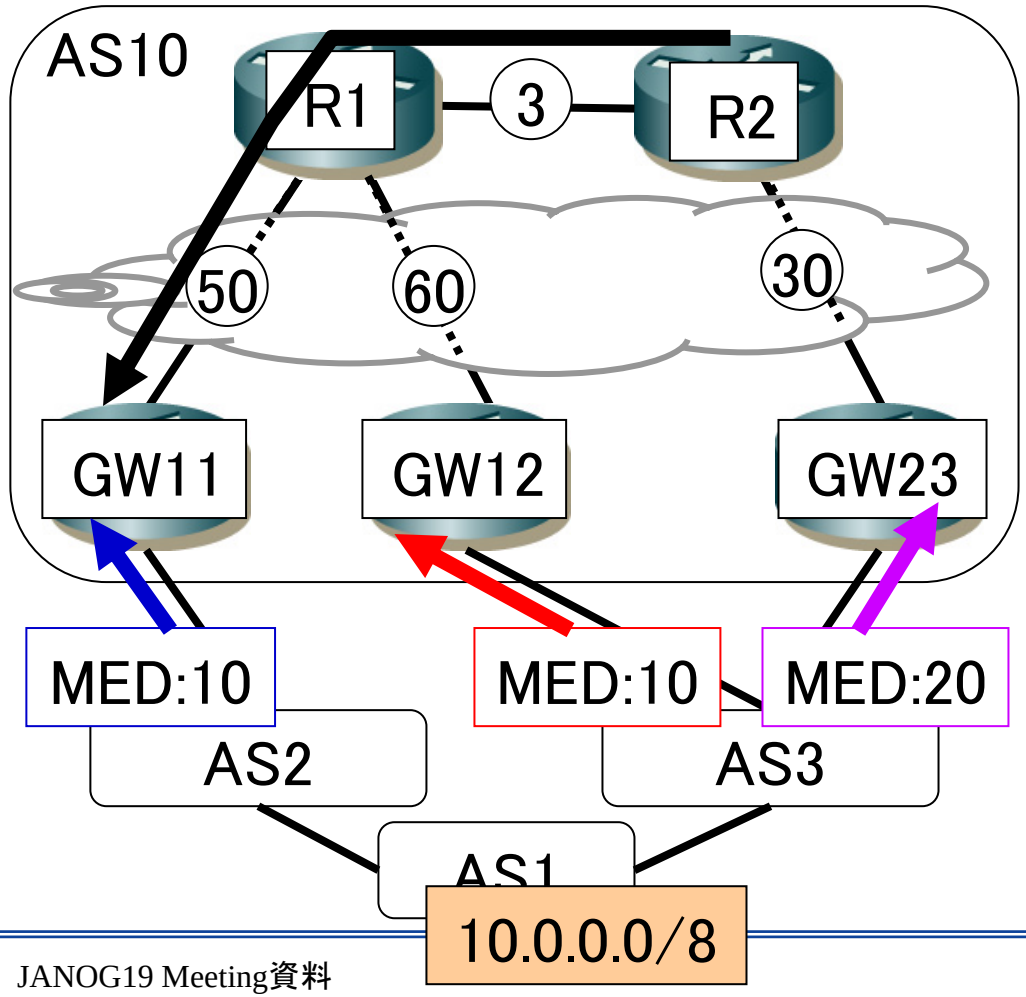
	NH	MED	AS-PATH
>	GW11(50)	10	2 1
	GW12(60)	10	3 1
	GW23(33)	20	3 1

[R2のBGPテーブル]

	NH	MED	AS-PATH
	GW23(30)	20	3 1
>	GW11(53)	10	2 1
	GW12(63)	10	3 1

4-3. ループしないようにするためには？
-MEDの振舞を変更する#2(bgp deterministic-med)

- “bgp deterministic-med”を有効にすると・・・
  - -BGPテーブルは、登録順序のままの表示(わかりにくい)



[R1のBGPテーブル]

	NH	MED	AS-PATH
>	GW11(50)	10	2 1
	GW12(60)	10	3 1
	GW23(33)	20	3 1

[R2のBGPテーブル]

	NH	MED	AS-PATH
	GW23(30)	20	3 1
>	GW11(53)	10	2 1
	GW12(63)	10	3 1

※R1,R2とも以下の動作
(赤と紫の勝者) vs 青

5-1. MEDに関連したRFCs
-RFC3345と、RFC4451（ともにInformational）

RFC3345・・・Border Gateway Protocol Persistent Route Oscillation Condition

・RR/Confederation環境で発生しうる、BGPの経路が収束しない問題と、その対処策

RFC4451・・・BGP MULTI_EXIT_DISC Considerations

・MEDの変更/削除と、経路選択プロセスの動作
・missing-MEDの扱い
　-なし、0、2^32-1問題
・IGPコストを、MEDに反映した場合のdisadvantage
　-Route Flap Damping、Update Packingへの影響

5-2. MEDに関連したRFCs
-RFC3345のType 1,2 Chrunとは？

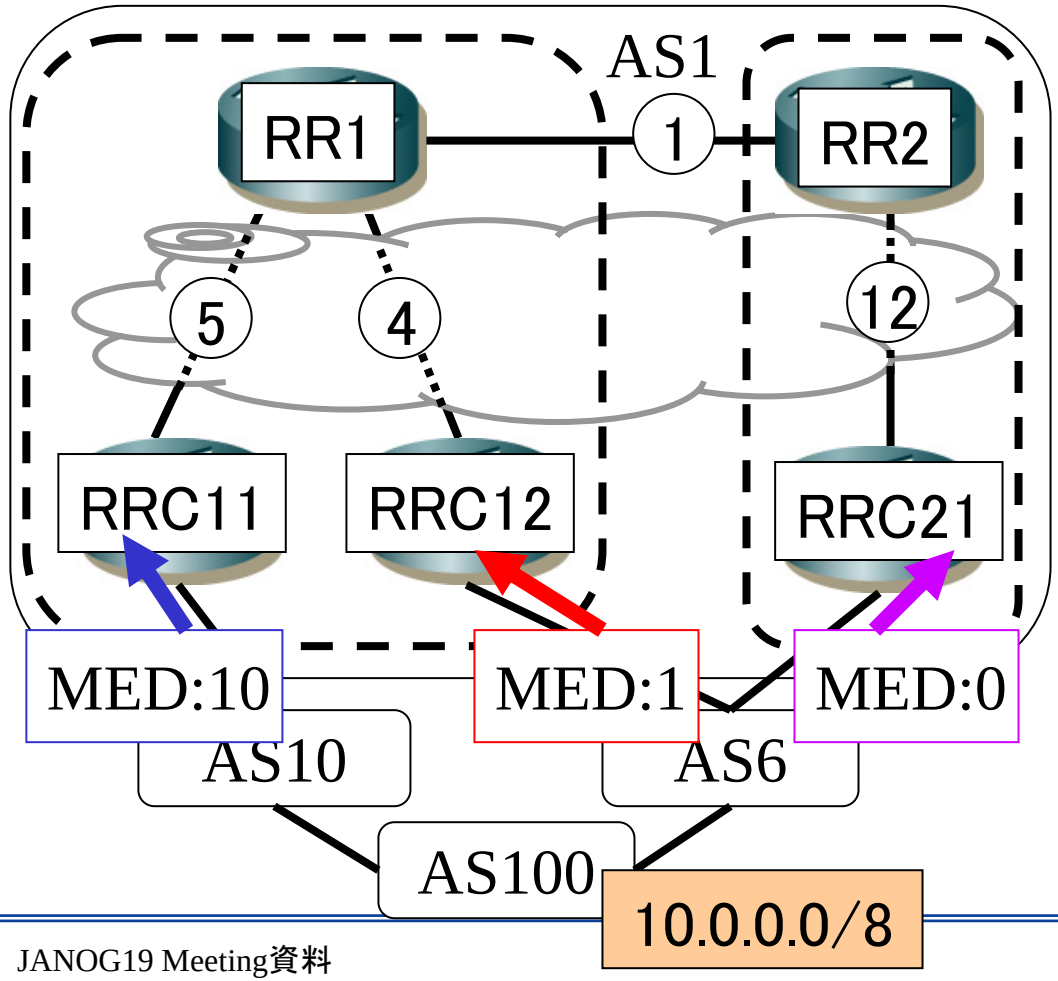
RFC3345では、BGPの経路情報が激しく変動する状況を、2つに分類。

- Type 1 Chrunの条件(AND)
  - 1つのPrefixを、2つ以上のASから受信
  - 任意のMED値を許可、または設定
  - 階層化していないRR/Confederation環境
- Type 2 Chrunの条件(AND)
  - 1つのPrefixを、2つ以上のASから受信
  - 任意のMED値を許可、または設定
  - 階層化したRR/Confederation環境

6-1. RR/Confederation環境での注意点 -RFC3345 Type 1 Churnについて

NTT PC COMMUNICATIONS

・“bgp deterministic-med”が有効でも、RR/Confederation環境では、ルーティングループが発生することがある。



○Type 1 Churnの条件(AND)

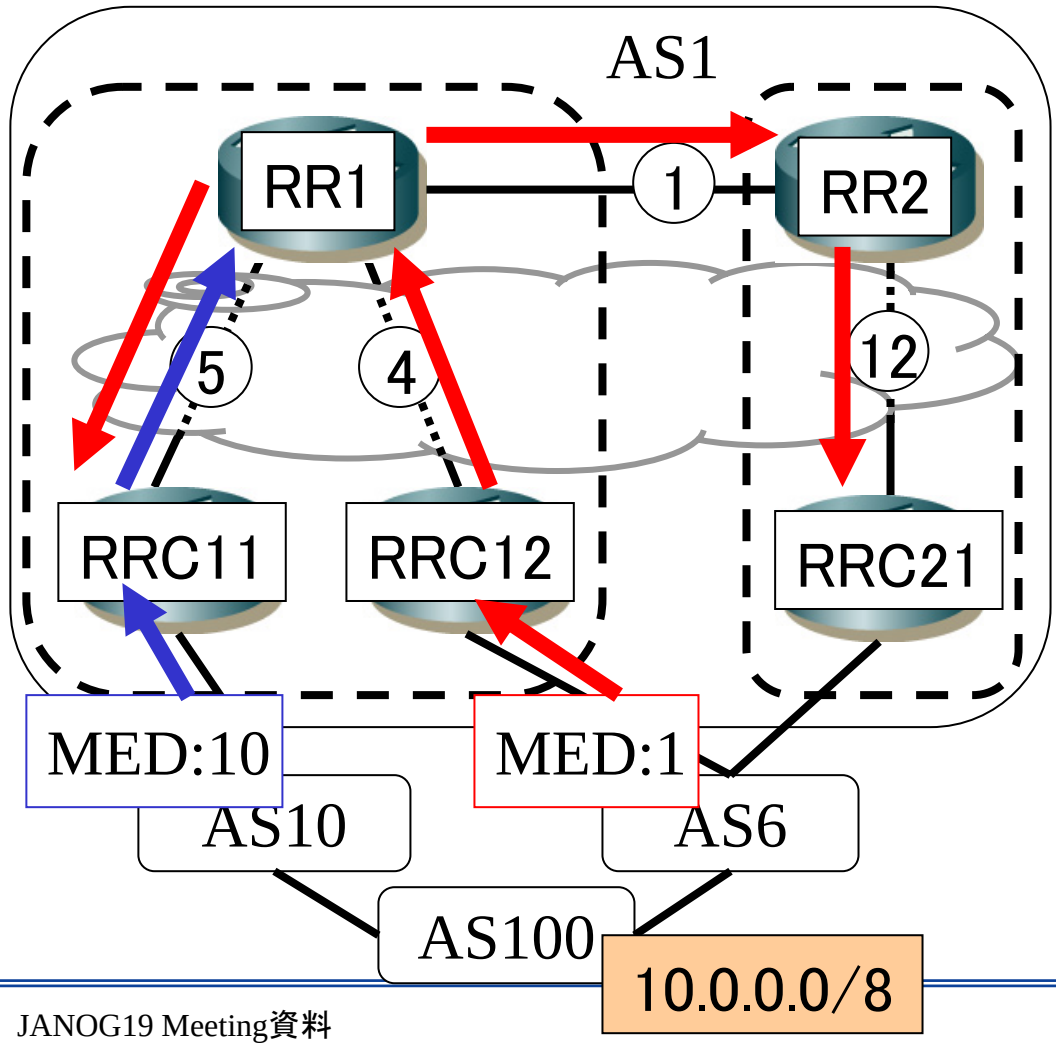
①RR/Confederationが、シングルレベル(階層化されていない)

②1つのPrefixを、2つ以上のASから受信し、任意のMEDを許可

6-2. RR/Confederation環境での注意点 -RFC3345 Type 1 Churnについて-続き

NTTPC COMMUNICATIONS

・始めの状態



RR1)sh ip bgp 10.0.0.0/8

NH	MED	AS-PATH
>RRC12(4)	1	6 100
RRC11(5)	10	10 100

RR2)sh ip bgp 10.0.0.0/8

NH	MED	AS-PATH
>RRC12(5)	1	6 100

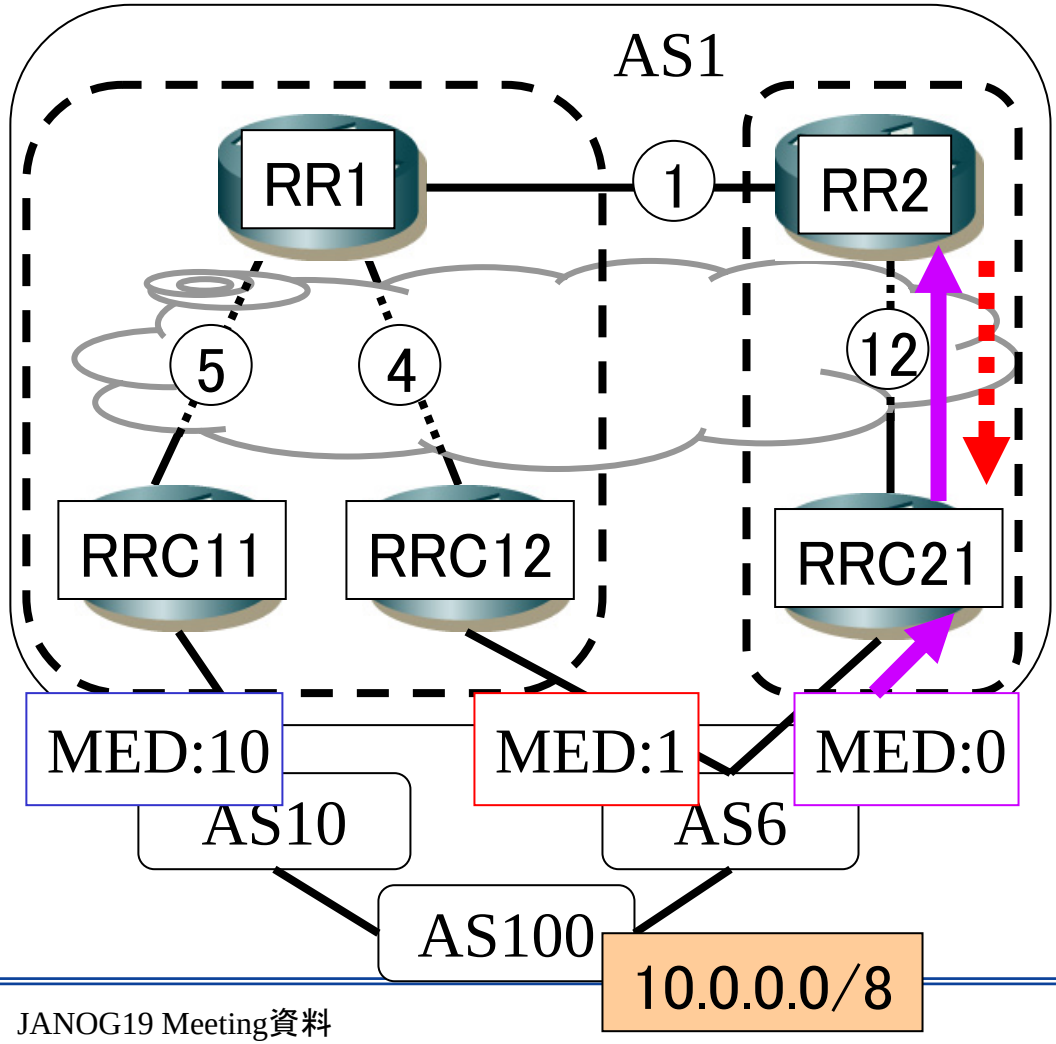
RRC21)sh ip bgp 10.0.0.0/8

NH	MED	AS-PATH
>RRC12(17)	1	6 100

6-3. RR/Confederation環境での注意点 -RFC3345 Type 1 Churnについて-続き

NTT PC COMMUNICATIONS

・RRC21が新しくAS6から紫の経路情報を受信すると・・・



RR2)sh ip bgp 10.0.0.0/8

NH	MED	AS-PATH
>RRC21(12)	0	6 100
RRC12(5)	1	6 100

RRC21)sh ip bgp 10.0.0.0/8

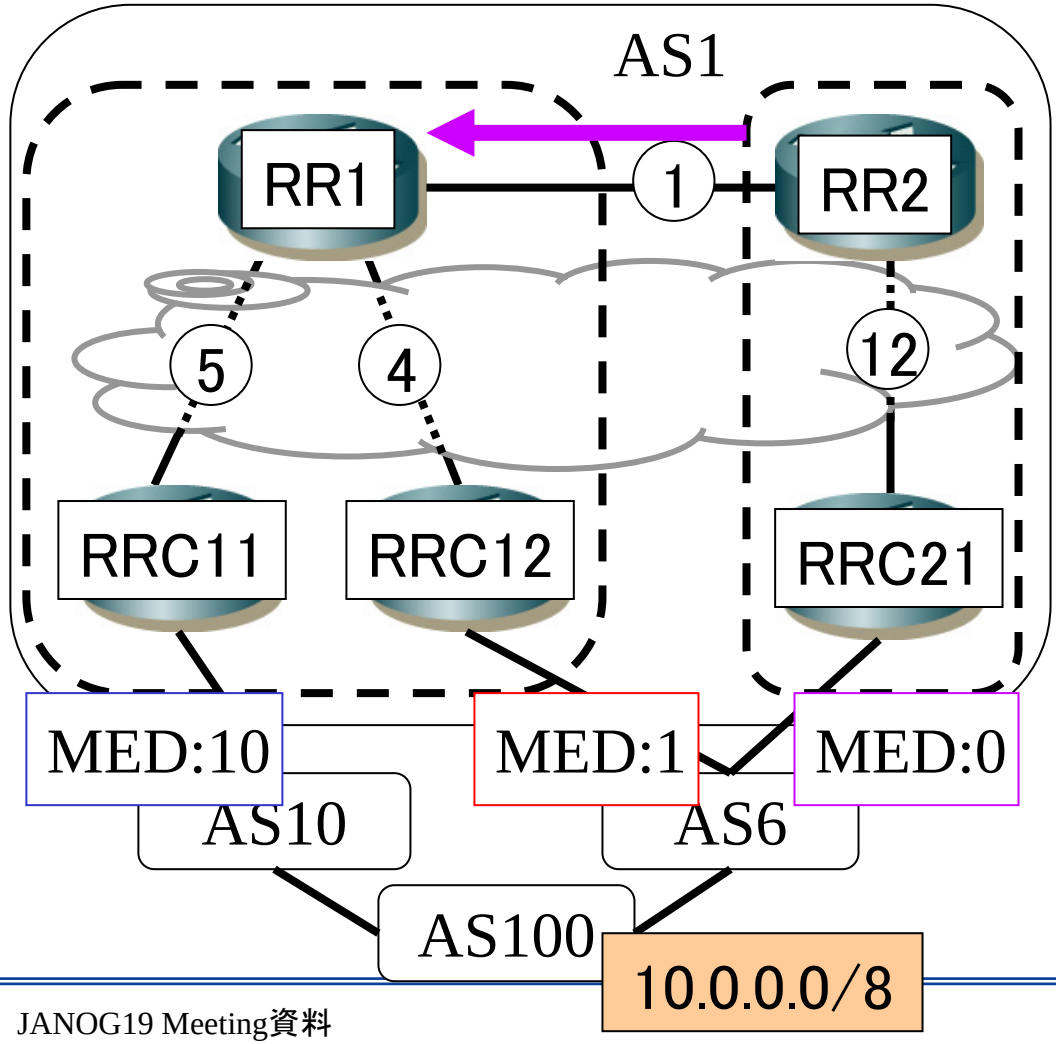
NH	MED	AS-PATH
>RRC21(0)	0	6 100
~~RRC12(17)~~	~~1~~	~~6 100~~

RRC21,RR2共に
紫の経路がベストパスに！

6-4. RR/Confederation環境での注意点 -RFC3345 Type 1 Churnについて-続き

NTT PC COMMUNICATIONS

・RR2が、RR1へ新しいベストパス(紫の経路)をUpdateすると



RR1)sh ip bgp 10.0.0.0/8

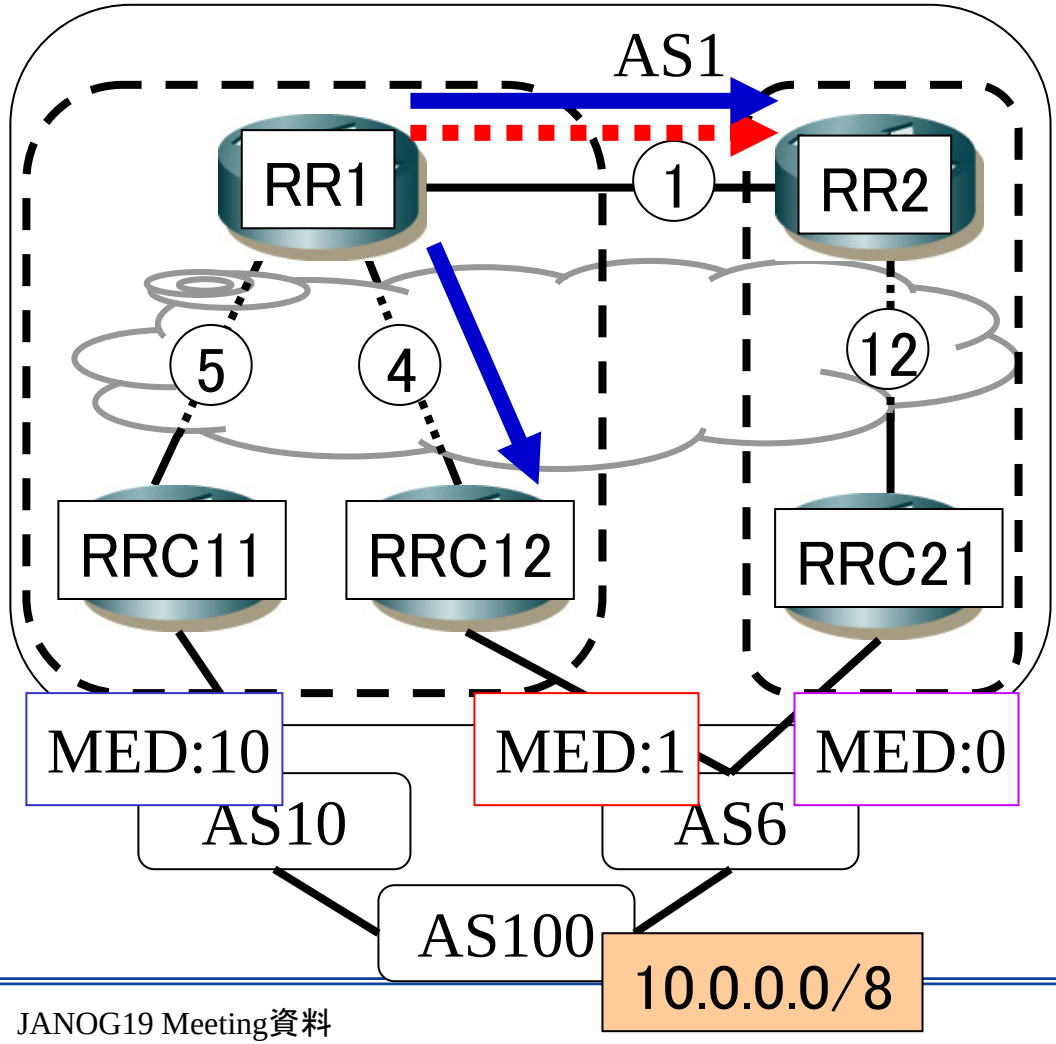
	NH	MED	AS-PATH
	RRC21(13)	0	6 100
	RRC12(4)	1	6 100
>RRC11(5)		10	10 100

1回目の比較：紫 vs 赤
-MEDにより、紫の勝利
2回目の比較：紫 vs 青
-IGPコストにより、青の勝利
青の経路がベストパス

6-5. RR/Confederation環境での注意点 -RFC3345 Type 1 Churnについて-続き

NTT PC COMMUNICATIONS

・RR1のベストパスが青の経路に変化すると



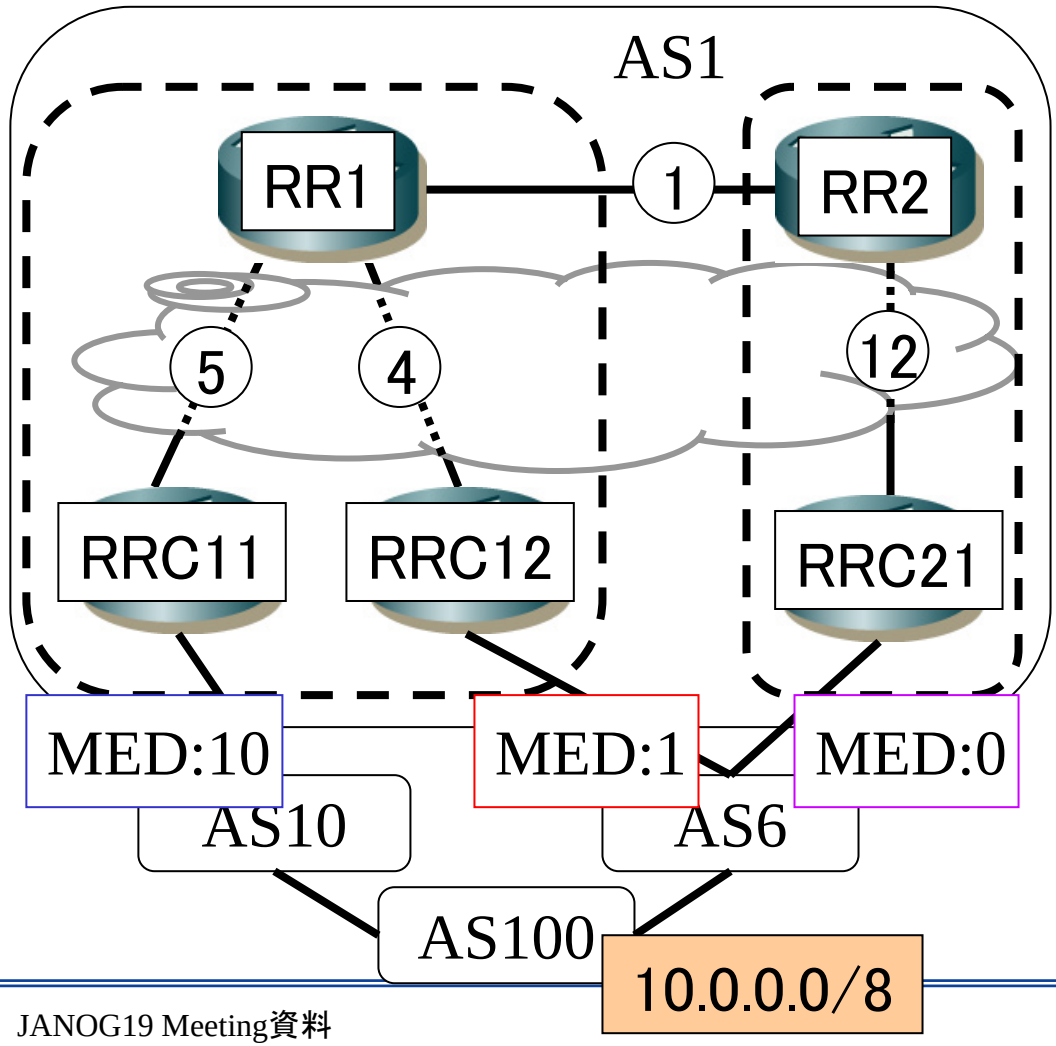
RR1)sh ip bgp 10.0.0.0/8

NH	MED	AS-PATH
RRC21(13)	0	6 100
RRC12(4)	1	6 100
>RRC11(5)	10	10 100

・新しくベストパスになった青の経路をUpdate
・旧のベストパスである赤の経路をWithdrawn

6-6. RR/Confederation環境での注意点 -RFC3345 Type 1 Churnについて-続き

・RR2が、RR1からのUpdate/Withdrawnを受信すると・・・



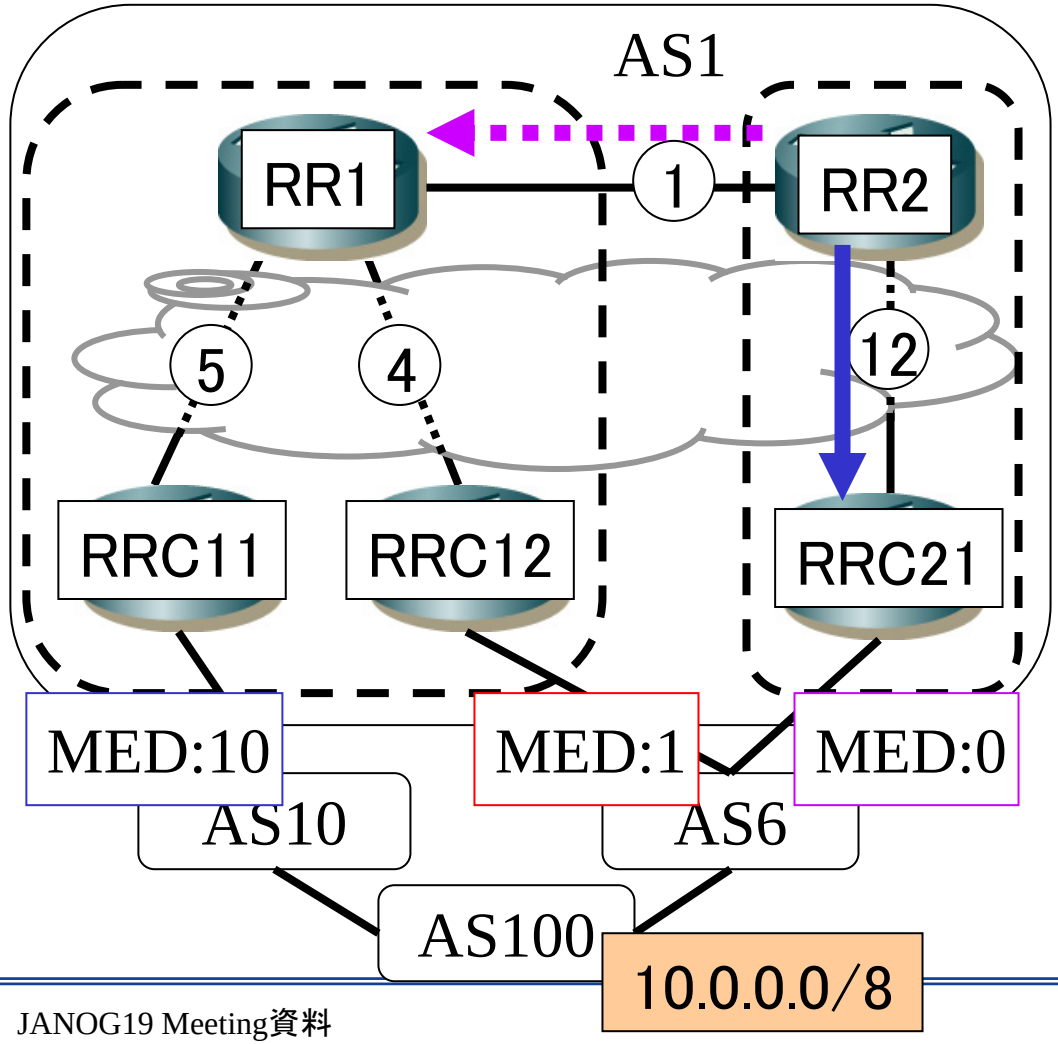
RR2)sh ip bgp 10.0.0.0/8

	NH	MED	AS-PATH
>RRC11(6)		10	10 100
RRC21(12)		0	6 100
~~RRC12(5)~~		~~1~~	~~6 100~~

青の経路がベストパスに！

6-7. RR/Confederation環境での注意点 -RFC3345 Type 1 Churnについて-続き

・RR2のベストパスが変更されると・・・



RR2)sh ip bgp 10.0.0.0/8

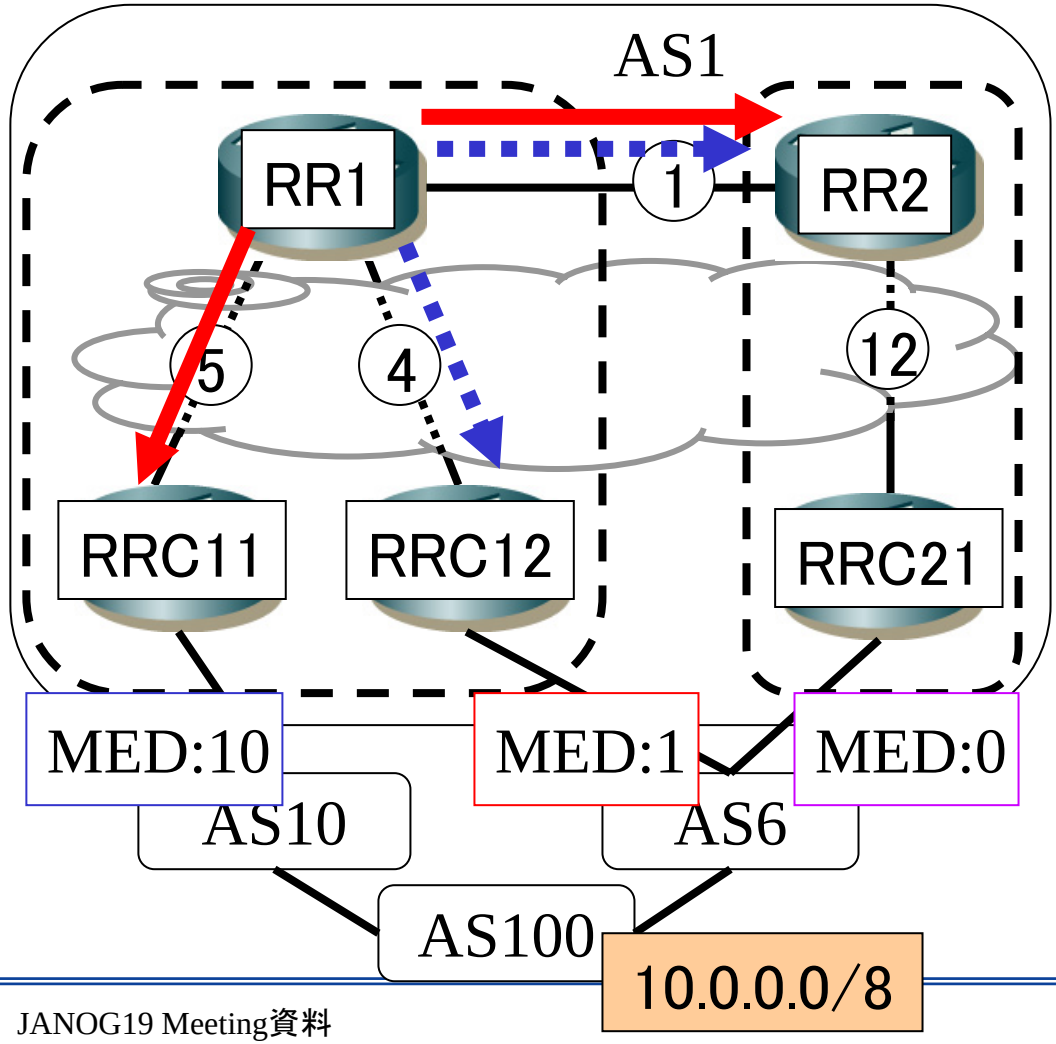
	NH	MED	AS-PATH
>	RRC11(6)	10	10 100
	RRC21(12)	0	6 100

・青の経路をUpdate
・紫の経路をWithdrawn

6-8. RR/Confederation環境での注意点 -RFC3345 Type 1 Churnについて-続き

NTT PC COMMUNICATIONS

・RR1が、紫の経路をwithdrawnされると・・・



RR1)sh ip bgp 10.0.0.0/8		
NH	MED	AS-PATH
~~RRC21(13)~~	~~0~~	~~6 100~~
>RRC12(4)	1	6 100
RRC11(5)	10	10 100

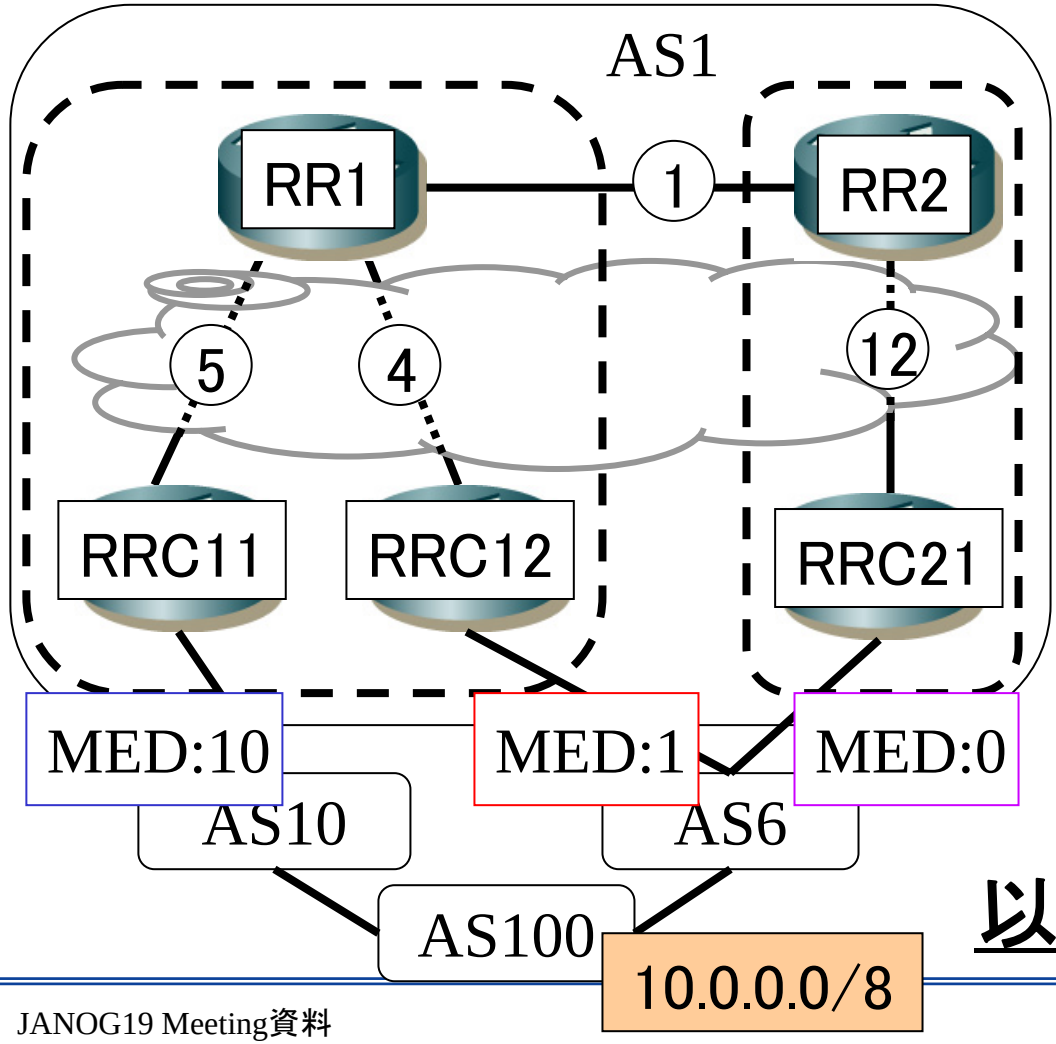
赤の経路がベストパスに！

・赤の経路をUpdate
・青の経路をwithdrawn

6-9. RR/Confederation環境での注意点 -RFC3345 Type 1 Churnについて-続き

NTT PC COMMUNICATIONS

・RR2がRR1からのUpdate/Withdrawnを受信すると・・・



RR2)sh ip bgp 10.0.0.0/8

	NH	MED	AS-PATH
	RRC12(5)	1	6 100
	~~RRC11(6)~~	~~10~~	~~10 100~~
>	RRC21(12)	0	6 100

・MEDにより、紫の経路をベストパスに選択

以降、endless loopは続く

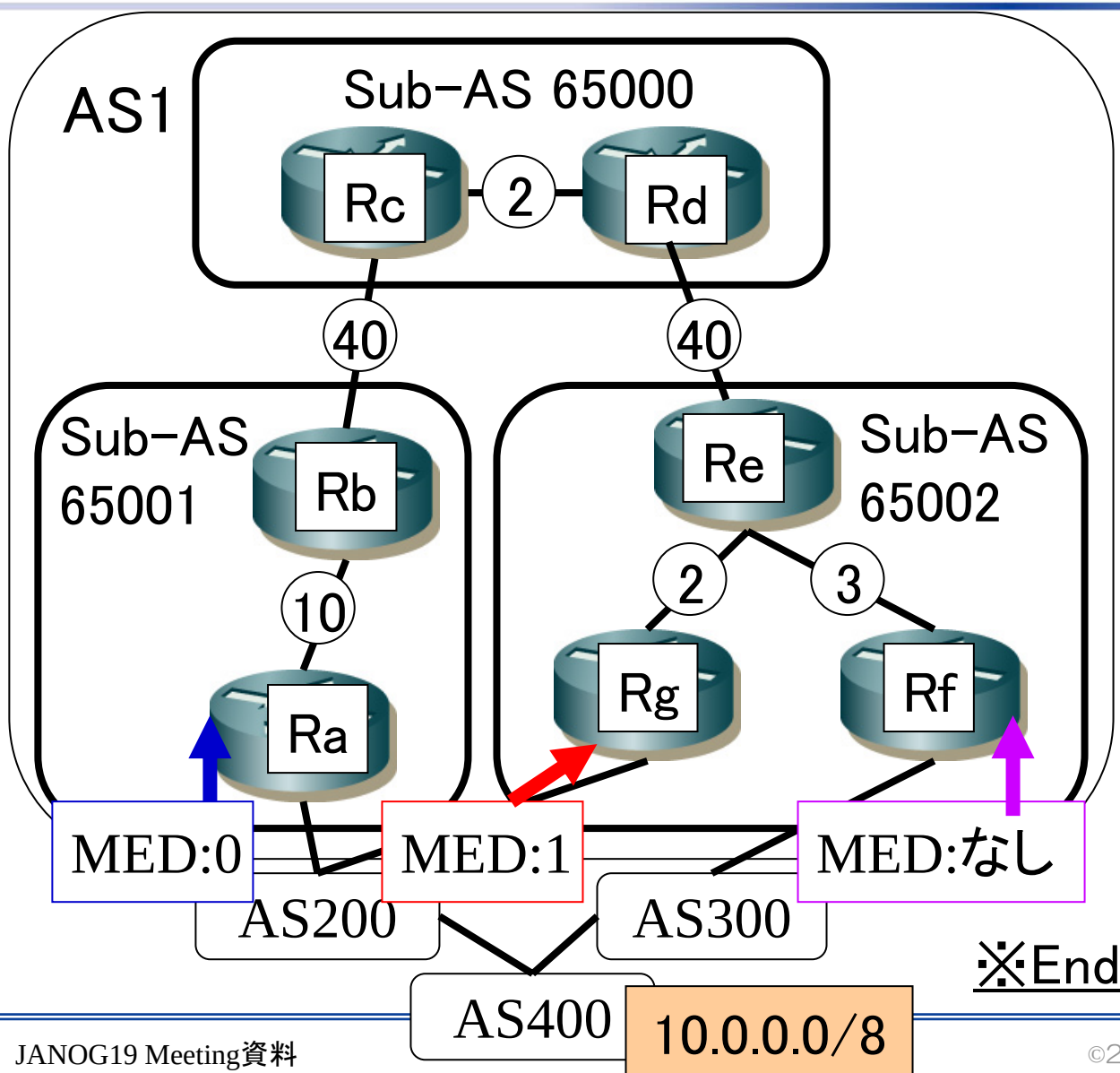
6-10. RR/Confederation環境での注意点 -RFC3345 Type 1 Churnについて-続き

NTT PC COMMUNICATIONS

・Type 1 Churn問題を回避する方法

1. RR環境の場合
   -cluster間IGPメトリック > cluster内IGPメトリック
2. Confederation環境の場合
   -sub-AS間IGPメトリック > sub-AS内IGPメトリック
3. MEDを評価させない
   -LOCAL_PREF
   -MEDをすべて同じ値に上書き
4. “bgp always-compare-med”を使用する
5. iBGPをフルメッシュに張る(?!!!)

6-11. RR/Confederation環境での注意点 -RFC3345 Type 2 Churnについて



○Type 2 Churnの条件 (AND)

①RR/Confederationが、階層化されている

②1つのPrefixを、2つ以上のASから受信し、任意のMEDを許可

※Endless loopの動作は省略

6-12. RR/Confederation環境での注意点 -RFC3345 Type 2 Churnについて-続き

NTT PC COMMUNICATIONS

・Type 2 Churn問題を回避する方法

1. MEDを評価させない
   -LOCAL_PREF
   -MEDをすべて同じ値に上書き
2. “bgp always-compare-med”を使用する
3. RR/Confederationを階層化しない(?!!!)
4. RR環境では、RRCとのiBGPをフルメッシュ
5. Confederation環境では、同じ階層にいる
   ボーダルータ間で、BGPをフルメッシュ
   （Type 2 Churnの例で、RbとRe間にeBGPを張る）

6-13. RR/Confederation環境での注意点 -RFC3345 Type 1,2 Churnは何故起きる？

以下の3つの特性が、Endless loopを引き起こす。

- 隣接ASが同じでないと、MEDは評価しない
- BGPプロトコルは、ベストパスしか通知しない
- RR/Confederation環境下では、経路選択に使用した情報が、Cluster/Sub-AS内に隠される
  →階層化すると、より顕著に！

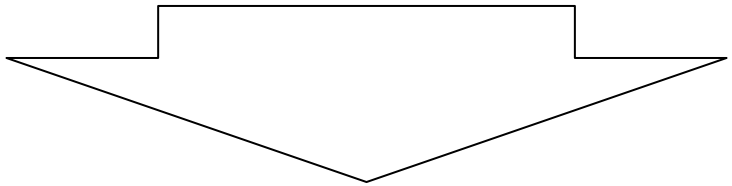
やっぱり、MEDって曲者？

7-1. まとめ
-BGPをスケールする環境に関係なく、有効な手段とは？

- MEDを評価させない#1
  - -LOCAL_PREF
    - MEDを含め、他の属性が評価されなくなる
- MEDを評価させない#2
  - -ネットワークの入り口で、MEDを同じ値に上書き
    - MED以外の属性はすべて評価される
- “bgp always-compare-med”を使用する
  - -MEDを書き換えている(ほとんどのISP？)場合に、最も運用しやすく、現実的？
- Tunnel技術と併用する
  - -遠いところを近く見せる(隠れた経路を見せてやる)。

7-2. まとめ
-MEDの振舞は、すぐには変えられない

通常のアルゴリズムから、“bgp deterministic-med”等、MEDの振る舞いを変更しようにも、現用のNWでは、各BGPスピーカが、別のベストパス選択アルゴリズムを動作させると、余計にループを引き起こすことになる。



①MEDの評価をさせないようにする
②各BGPスピーカの、ベストパス選択アルゴリズムを変更する
③新しいポリシーを、NWに展開する

7-3. まとめ
-MEDってどう運用すべき?

BGPを運用していく上で、MEDに振り回されないためのベストプラクティスについて、議論しましょう。

- MEDの扱い
  - 使う/使わない
  - ピアからのMEDを受け入れる/書き換える
  - 隣接ASの差異を考慮する/しない
- BGPをスケールしていく時の注意点は？
  - iBGP full-mesh/RR/Confederation
  - IGPとの関わり
- その他、思わぬ落とし穴とか・・・