サービスプロバイダ バックボーン設計入門

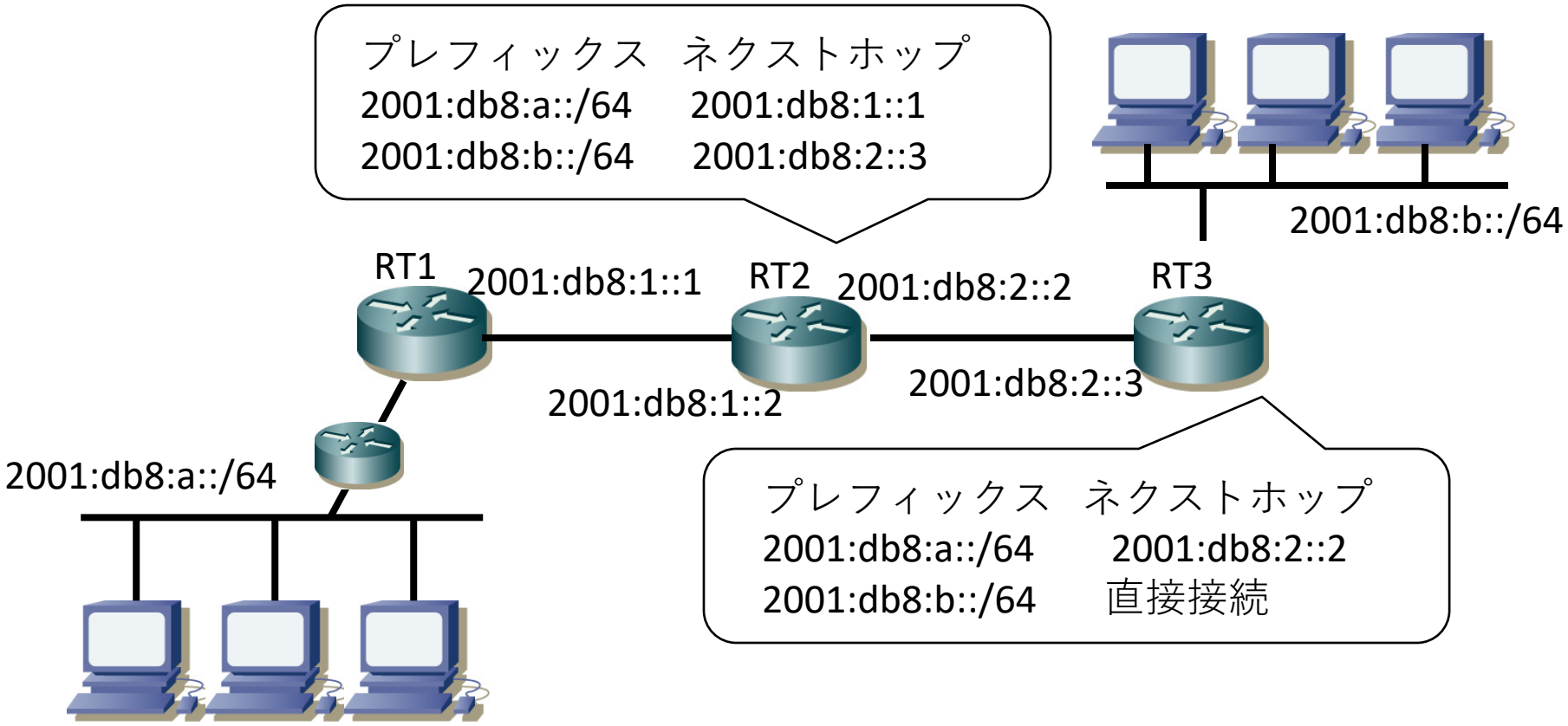
BGP概論

Matsuzaki ‘maz’ Yoshinobu

<maz@iij.ad.jp>

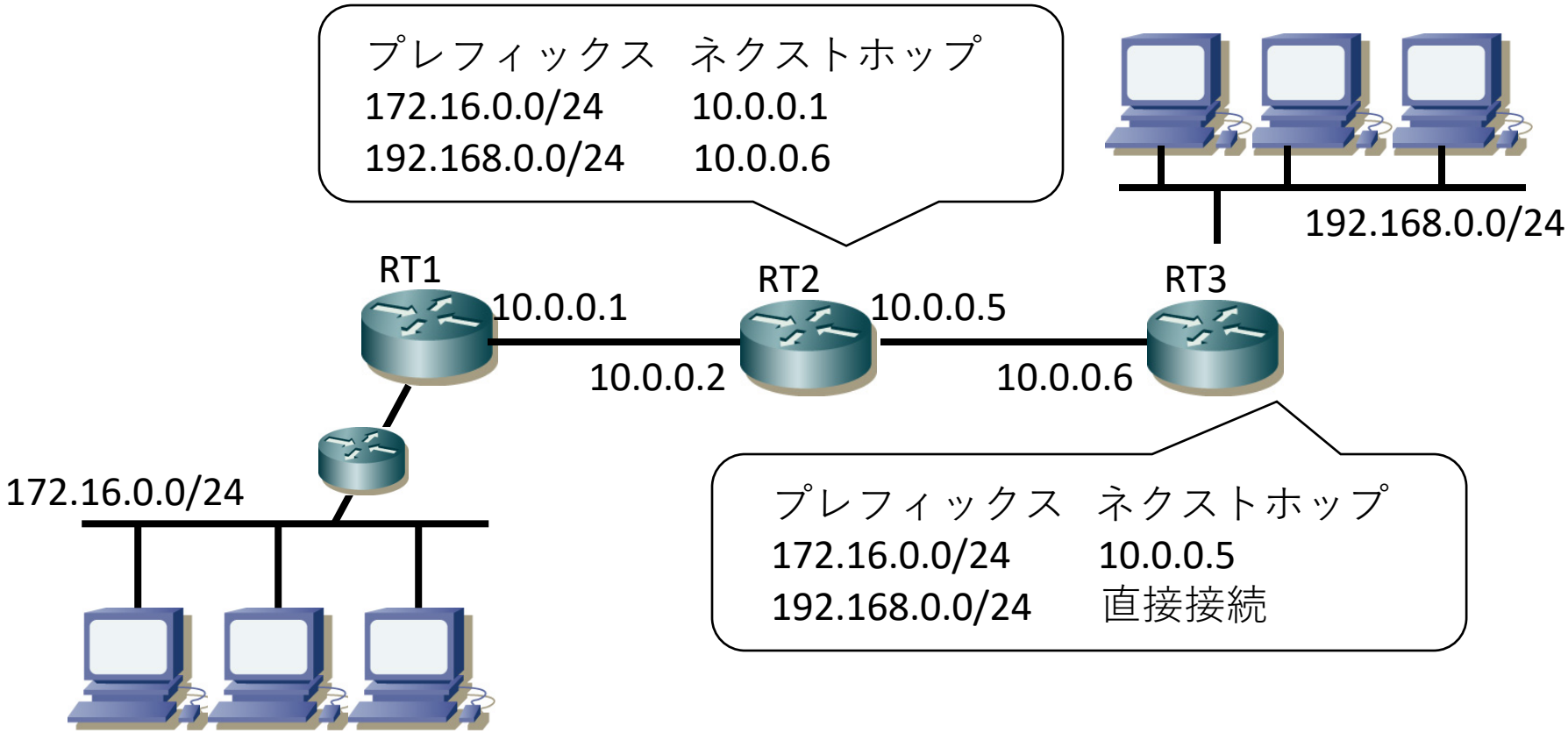
経路情報

- 宛先プレフィックス＋ネクストホップの集合



経路情報

- 宛先プレフィックス＋ネクストホップの集合



経路の優先順位

1. prefix長が長い(経路が細かい)ほど優先

   長い ⟵ prefix長 ⟶ 短い

   ホスト経路(/128)⟷default経路(::/0)

   ホスト経路(/32) ⟷default経路(0.0.0.0/0)

   優先 ⟵ 優先度 ⟶ 非優先

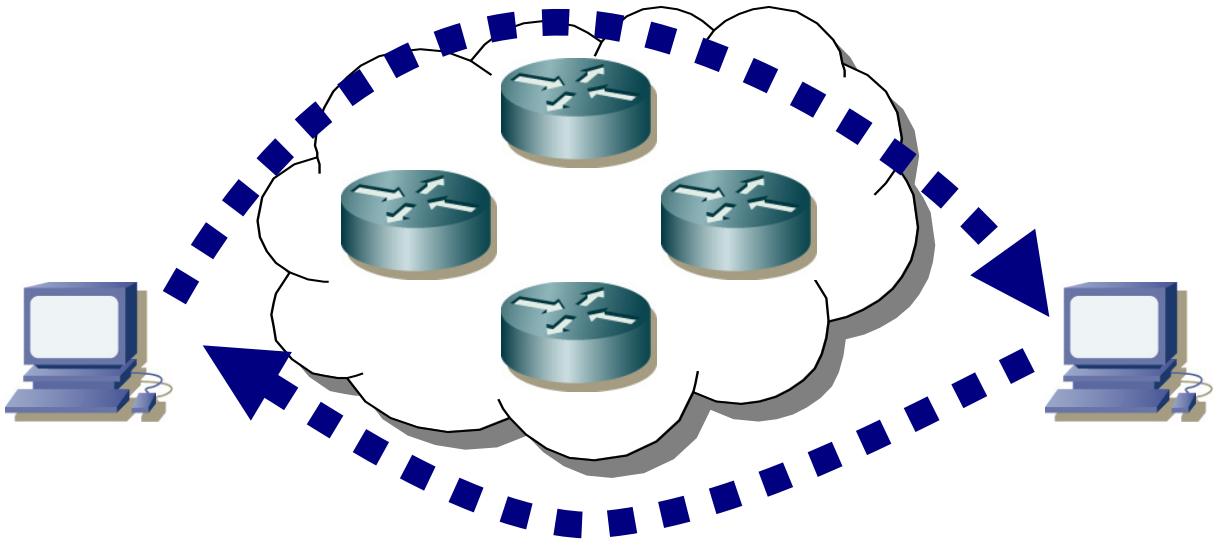
2. 経路種別で優先
   - ① connected経路
   - ② static経路
   - ③ 動的経路(ospf, bgp, etc...)
     - 内訳はベンダ依存

経路の種類

- 静的経路
  - connected経路
    - ルータが直接接続して知っている経路
  - static経路
    - ルータに静的に設定された経路
- 動的経路
  - ルーティングプロトコルで動的に学習した経路
    - OSPFやIS-IS、BGPなどで学習した経路
- これらを組み合わせて適切な経路制御を実現

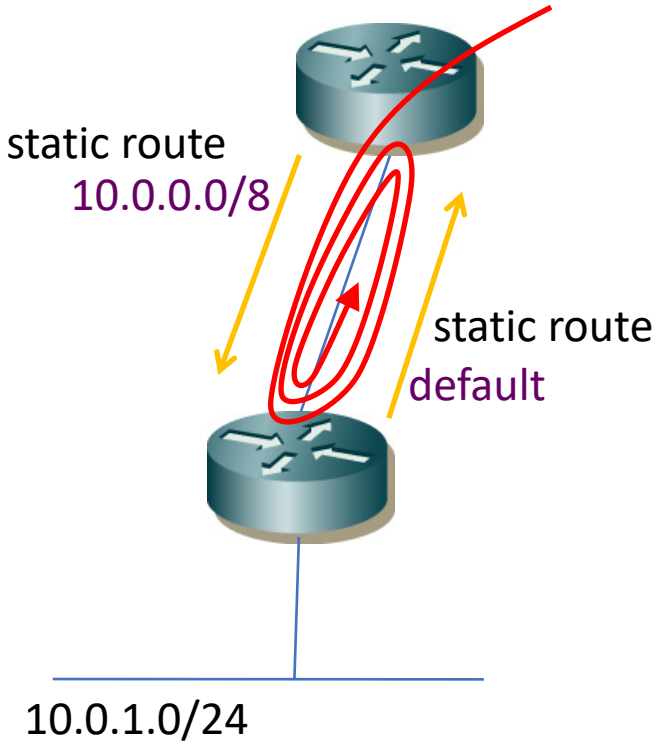
パケットと経路

- 送信元から宛先まで経路に矛盾が無ければ、宛先にパケットが届く
- 双方向で問題が無ければ、相互に通信できる
  - 行きと帰りの経路は違うかもしれない



経路ループ

- 起こしちゃダメ
  - 簡単に回線帯域が埋まる
- 大抵設定/設計ミス
  - 矛盾のあるstatic経路
  - 無茶な設定の動的経路制御

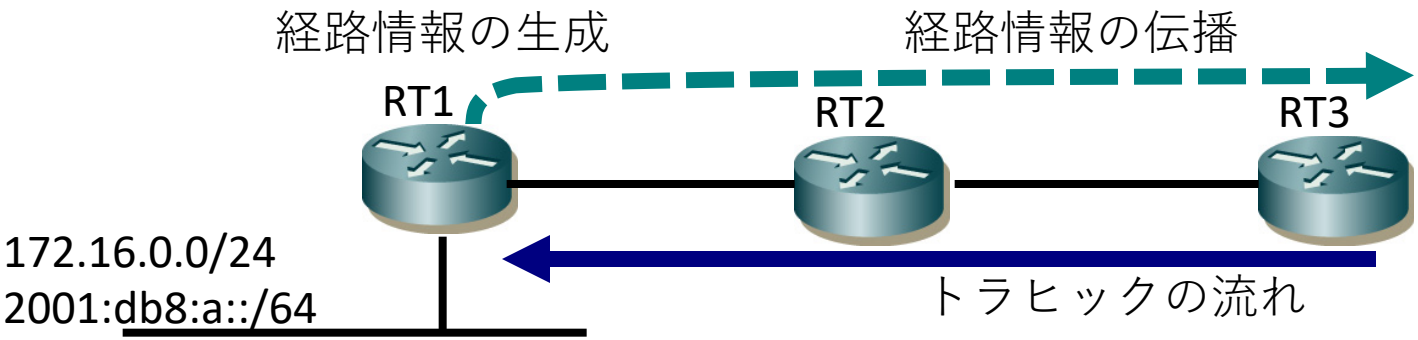


動的経路制御の必要性

- ネットワーク変化を経路情報に反映
  - 自動化 :)
  - ネットワークの拡張が容易
- ISPのバックボーン運用では必須
  - インターネットは変化し続けてる
  - うまく冗長設計すると障害時も綺麗に自動迂回
- 大事なこと
  - プロトコルごとの得手不得手を把握しておく
  - 何を設定しているのか理解しておく

動的経路制御の基本アイディア

- 検知 – ルータがネットワークの変化を検知
- 通知 – 情報を生成し他のルータに伝達
- 構成 – 最適経路で経路テーブルを構成



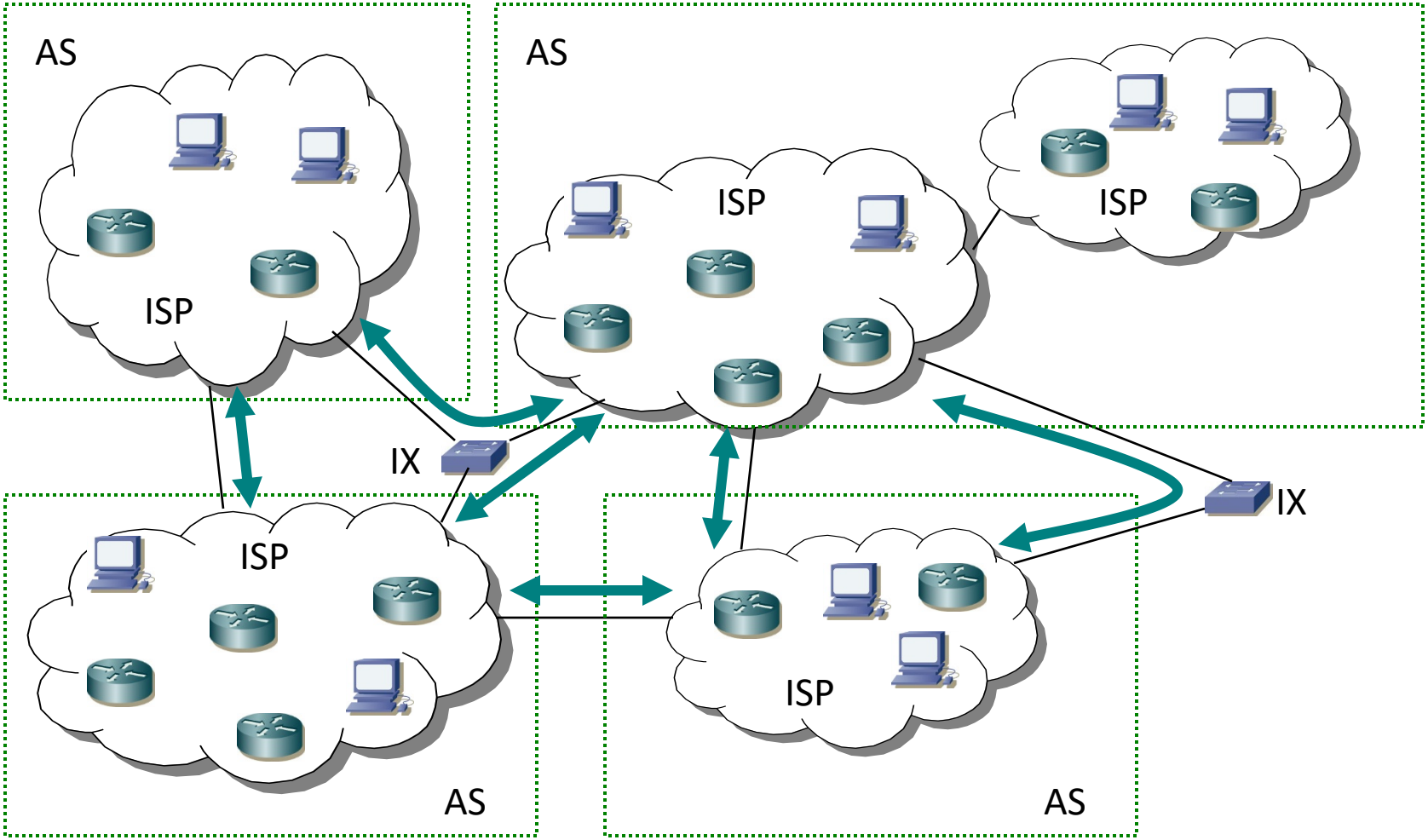
経路情報の伝搬の方向とトラヒックの流れは逆になる

動的経路制御の種類

- ディスタンスベクタ(distance vector)
  - RIPなど、距離と方向で運用するプロトコル
- リンクステート(link state)
  - OSPFやIS-ISなど、ルータに繋がっているリンク状態を収集して運用するプロトコル
- パスベクタ(path vector)
  - BGPなど、パス属性と方向で運用するプロトコル

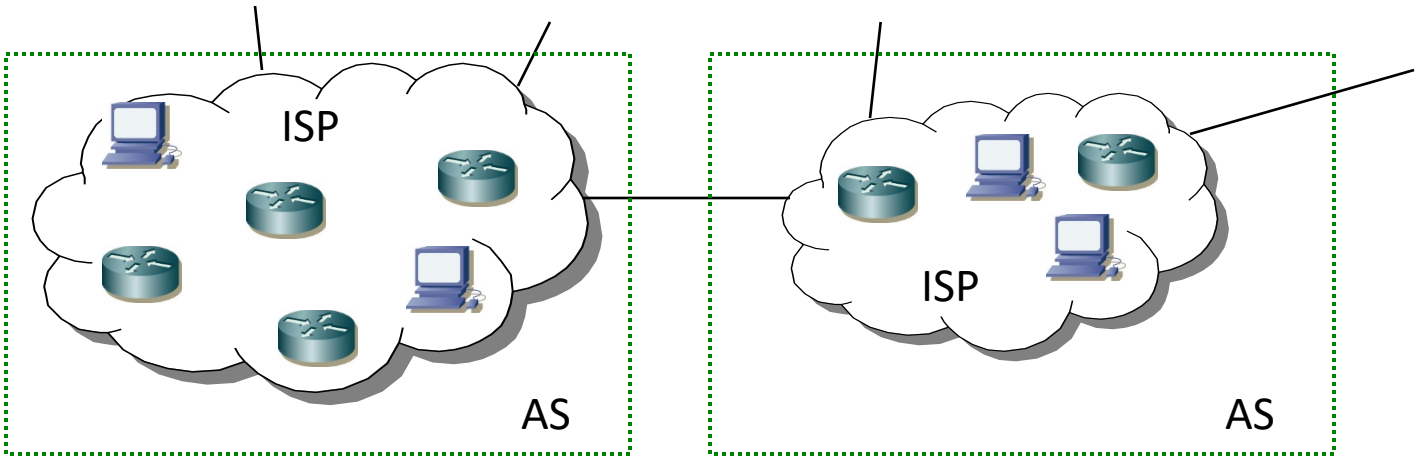
BGP

インターネットの構成



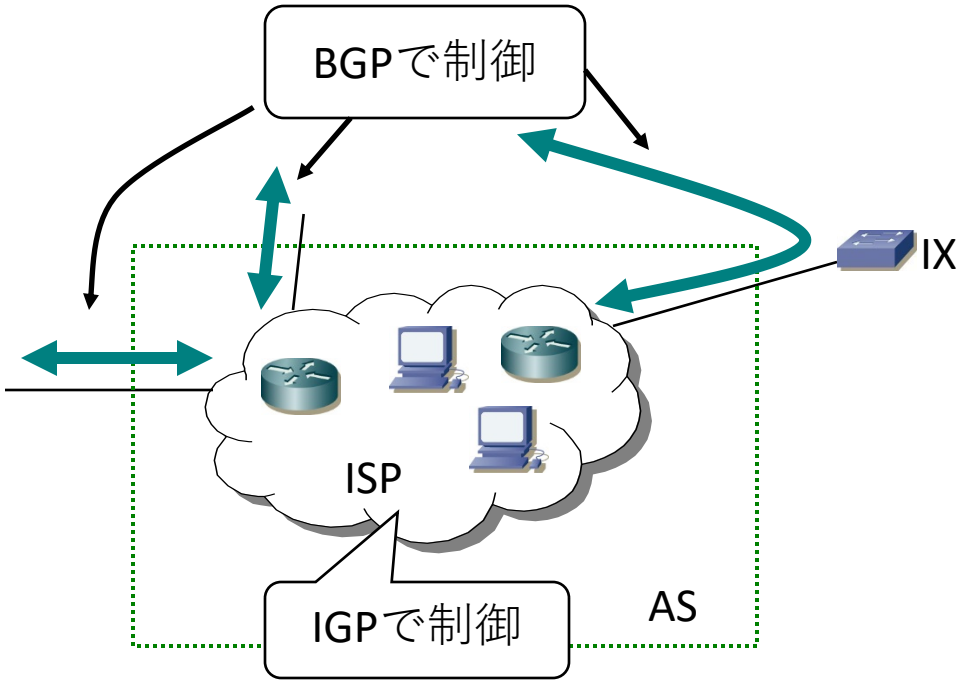
AS

- Autonomous System
- 統一のルーティングポリシのもとで運用されているIPプレフィックスの集まり
- ASの識別子として、インターネットではIRから一意に割り当てられたAS番号を利用する
  - IR: JPNICとかAPNICとかのインターネットレジストリ



IGPとEGP

- IGP
  - OSPF、IS-IS、BGP等
  - AS内
- EGP
  - 事実上BGPのみ
  - AS間



- 最近は網内でトポロジ情報の交換に使うプロトコルをIGPとして認識している場合も多い

BGP概要

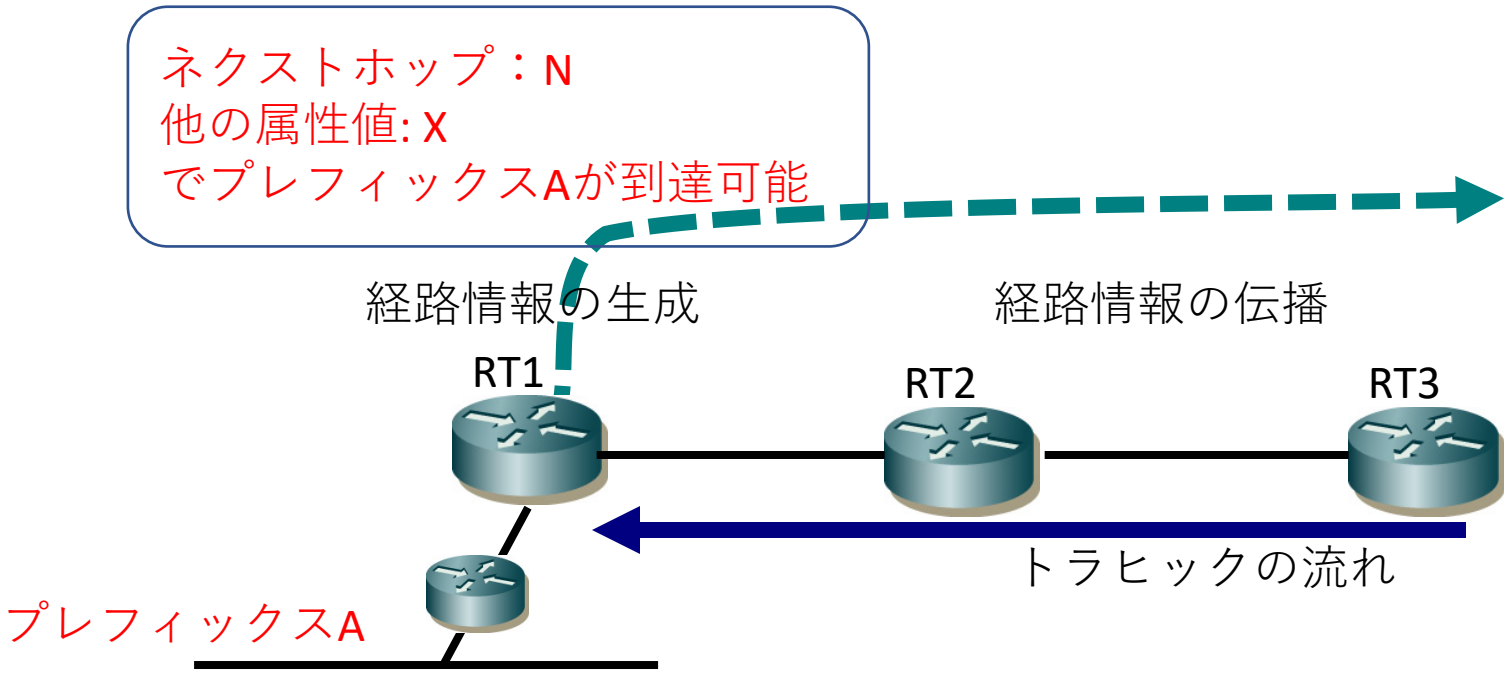
- パスベクタ型プロトコル
  - プレフィックスに付加されたパス属性で経路制御
- AS番号によって組織間、組織内を認識する
- 経路交換にTCPを利用
  - データの到達や再転送はTCP任せ
- 変更があった場合にのみ通知
  - ベスト経路のみを通知する
- 現在のバージョンは４ (BGP4)

BGPの基本アイディア

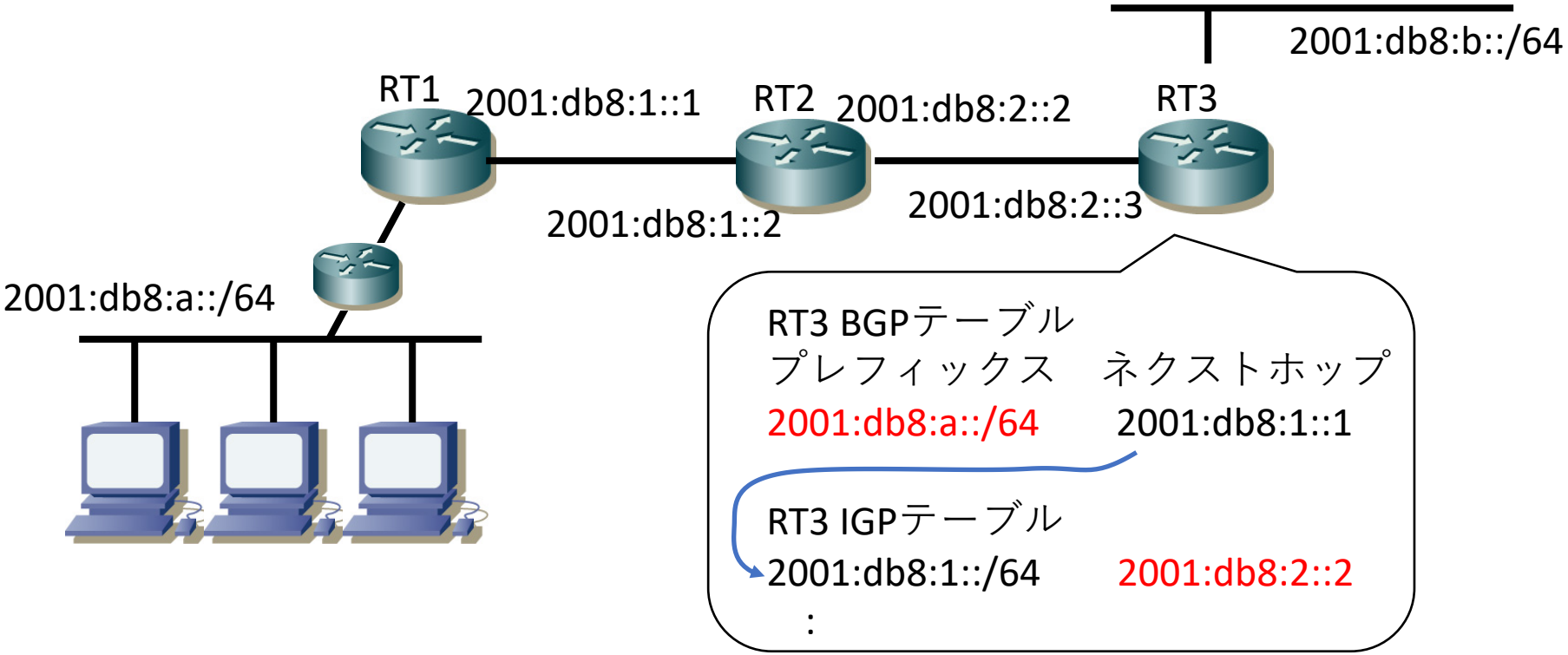
- 準備
  - 経路交換したいBGPルータとTCPでネイバを構築
  - (ネイバ|ピア|BGPセッション)を張るとも言う
- 通知
  - 最適経路に変更があればUPDATEとしてネイバに広報
  - 受信した経路は幾つかの条件を経て、他のネイバに広報
- 構成
  - 各ルータが受信経路にポリシを適用し、パス情報を元に最適経路を計算して経路情報を構築
  - 経路情報に従ってパケットを転送

BGPの経路広告ざっくり

- ネクストホップ + 他の属性値+プレフィックス

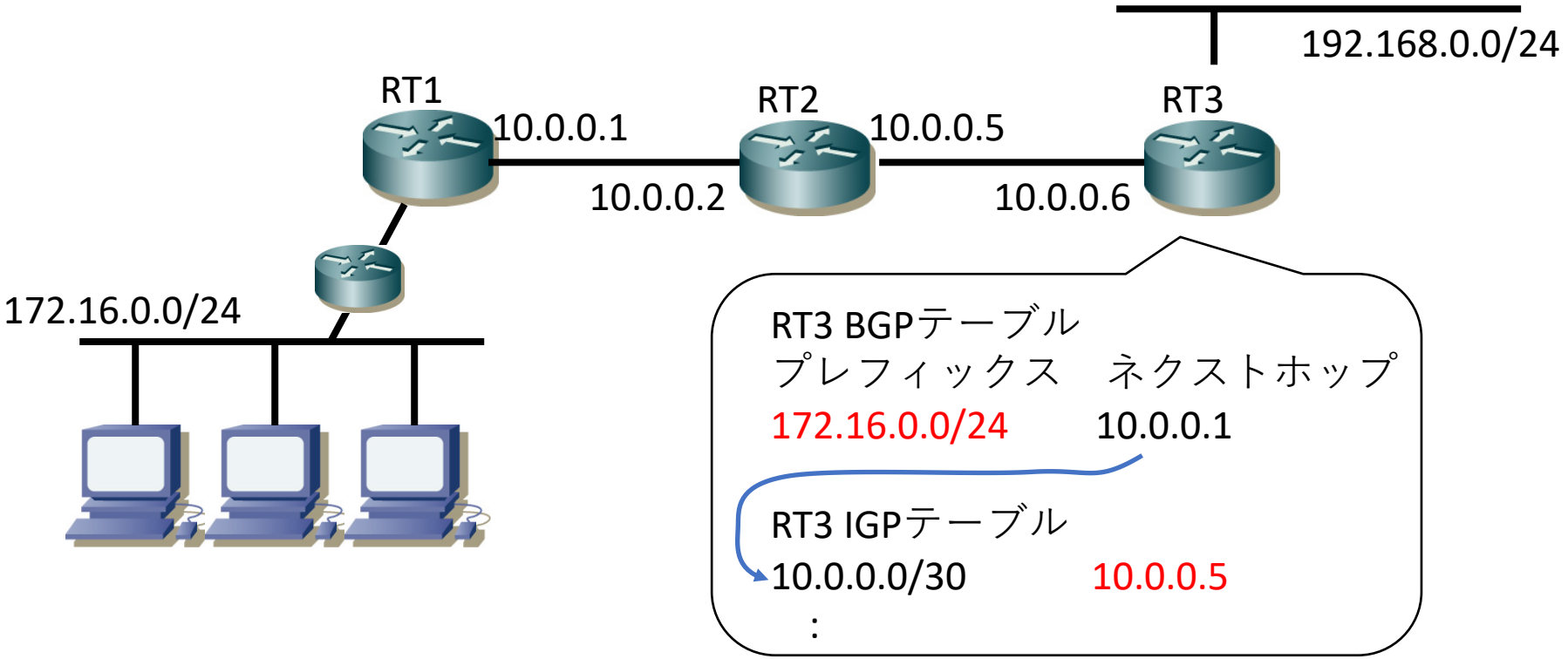


BGPと再帰経路



BGPで学習したネクストホップアドレスをさらに
経路情報で再帰的に探して、ルータが実際に
パケットを送出する隣接ノードを見つけ出す
「2001:db8:a::/64宛は2001:db8:2::2(RT2)にフォワード」

BGPと再帰経路



BGPで学習したネクストホップアドレスをさらに経路情報で再帰的に探して、ルータが実際にパケットを送出する隣接ノードを見つけ出す
「172.16.0.0/24宛は10.0.0.5(RT2)にフォワード」

経路優先度

1	NEXT_HOP	NEXT_HOP属性のIPアドレスが到達不可能な経路は無効
2	AS loop	AS Path属性に自身のAS番号が含まれている経路は無効
3	LOCAL_PREF	LOCAL_PREF属性値が大きい経路を優先 (LOCAL_PREF属性が付加されていない場合は、ポリシに依存)
4	AS_PATH	AS_PATH属性に含まれるAS数が少ない経路を優先 (AS_SETタイプは幾つASを含んでも1として数える)
5	ORIGIN	ORIGIN属性の小さい経路を優先 (IGP < EGP < INCOMPLETE)
6	MULTI_EXIT_DISC	同じASからの経路はMED属性値が小さな経路を優先 (MED属性が付加されていない場合は、最小(=0)として扱う)
7	PEER_TYPE	IBGPよりもEBGPで受信した経路が優先
8	NEXT_HOP METRIC	NEXT_HOPへの内部経路コストが小さい経路が優先 (コストが算出できない経路がある場合は、この項目をスキップ)
9	BGP_ID	BGP IDの小さなBGPルータからの経路が優先 (ORIGINATOR_IDがある場合は、これをBGP IDとして扱う)
10	CLUSTER_LIST	CLUSTER_LISTの短い経路が優先
11	PEER_ADDRESS	ピアアドレスの小さなBGPルータからの経路を優先

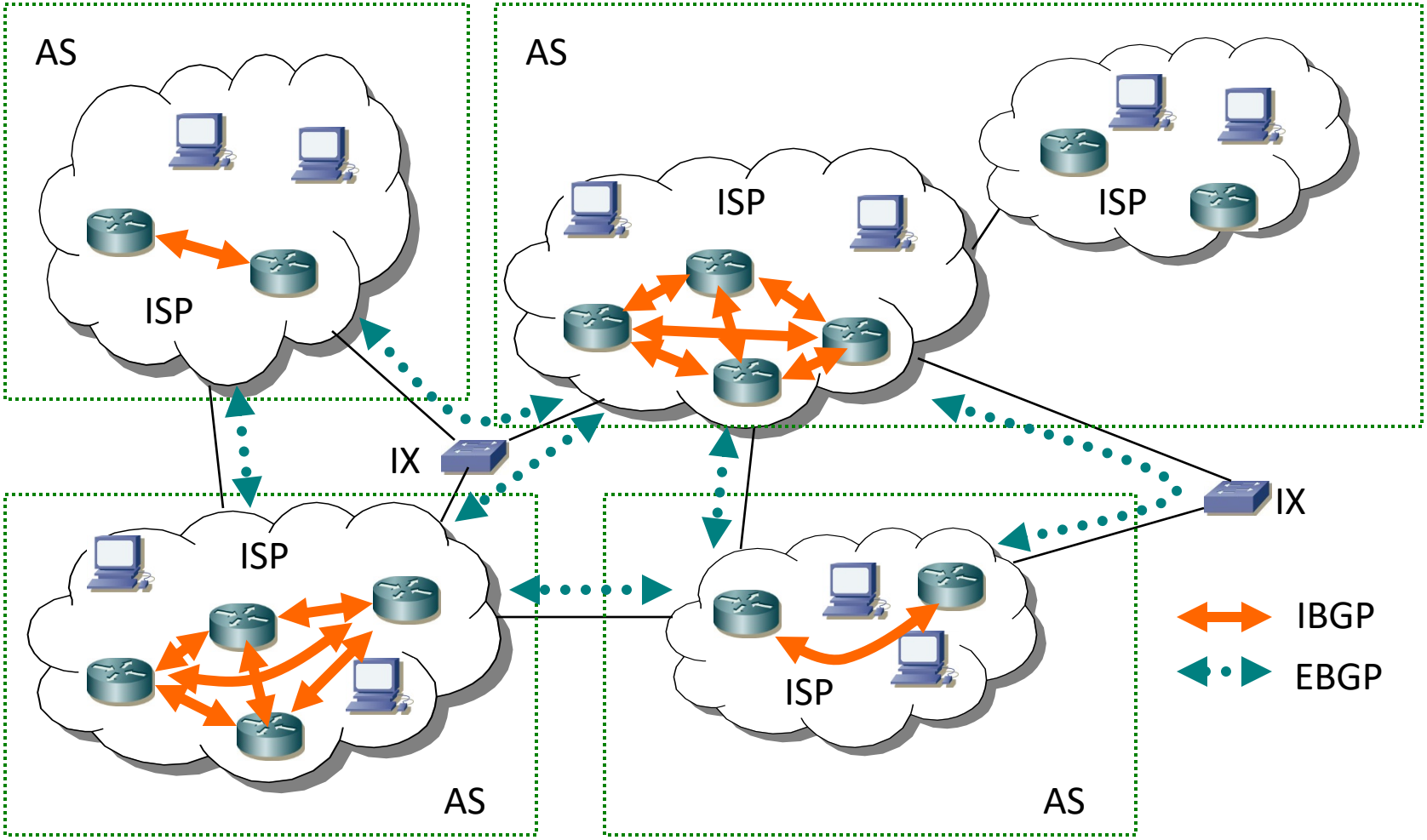
BGP RFCs

- 基本
  - [RFC4271] A Border Gateway Protocol 4 (BGP-4)
- この他にもいっぱい
  - [RFC1997] BGP Communities Attribute
  - [RFC3065] AS Confederations for BGP
  - [RFC4451] BGP MED Considerations
  - [RFC4456] BGP Route Reflection
  - [RFC6286] AS-Wide Unique BGP Identifier for BGP-4
  - [RFC6793] BGP Support for Four-Octet AS Number Space
  - [RFC7606] Codification of AS 0 Processing
  - [RFC8092] BGP Large Communities Attribute
  - [RFC8212] Default EBGP Route Propagation Behavior without Policies

BGP用語

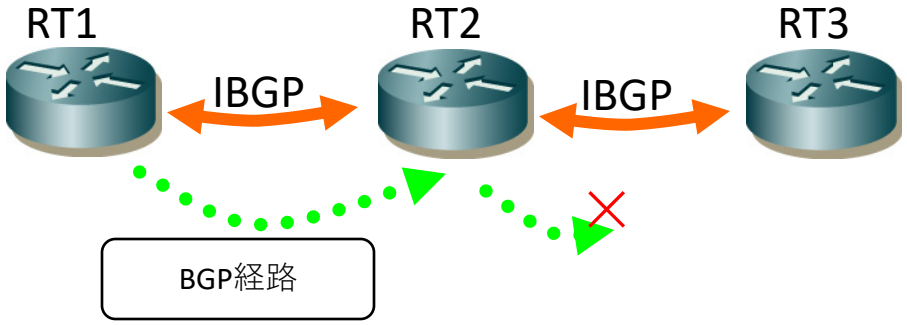
- BGP ID
  - ルータを識別する32bitの数値
    - AS内で一意である必要がある [RFC6286]
  - インタフェースの何れかのIPアドレスから選ばれる
  - 変更が発生しないように明示的に指定するか、loopbackインタフェースに付与したIPv4アドレスを利用する場合が多い
- NLRI
  - Network Layer Reachability Information
  - ネットワーク層到達可能性情報
  - prefixで示される宛先のこと

BGPの世界



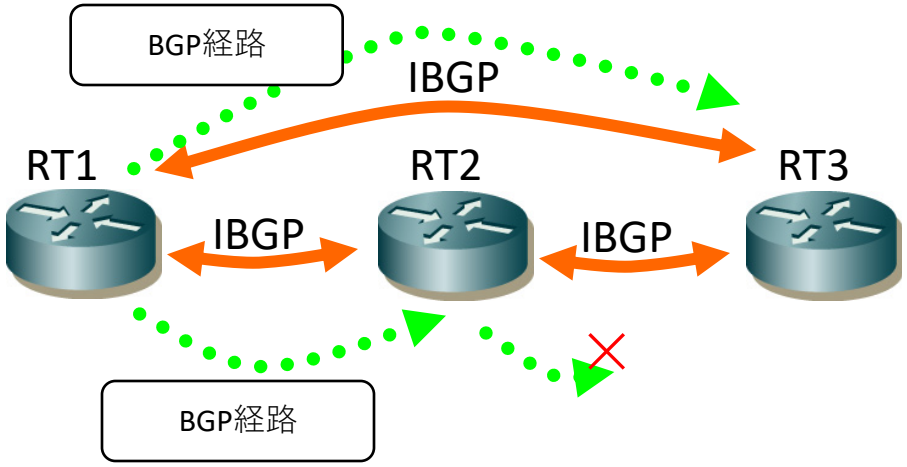
IBGP(Internal BGP)

- 同じAS内でのBGP接続
- IBGPで受信した経路は他のIBGPルータに広報されない
  - 全ての経路を伝えるには、AS内の全BGPルータがfull-meshでIBGPを張る必要がある



IBGP full-mesh

- AS内の全BGPルータが全ての経路を交換できるようにするためには、AS内の全BGPルータがfull-meshでIBGPを張る必要がある

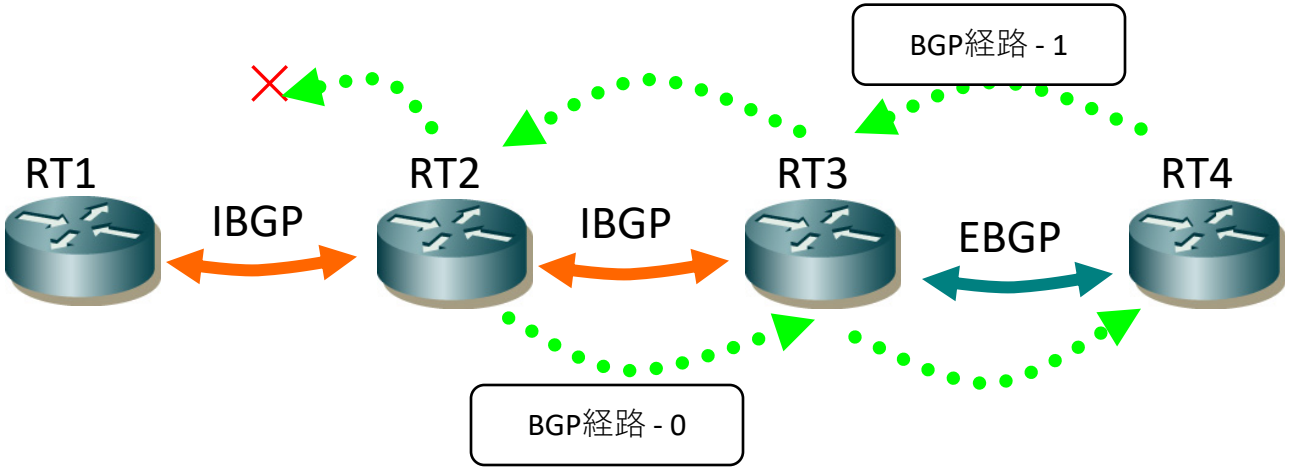


IBGPの基本

- 通常、Loopbackインタフェースを利用
  - どれか物理インタフェースが生きてたら到達可能
  - IGPでLoopback間の到達性を確保
- 経路情報をそのまま伝える
  - 基本的にパス属性を操作しない
    - MEDやLocal Preference等の優先度、ネクストホップ
  - 下手にいじると経路ループする
- 基本的に全てを広報し、全てを受け取る
  - 特段の理由が無ければ経路フィルタしない

EBGP(External BGP)

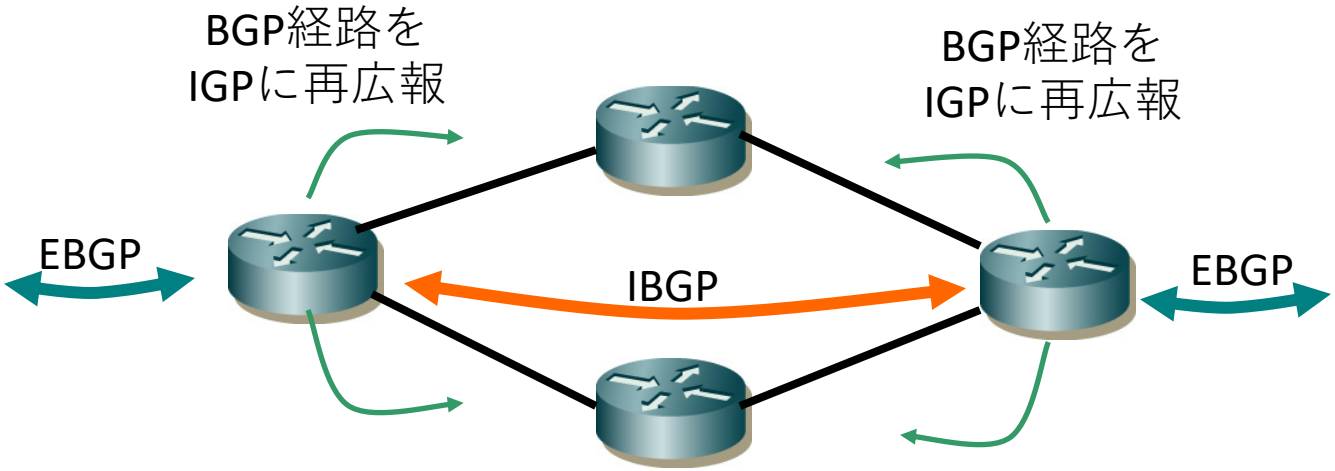
- 異なるASとのBGP接続
- EBGPから受信した経路は、他のBGPルータに広報する
  - IBGPから受信した経路もEBGPには広報する



EBGPの基本

- 通常、物理接続してるインターフェースで張る
- ポリシの実装をするならここ
  - 受信のポリシ
    - 不要な経路のフィルタやタグ付け
    - MEDやlocal preferenceによる優先制御
  - 広報のポリシ
    - 不要な経路のフィルタと必要な経路の広報
    - MEDやprependによる優先制御
- ポリシが違うところは網内でもEBGPが便利
  - Private AS番号の利用など
  - 64512-65534, 4200000000-429496729

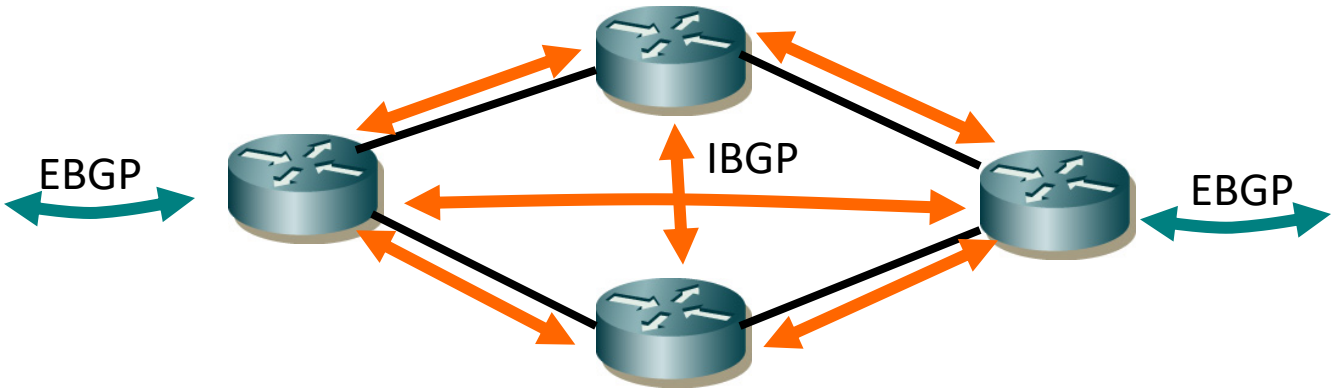
BGPのいにしえのモデル



- EBGPを張るルータのみがBGPルータとなる
- BGP経路をIGP(OSPFやIS-IS)に再広報してAS内部はIGPで経路制御
  - 内部にIGPのみのトラヒック中継ルータが居るため、bgp synchronizationが必要だった

・・・経路数が増大すると破綻

今時のBGPモデル



- 主要なルータは全てBGPルータ
- IGPはトポロジと最低限の経路を運び、BGPでその他の全ての経路を運ぶ

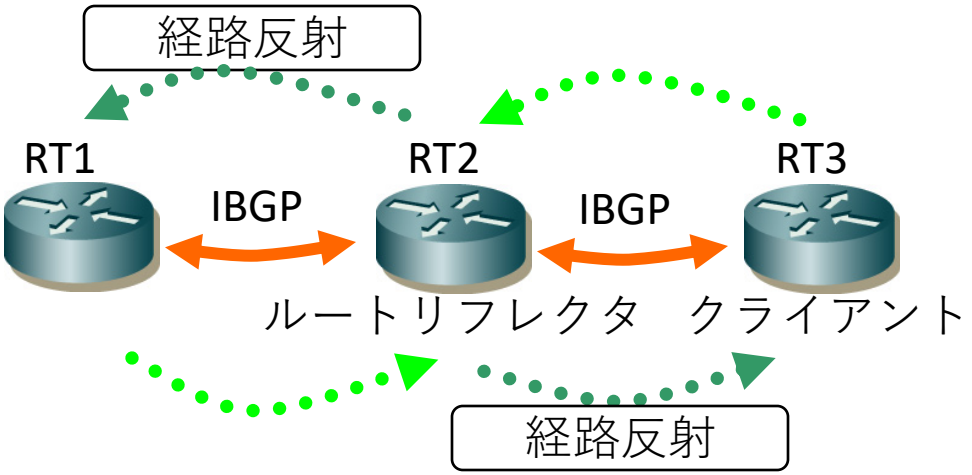
・・・IBGP接続の増大

IBGP full-mesh n*(n-1)/2

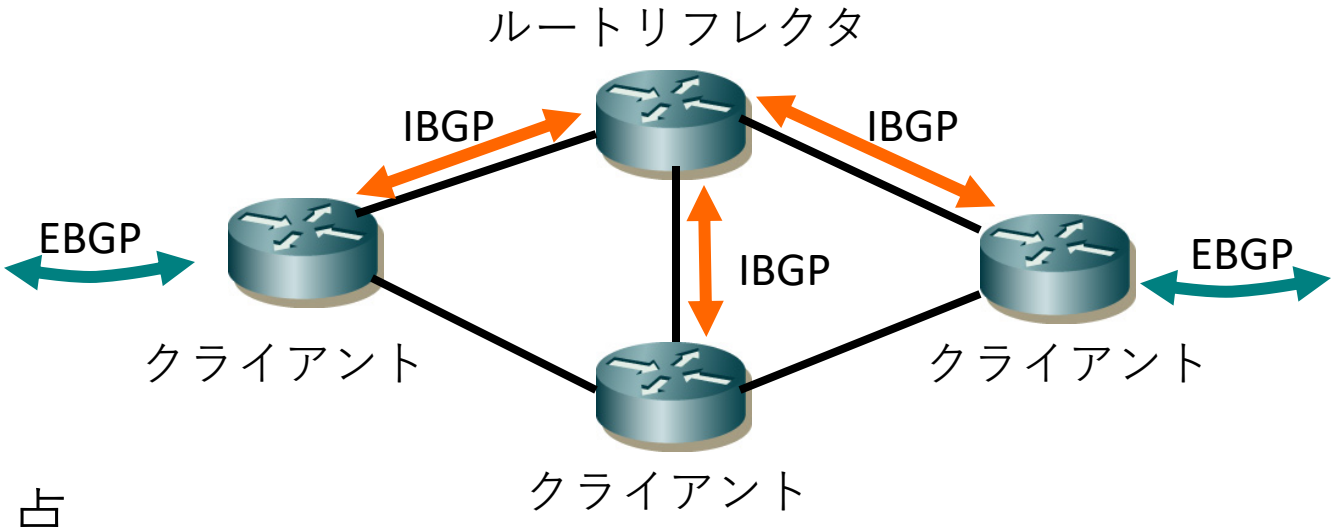
- AS内にBGPルータが増える毎にIBGP接続が増大していく
  - 20台目のBGPルータが接続すると19接続追加
  - ルータリソースの問題、設定負荷の問題
- 解決策の模索
  - [RFC4456] ルートリフレクタ
  - [RFC3065] コンフェデレーション
  - 気にせずリソースを強大にする
  - ルータを減らす

ルートリフレクタ

- IBGPで受信した経路の転送ルールを変更
- ルートリフレクタの機能
  - BGP接続ごとに設定される
  - クライアント以外のIBGPで受信した経路をクライアントに送信
  - クライアントから受信した経路を他のIBGPルータに送信
- ベスト経路のみを広報するルールは変わらない



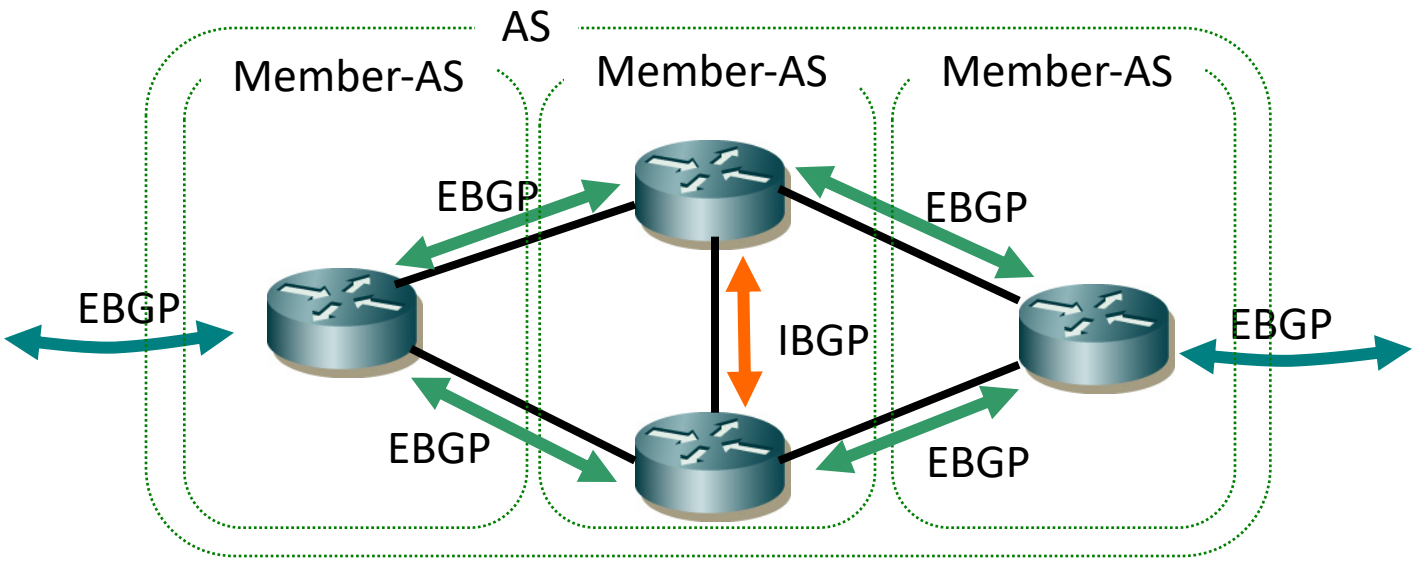
ルートリフレクタの利点と欠点



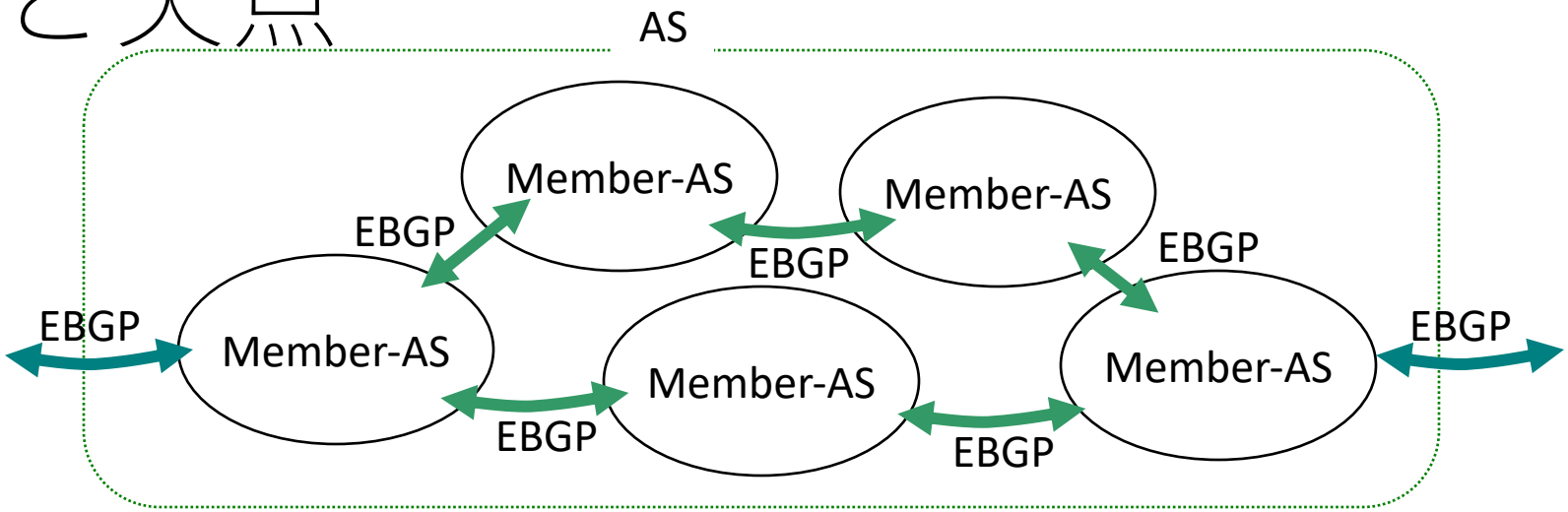
- 利点
  - IBGP接続数が削減できる
  - 比較的容易に導入できる
- 欠点
  - 経路削除時に、UPDATEが増える可能性がある
  - 経路情報が隠蔽されるため最適ではない経路を選ぶ可能性がある
    - リフレクタの階層はできるだけ物理トポロジに合わせるべし！

コンフェデレーション

- 外部からは一つのASのままだが、内部を複数のメンバASで構成する
- メンバAS間のBGP接続はEBGPに似た挙動をする
- メンバASにはプライベートASを使うのが一般的



コンフェデレーションの利点と欠点



- 利点
  - IBGP接続数が削減できる
  - 管理区分を分けられる
- 欠点
  - 経路削除時にUPDATEが増える可能性がある
  - 経路情報が隠蔽されるため最適ではない経路を選ぶかもしれない

IBGP - 網内構成

ISPでのプロトコルの利用法

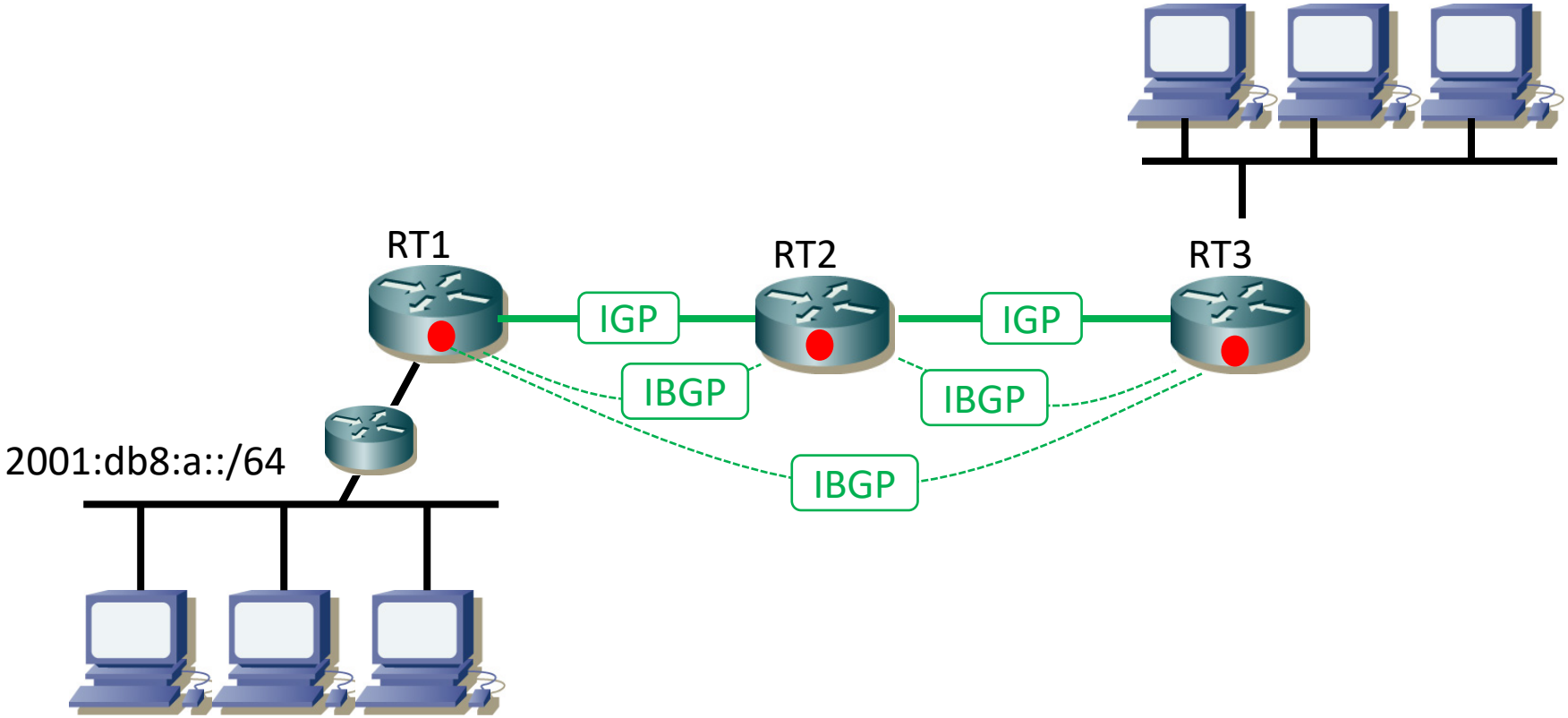
- OSPF or IS-IS
  - ネットワークのトポロジ情報
  - 必要最小限の経路で動かす
  - 切断などの障害をいち早く通知、迂回
- BGP
  - その他全ての経路
    - 顧客の経路や他ASからの経路
  - 大規模になっても安心
  - ポリシに基づいて組織間の経路制御が可能

IBGP構成の考慮点

- IBGPを張るインタフェース(IPアドレス)
  - ふつーはLoopback
- 経路を網内に流すときのネクストホップ
  - 外側のアドレスか内側のアドレスか
- IBGP full-meshをどうするか
  - 気にしない
  - ルートリフレクタ構成
  - プライベートASで分離

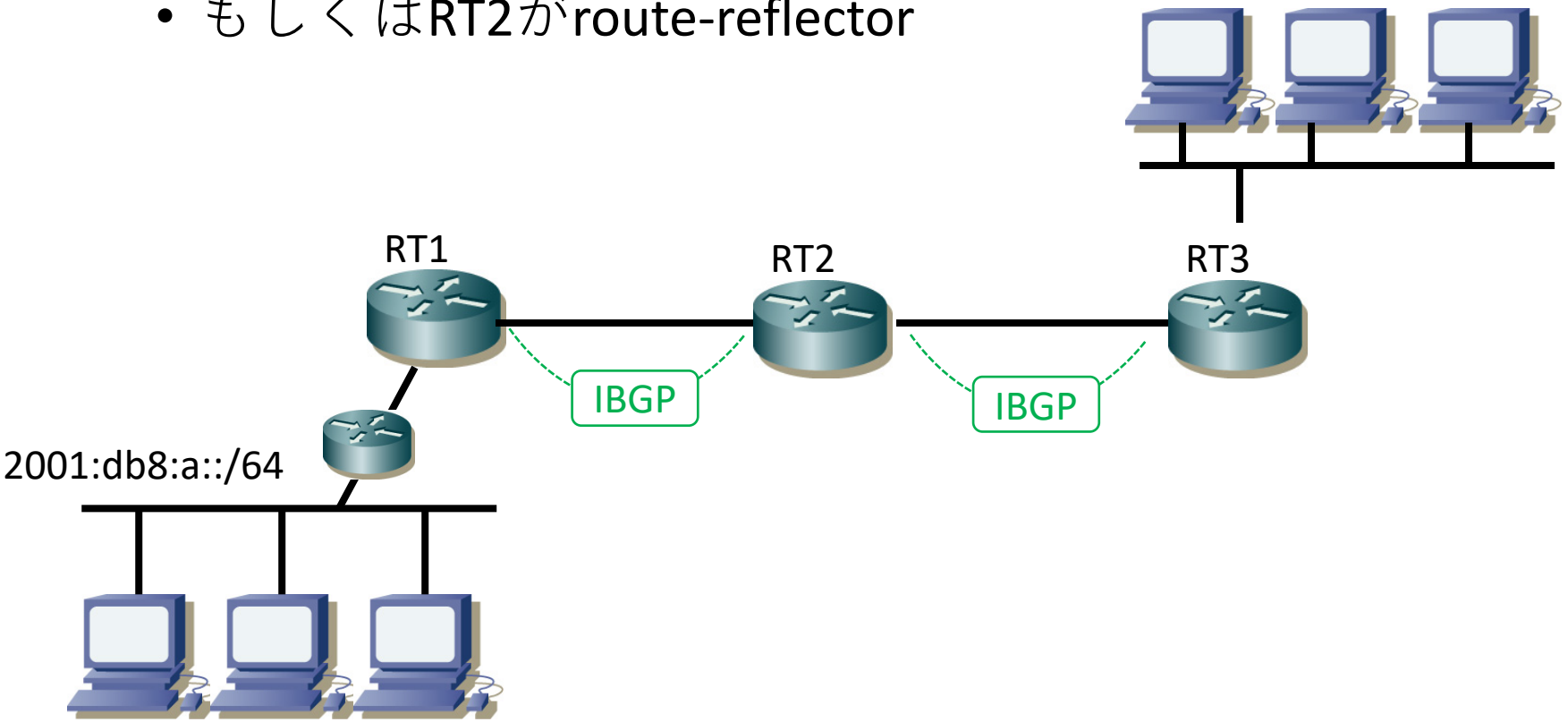
ふつーの網内構成

- 最低限Loopbackの経路情報をOSPF/IS-ISで交換
- ルータのLookback間でIBGPを構成

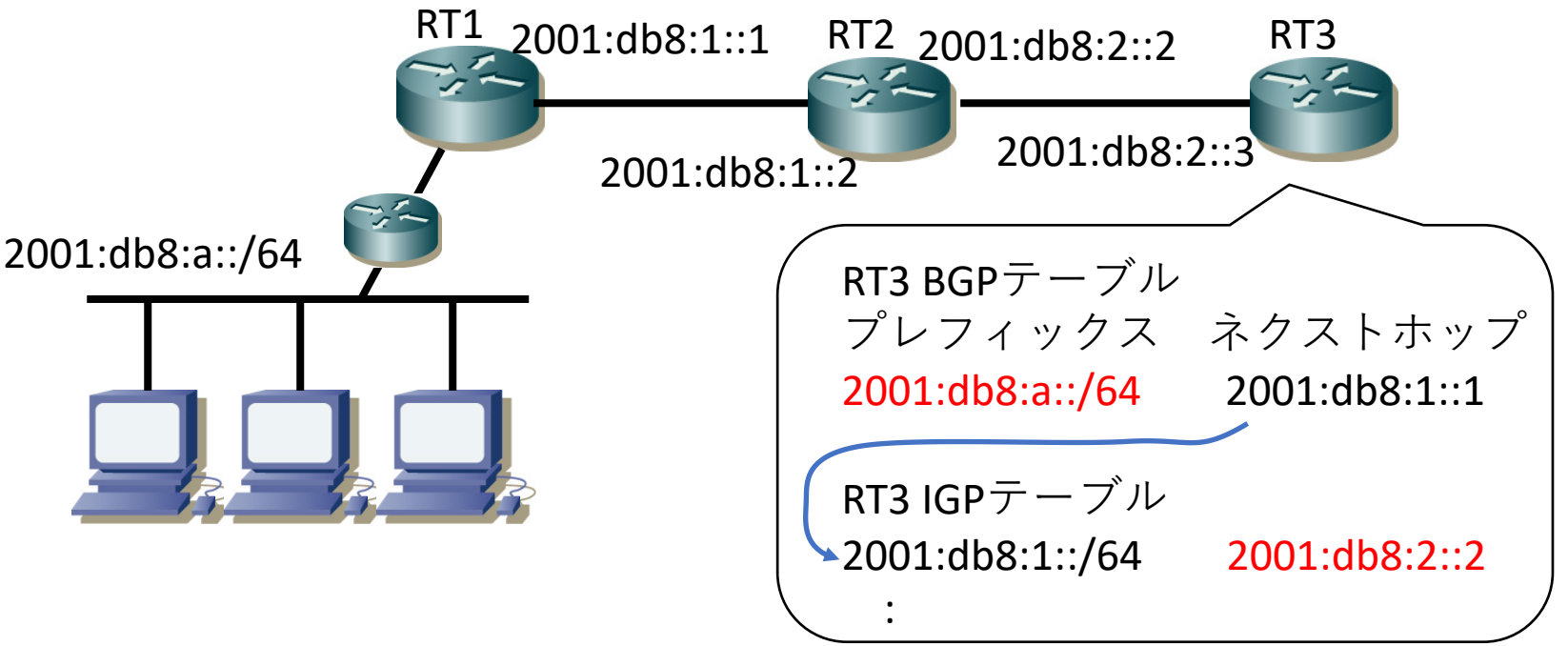


IBGPの構成の自由さ(用途次第)

- ルータの実インタフェース間でIBGPを張る
  - 全ルータが異なるプライベートAS
  - もしくはRT2がroute-reflector

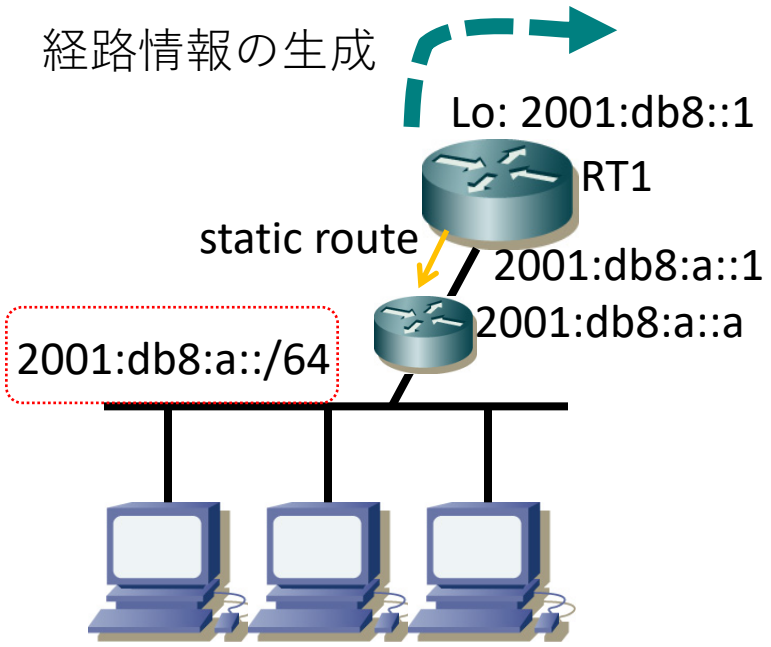


BGPネクストホップと再帰検索



BGPで学習したネクストホップアドレスをさらに
経路情報で再帰的に探して、ルータが実際に
パケットを送出する隣接ノードを見つけ出す
「2001:db8:a::/64宛は2001:db8:2::2(RT2)にフォワード」

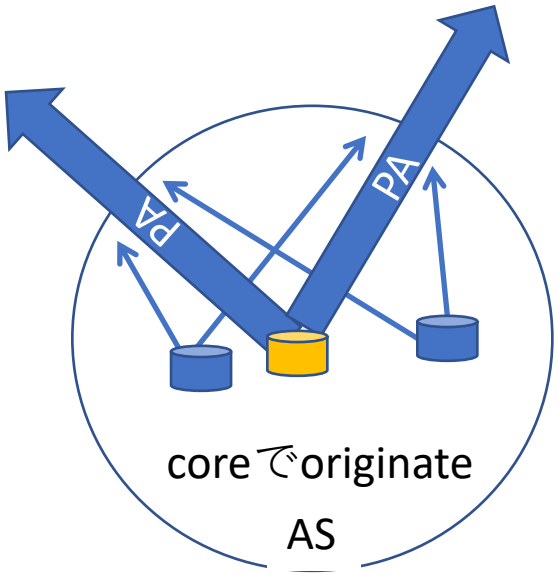
経路生成時のネクストホップ



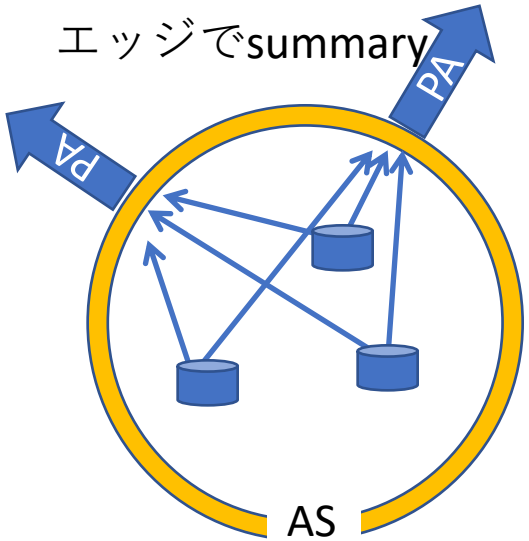
- RT1のLoopback
  - 2001:db8::1
  - next-hop-selfなどのBGP設定
  - LoopbackをIGPで網内に広報しておく
- staticの向き先
  - 2001:db8:a::a
  - 2001:db8:a::/64を何らか網内に広報しておく
- RT1の網内実インタフェース
  - 2001:db8:1::1
  - ふつーあまりやらない。実インターフェースでIBGPを張っているなど特殊な構成時のみ

PA経路の生成

1. 内部のルータでnull向けstatic経路から生成
2. EBGPルータでsummary経路として生成



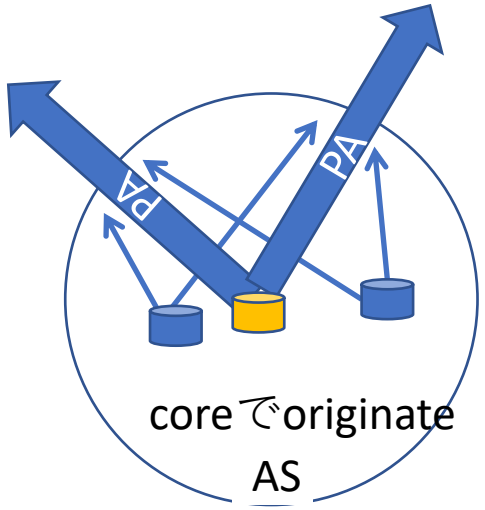
内部で生成方式



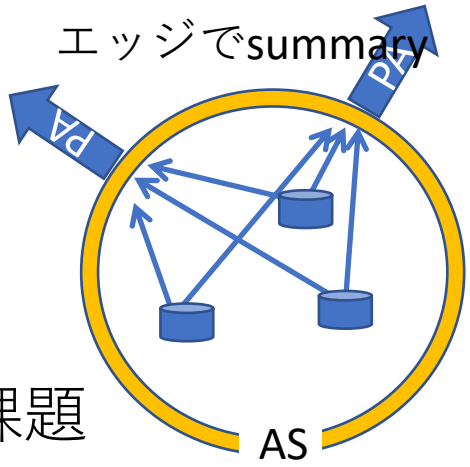
エッジで生成方式

経路の生成：内部で生成方式

- 想定障害
  - 経路生成ルータでの障害や到達性障害
  - ネットワーク分断
  - -> 経路をいかに広報し続けるかが課題
- 対策案
  - 複数台での経路生成
  - 内部ネットワークで頑強な接続性を保持しているルータで経路生成
  - なるべく実ネットワーク利用の近傍で経路生成



経路の生成：エッジで生成方式



- 想定障害
  - 経路生成ルータの孤立
  - ネットワーク分断
  - -> 障害時にうまく広報を止めることが課題
- 対策案
  - 障害時の影響を小さくするため、地域ごとに利用するprefixを分ける
  - 他のIGPとBGPで運ぶ経路情報を棲み分ける
- 他ASと隣接する全EBGPルータで設定が必要
  - 顧客向けやピア収容ルータで忘れないように

customer持ち込みのPI経路生成

- 回線向けのstatic経路から生成
  - 回線が落ちると経路が消える
    - multiple origin ASしている場合には必須機能
  - 回線がflapするとdampeningペナルティがあるかも
- null向けstatic経路から生成
  - customerとの回線が落ちても経路は消えない
  - BGP的には安定

トラヒック増加対応

- １インタフェースの上限速度がある
  - 今のところ、10GEが標準的
  - 100GEがようやく使われ始めたけどまだ高い
- ISP間、ルータ間は10G以上のトラヒック
  - 実効帯域を何とかして増やしたい
  - しかも、冗長構成は必須
- 次のインタフェースがあんまりない
  - 400Gbps?
  - 中途半端なので、多分100Gを束ねて使う方が楽

link aggregation

- 回線を束ねて、論理的に一つの回線に見せる
  - 複数の回線を束ねられる
  - 束ねられる回線数には実装により、上限あり
- 回線が切れると迂回路に回る
  - 用意した帯域の半分程度しか利用できない
  - 実トラヒック量が許すなら、構成回線が全て切れるまで断と見なさない運用も可能

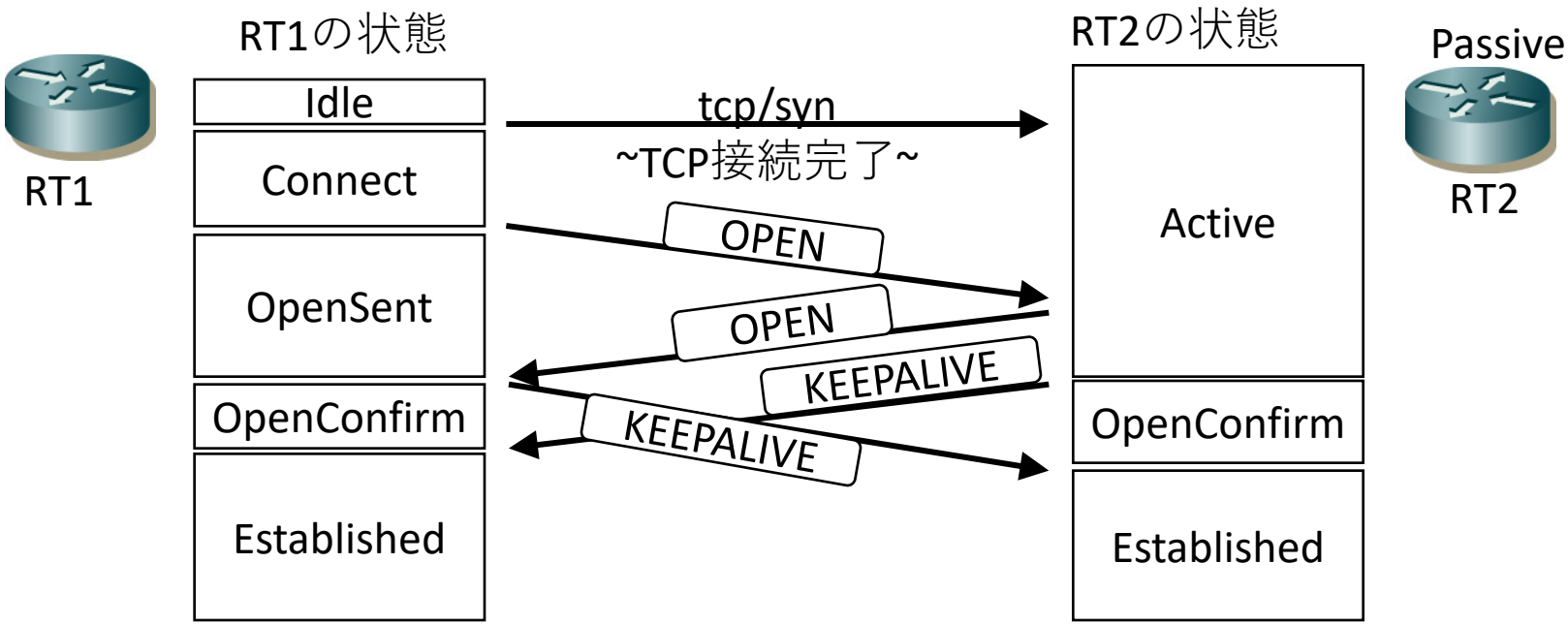
multipath

- OSPF Multipath
  - ISP(AS)内での分散に利用可能
  - 標準技術
- BGP Multipath
  - 非標準技術だが、多くのベンダが採用
  - 構成をきちんと組めば、ISP(AS)間にも有効
- 帯域の利用効率が良い
  - IPアドレスやポート番号にルータ毎のsaltを加えたhashでフォワードする回線を選ぶ場合が多い
    - flowベースで同じ回線を通る
    - 多段にmultipathしてもそこそこ分散するように

BGPパケット

BGPのプロトコルパケットの
フォーマットを解説する

BGP接続の確立



Idel – 初期状態

Connect – TCPの接続完了待ち

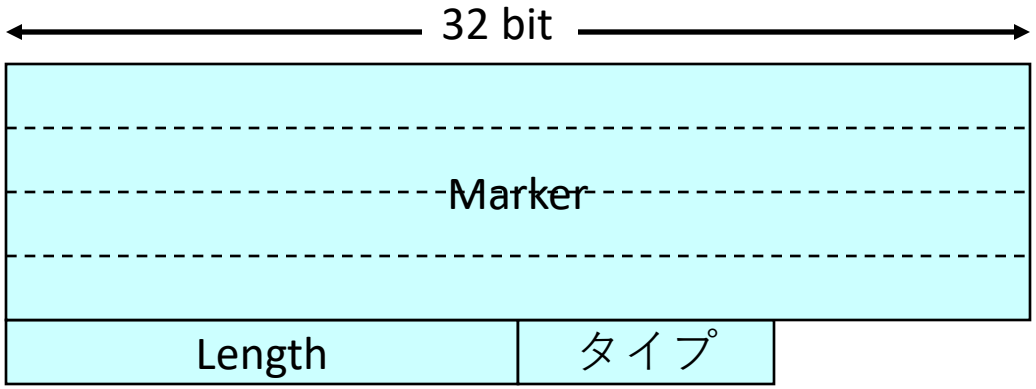
Active – 隣接からのTCP接続を待つ

OpenSent – OPEN送信後、隣接からのOPENを待つ

OpenConfirm – OPEN受信後、隣接からのKEEPALIVEを待つ

Established – BGP接続完了、経路交換の開始

BGP Message header



- Marker(マーカ)
  - 16-octetの全bitが１
  - 過去との互換性のため
- Length
  - 2-octetのメッセージ長
  - 19～4096

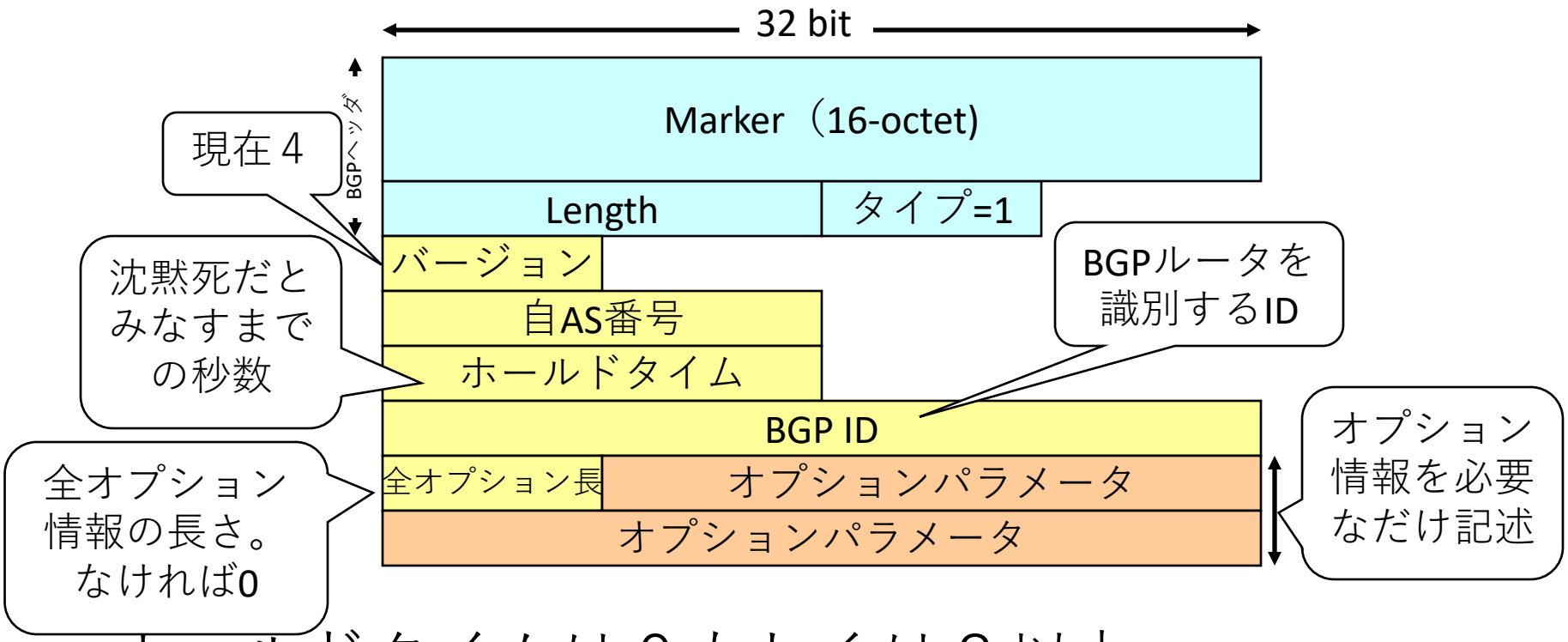
- タイプ（1-octet)
  1. OPEN
  2. UPDATE
  3. NOTIFICATION
  4. KEEPALIVE
  5. ROUTE_REFRESH

タイプ１ OPENメッセージ

- TCP接続が確立後、最初にやりとりされる
- パラメタの交換
  - バージョン、AS番号やBGP ID、ホールドタイム
  - オプションパラメータで各種機能を通知しあう

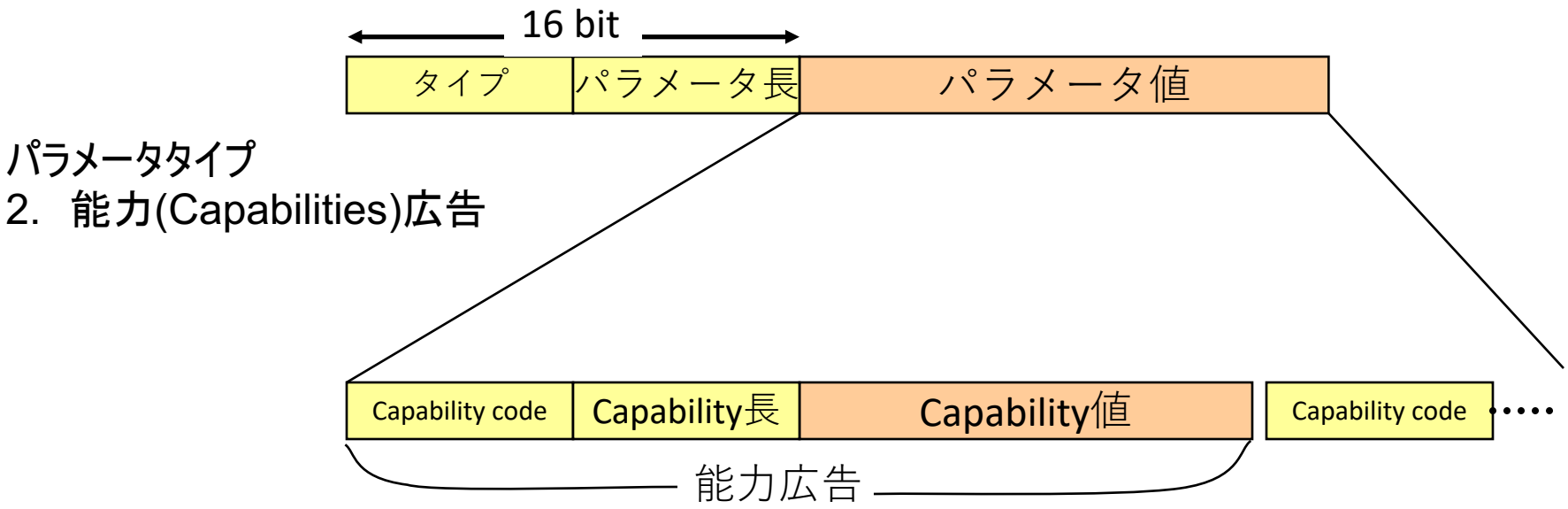
- タイプ4 KEEPALIVEで接続確立

タイプ１ OPENメッセージ



- ホールドタイムは０もしくは３以上
  - 小さな値が採用される
  - ０の場合、セッション維持にKEEPALIVEを利用しない

オプションパラメータフォーマット



- 今のところ能力広告のためだけに利用
  - 利用可能な機能をピア先へ通知する

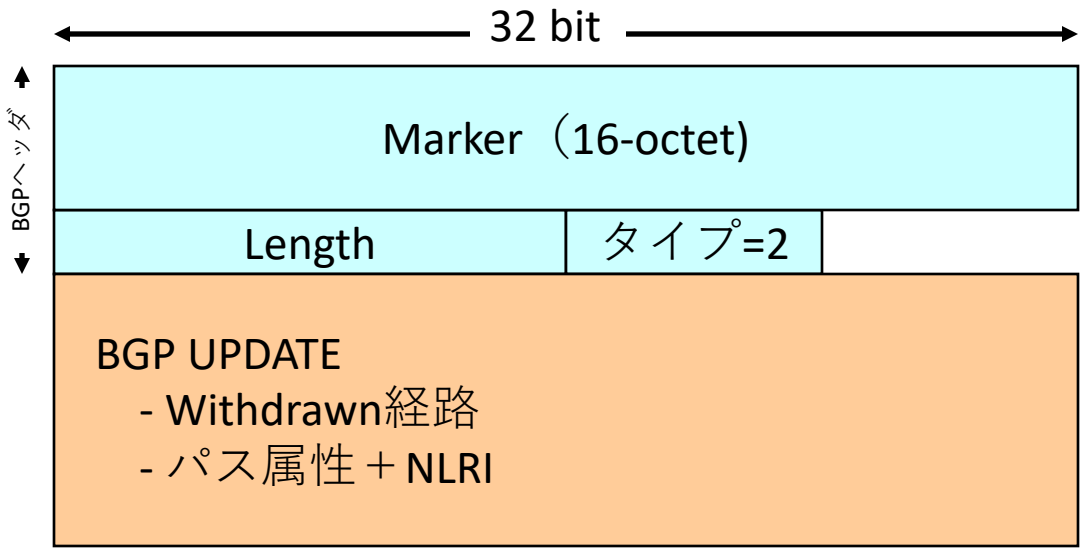
Capabilityコード

1	Multiprotocol Extension	サポートする<AFI, SAFI>の広告
2	Route Refresh	rfc版のRoute Refresh機能広告
3	Cooperative Route Filtering	
4	Multiple routes to a destination	
64	Graceful Restart	
65	Support for 4-octet AS number	
67	Support for Dynamic Capability	
128	Route Refresh(cisco)	Cisco独自のRoute Refresh機能広告

タイプ2 UPDATEメッセージ

- 経路情報を運ぶ
- 一つのメッセージで以下の情報を運べる
  - 複数のWithdrawn(取り消された)経路
  - 同じパス属性を持つ複数のNLRI
    - Withdrawn経路に含まれる経路は、同じメッセージ中でNLRIに含まれてはならない
- 情報の伝播保証はTCP任せ

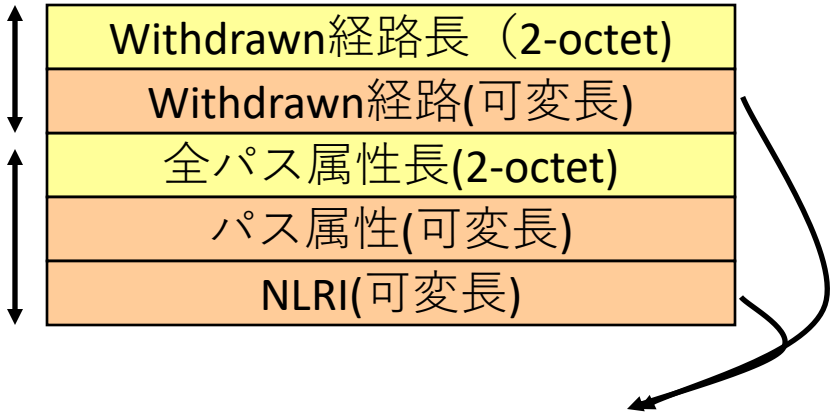
タイプ2 UPDATEメッセージ



- パス属性が異なるNLRIは、異なるUPDATEメッセージで運ばれる

BGP UPDATEフォーマット

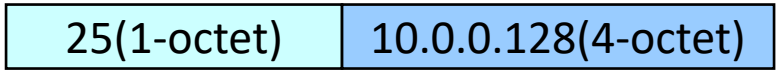
- Withdrawn経路
  - Withdrawnの長さ(2-octet)
  - Withdrawn経路の列挙
- 到達可能経路
  - 全パス属性の長さ(2-octet)
  - パス属性の列挙
  - NLRIの列挙



プレフィックスの格納形式

- 例：10.0.0.0/8

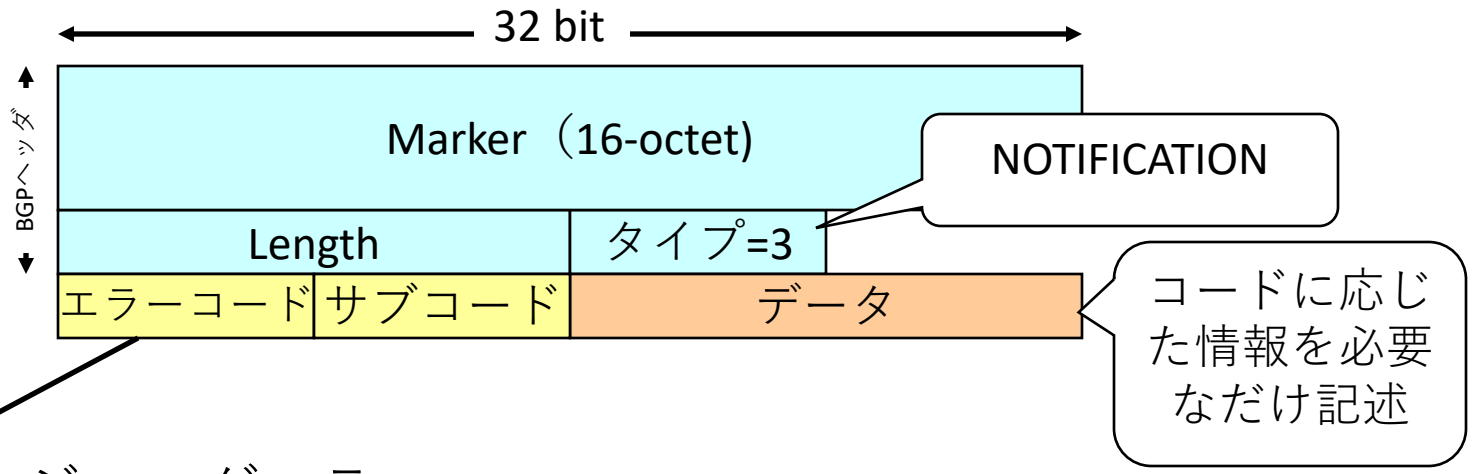
- 例：10.0.0.128/25



タイプ3 NOTIFICATIONメッセージ

- エラーを検出すると送信する
  - 送信後、すぐにBGP接続を切断する
- エラー内容がエラーコードとエラーサブコードで示される
  - 必要であれば、追加のデータも通知される

タイプ3 NOTIFICATIONメッセージ

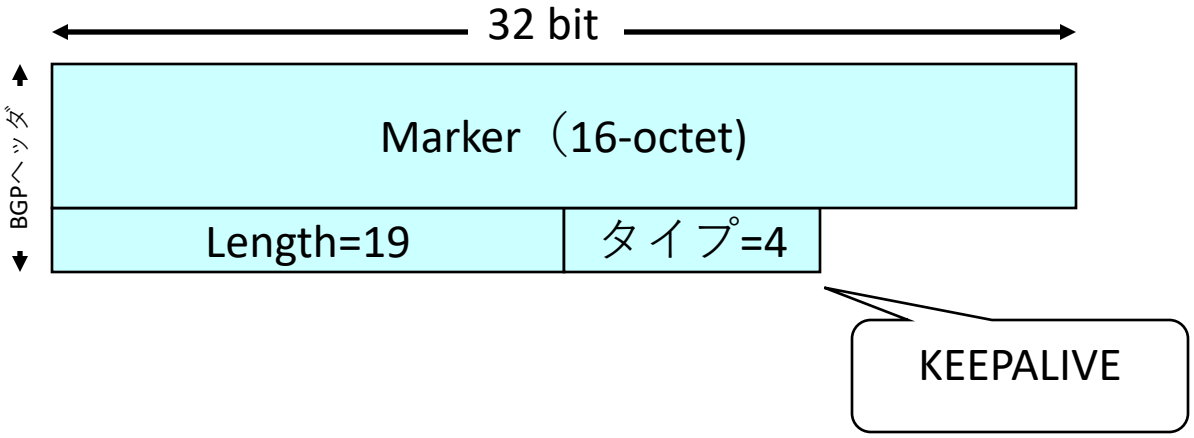


1. メッセージヘッダエラー
2. OPENメッセージエラー
3. UPDATEメッセージエラー
4. HoldTime超過
5. 状態遷移エラー
6. Cease
7. ROUTE-REFRESHエラー

タイプ4 KEEPALIVEメッセージ

- BGP接続を確立させる
- BGP接続を維持する
  - 送信間隔内にUPDATEが無ければ送信
  - 送信間隔はホールドタイムの１/3程度
    - 最小で１秒
  - ホールドタイムが０の場合は送信してはならない

タイプ4 KEEPALIVEメッセージ

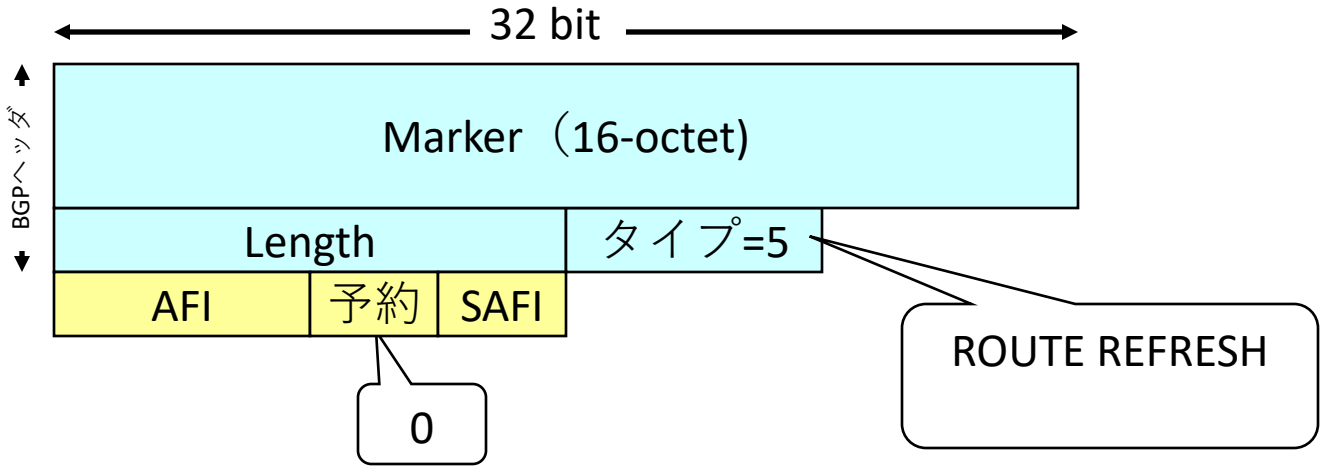


- KEEPALIVEであること以外、何も運ばない
- 最小のBGPメッセージ

タイプ5 ROUTE-REFRESHメッセージ

- 全経路の再広報を依頼する
  - <AFI, SAFI>を指定 (IPv4 unicastなど)
- 受信時、知らない<AFI, SAFI>であれば無視
- メッセージを送信するには、OPENメッセージのCapability広告でROUTE_REFRESH機能が通知されている必要がある

タイプ5 ROUTE-REFRESHメッセージ

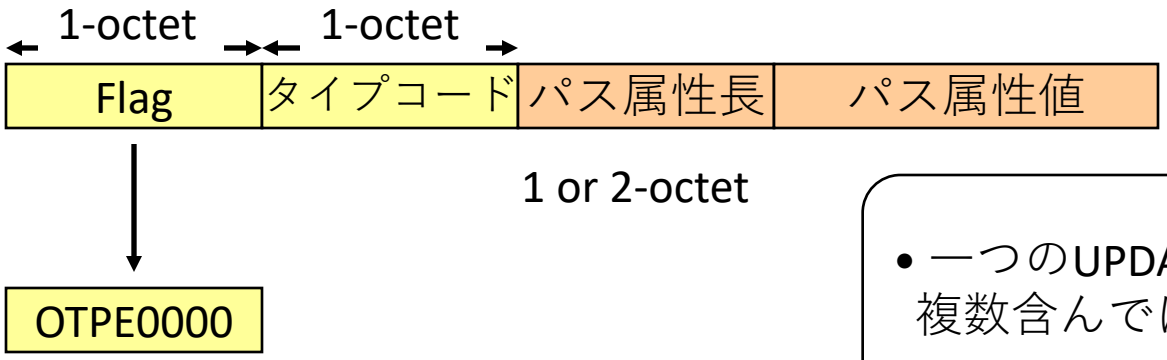


- AFI = Address Famiry Identifier
  - IPv4やIPv6など
- SAFI = Subsequent Address Famiry Identifier
  - UnicastやMulticastなど

パス属性

パス属性の構成と主要なパス属性について解説する

パス属性フォーマット



- 一つのUPDATEに同じパス属性を複数含んではいけない

O bit: Optional(パス属性の種別)

0=Wellknown, 1=optional

T bit: Transitive(パス属性の転送)

0=non-transitive, 1=transitive

P bit: Partial(パス属性の処理)

0=complete, 1=partial

E bit: Extended length

0=パス属性長は1-octet

1=パス属性長は2-octet

- Partial bit
  - オプション属性が、経路が広報されてから経由した全てのルータで解釈されたかどうかを示す
  - 0:全てのルータで解釈された
  - 1:解釈されなかったルータあり

パス属性の４つのカテゴリ

- 周知必須 - well-known mandatory [T]
  - 全てのBGPルータで解釈可能
  - NLRI情報があれば必ずパス属性に含まれる
- 周知任意 - well-known discretionary [T]
  - 全てのBGPルータで解釈可能
  - 必ずしも含まれない
- オプション通知 - Optional transitive [OT]
  - 一部のBGPルータでは解釈できないかもしれない
  - 解釈できなくても、そのまま他のルータに広報する
    - この際、Partial bitを1にセットする
- オプション非通知 - Optional non-transitive [O]
  - 一部のBGPルータでは解釈できないかもしれない
  - 解釈できない場合は、他のルータに広報するとき属性を削除する

ORIGIN属性値

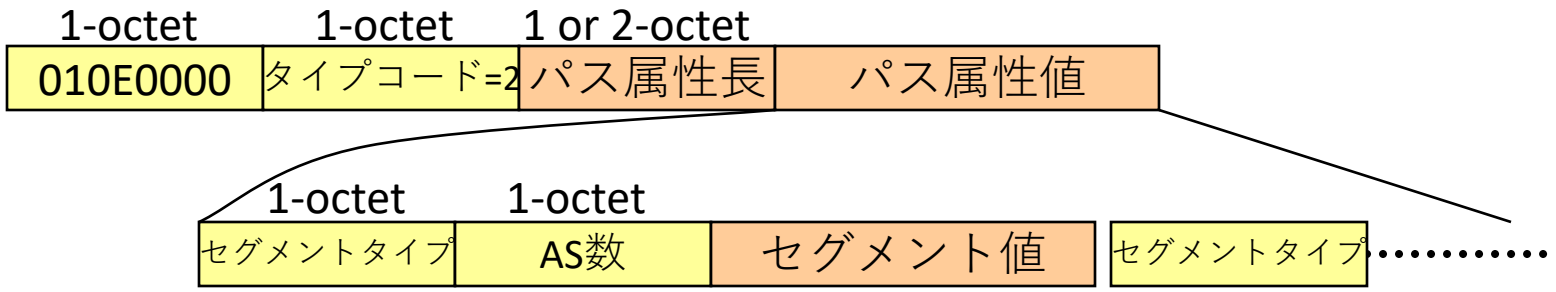
- 周知必須
- NLRIの起源を示す３つのタイプ
- 経路生成元で付加され、その後変更されない

0 – IGP ・・・ AS内部で生成
1 – EGP ・・・ EGP[RFC904]から生成
2 – INCOMPLETE ・・・その他の方法で生成

AS_PATH属性

- 周知必須
- NLRIが通過してきたAS番号のリスト
  - 例えば“10 20 30”
  - 一番右は経路を生成したAS番号
  - 他のASに広報するときに先頭に自AS番号を付加
- 用途に応じてセグメントが用意されている
  - 通常はAS_SEQUENCEを利用する
  - 異なるAS_PATHを集約した場合はAS_SET
  - AS_SETは{}でくくられる表記が多い
    - 例えば”10 20 30 {40 41}”

AS_PATH属性フォーマット



セグメントタイプ

1: AS_SET
UPDATEが経由したAS番号。順序は意味を持たない
異なるAS Pathの経路を集約したときに生成される

2: AS_SEQUENCE
UPDATEが経由したAS番号。順序に意味がある
経由した最新のAS番号はセグメント値の一番左

AS数

octet数ではなく、AS数
つまり、255個のASまで

セグメント値

2-octetのAS番号のリスト

- 新しいセグメントは先頭(左)に付加される
- ふつーはAS_SEQUENCEのみ

AS_PATH属性の処理

- 経路を転送する場合

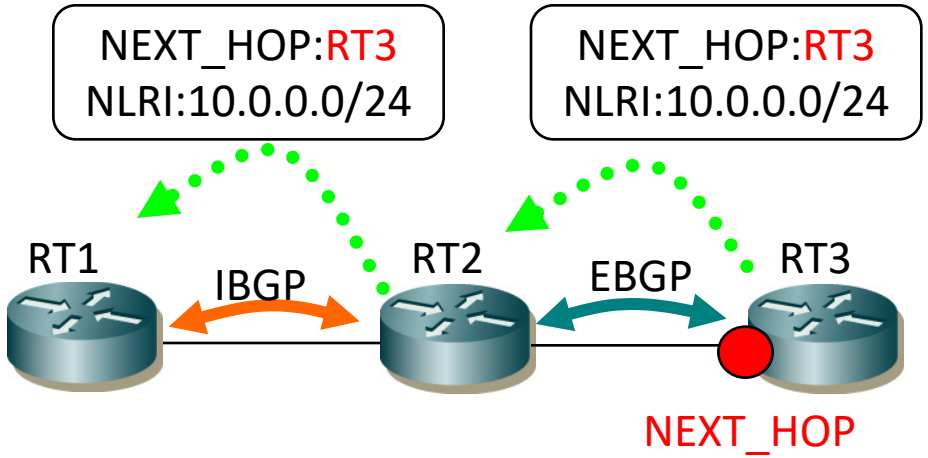
広報先	
IBGP	変更しない
EBGP	自AS番号をAS_SEQUENCEタイプでAS_PATH属性の先頭に付加する

- 経路を生成する場合

広報先	
IBGP	空のAS_PATH属性を生成する
EBGP	AS_SEQUENCEタイプで自AS番号のみのAS_PATH属性を生成する

NEXT_HOP属性

- 周知必須
- NLRIへ到達するためのネクストホップIPアドレス



NEXT_HOP属性の処理

- IBGPに経路を転送するときは
  - 変更しない
  - ただし、設定で自身のIPアドレスに変更することも可能
- IBGPに生成した経路を広報するときは
  - その宛先に到達するためのネクストホップを設定する
  - ただし、自身のIPアドレスを設定することも可能
- EBGPに経路を広報するときは
  - BGP接続に利用している自身のIPアドレスを設定する
  - ただし、宛先のネクストホップがEBGPルータと共通のサブネットに属する場合は、他のルータのIPアドレスや自身の別なインタフェースのIPアドレスを設定することも可能

MULTI_EXIT_DISC(MED)属性

- 周知任意
- 隣接ASとの距離を表す４-octetの数値
  - 小さいほど優先される
  - 付加されていないと最小の０と見なす[RFC4271]
- EBGPで受信したMEDは、他のEBGPにそのまま広報してはならない
- 幾つかの注意点
  - BGP MED Considerations [RFC4451]　など

LOCAL_PREF属性

- 周知
- AS内での優先度を示す4-octetの数値
  - 大きいほど優先される
- IBGPとEBGPで取り扱いが異なる
  - IBGPへの広報では付加されるべき
  - EBGPへの広報では付加してはならない
    - 付加されていた場合は無視
    - コンフェデレーションのSubAS間の場合は例外

COMMUNITIES属性

- オプション通知
- NLRIに32bitの数値で情報を付加する
  - この情報を元に予め実装したポリシ等を適用
- 上位16bitと下位16bitに分けた表記が一般的
  - 10進数で ”上位:下位” の様に表記する
  - 自ASでの制御は上位に自AS番号を用い、下位で制御の情報を付加するのが一般的
  - つまり ”asn:nn”

Well-Known-community

- (0xFFFFFF01)　**NO_EXPORT**
  - 他ASに広報しない
  - コンフェデレーション内のメンバASには広報する
- (0xFFFFFF02)　**NO_ADVERTISE**
  - 他BGPルータに広報しない
- (0xFFFFFF03)　**NO_EXPORT_SUBCONFED**
  - 他ASに広報しない
  - コンフェデレーション内でメンバASにも広報しない
- (0xFFFFFF04)　**NOPEER** [RFC3765]
  - 対等ピアには広報しない
  - まだ実装は無さそう

LARGE COMMUNITIES属性

- オプション通知
- NLRIに96bitの数値で情報を付加する
  - この情報を元に予め実装したポリシ等を適用
  - 単に大きくなったBGP COMMUNITIY
- “32bit:32bit:32bit”の10進表記
  - <自AS>:<タグ１>:<タグ２>や
  - <自AS>:<制御>:<対象AS>な利用を想定
- COMMUNITY属性と併用することを想定

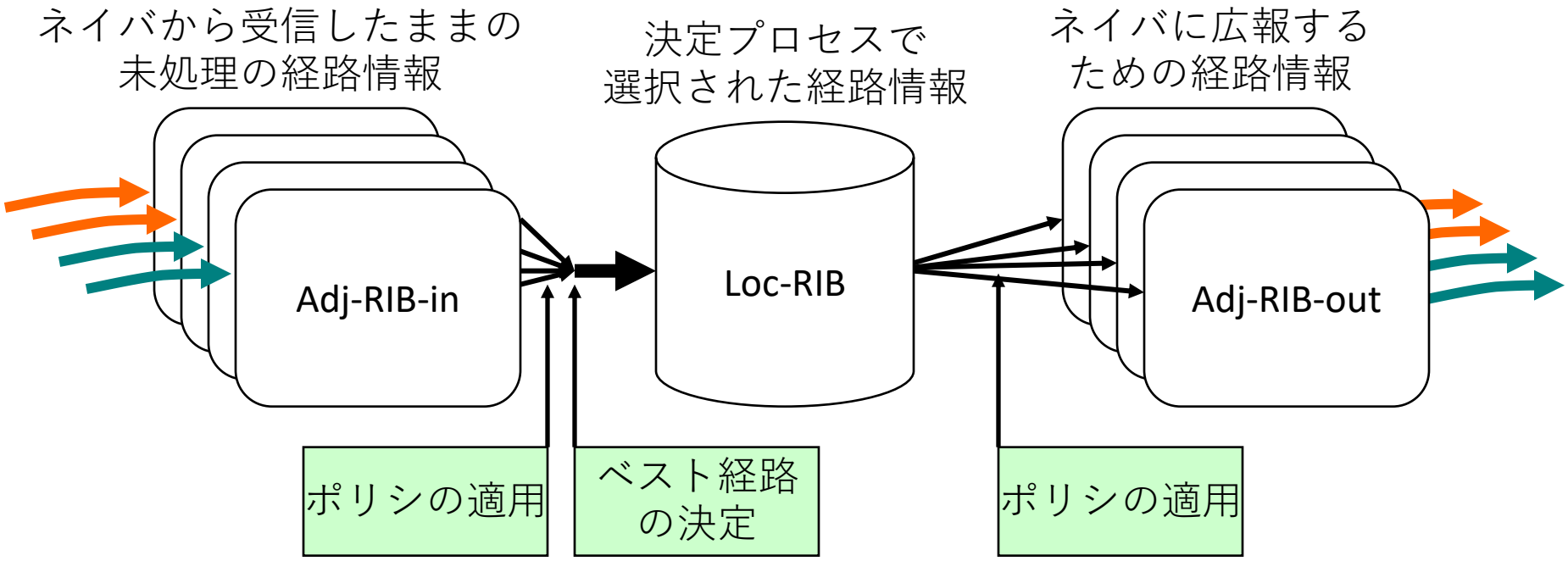
EBGP＆IBGPとパス属性

パス属性	EBGP	IBGP
ORIGIN	必須	必須
AS_PATH	必須	必須
NEXT_HOP	必須	必須
MULTI_EXIT_DISC	任意	任意
LOCAL_PREF	不許可	付加すべき
COMMUNITIES	任意	任意

BGPの経路選択

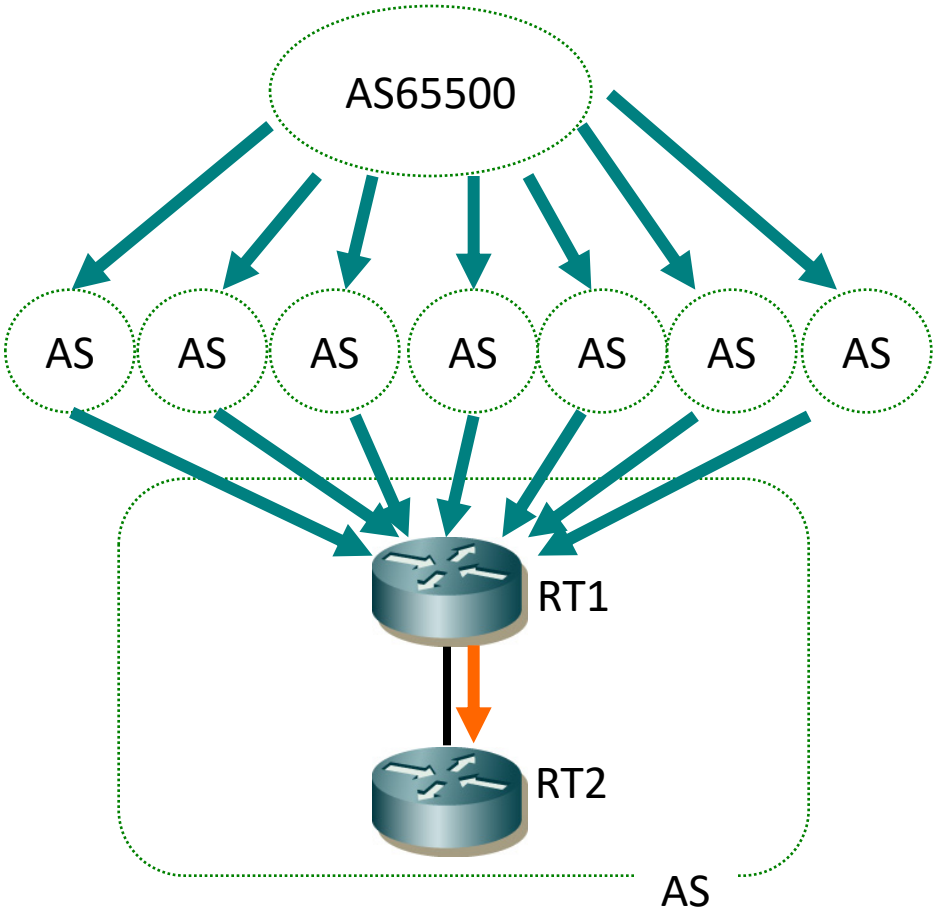
経路処理方法や、経路選択ルールを解説する

BGPの経路処理



- ポリシは設定/実装依存
- 無理なポリシを適用すると、経路ループを引き起こす可能性があるので注意

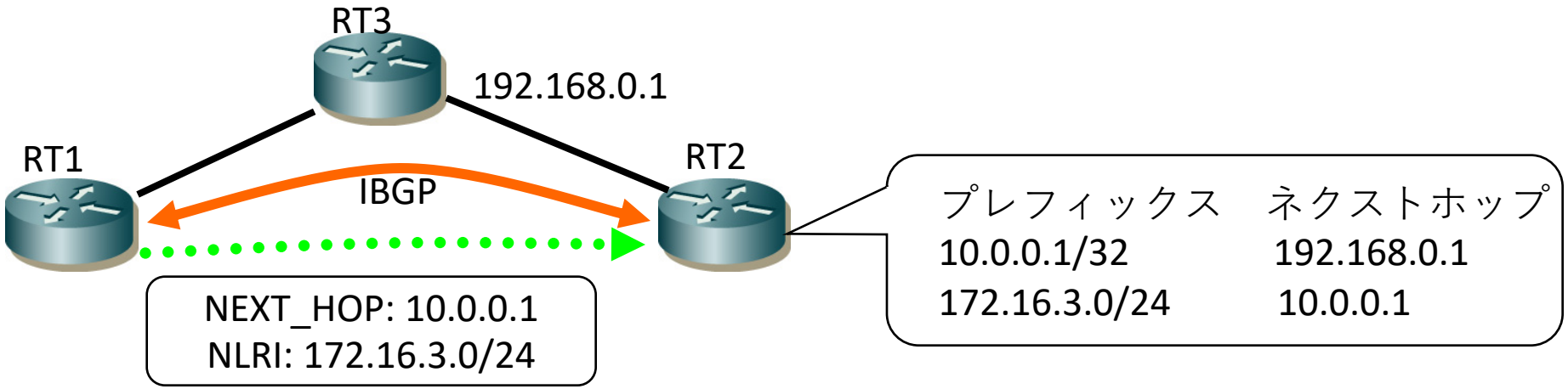
ベスト経路のみを広報



- RT1では７経路見える
  - ただし利用している経路はベストの１つだけ
- RT2へ広報されるのはRT1で選択されたベスト経路のみ
  - 経路に変更があるって最適経路が変わると、それが広報されて上書きされる

NEXT_HOP解決

- NEXT_HOP属性のIPアドレスまで到達可能であること
  - BGPも含めた経路で再帰解決して、最終的にBGPルータの隣接するネクストホップが得られる必要がある[RFC4271]



経路優先度

1	NEXT_HOP	NEXT_HOP属性のIPアドレスが到達不可能な経路は無効
2	AS loop	AS Path属性に自身のAS番号が含まれている経路は無効
3	LOCAL_PREF	LOCAL_PREF属性値が大きい経路を優先 (LOCAL_PREF属性が付加されていない場合は、ポリシに依存)
4	AS_PATH	AS_PATH属性に含まれるAS数が少ない経路を優先 (AS_SETタイプは幾つASを含んでも1として数える)
5	ORIGIN	ORIGIN属性の小さい経路を優先 (IGP < EGP < INCOMPLETE)
6	MULTI_EXIT_DISC	同じASからの経路はMED属性値が小さな経路を優先 (MED属性が付加されていない場合は、最小(=0)として扱う)
7	PEER_TYPE	IBGPよりもEBGPで受信した経路が優先
8	NEXT_HOP METRIC	NEXT_HOPへの内部経路コストが小さい経路が優先 (コストが算出できない経路がある場合は、この項目をスキップ)
9	BGP_ID	BGP IDの小さなBGPルータからの経路が優先 (ORIGINATOR_IDがある場合は、これをBGP IDとして扱う)
10	CLUSTER_LIST	CLUSTER_LISTの短い経路が優先
11	PEER_ADDRESS	ピアアドレスの小さなBGPルータからの経路を優先

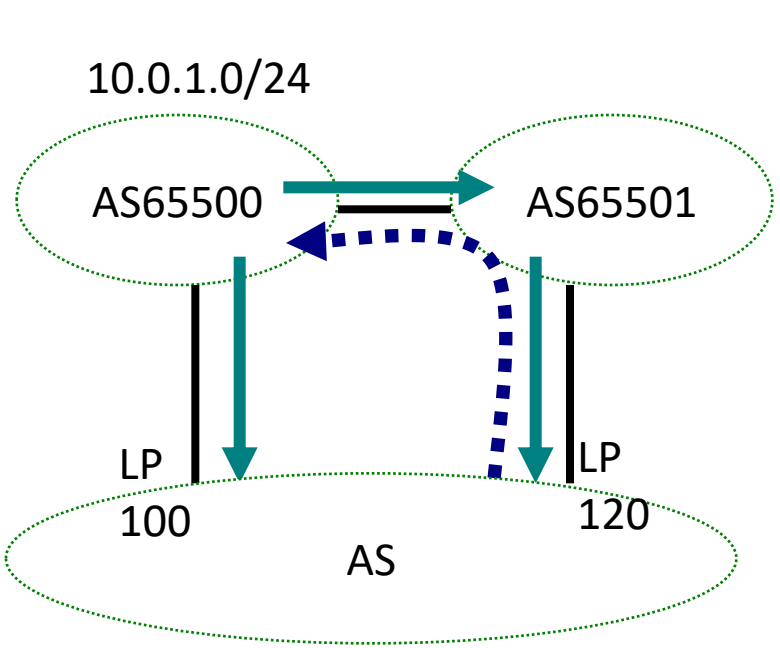
属性値の評価

属性値がどう評価されるかを
解説する

受信経路で重要な属性値

- Local Preference
  - 受信時に設定する
- AS_PATH
  - 相手ASから広報される
- MED
  - 相手ASから設定されて広報される、もしくは受信時に上書き設定する
- NEXT_HOP Cost
  - AS内部のトポロジに依存する

Local Preference

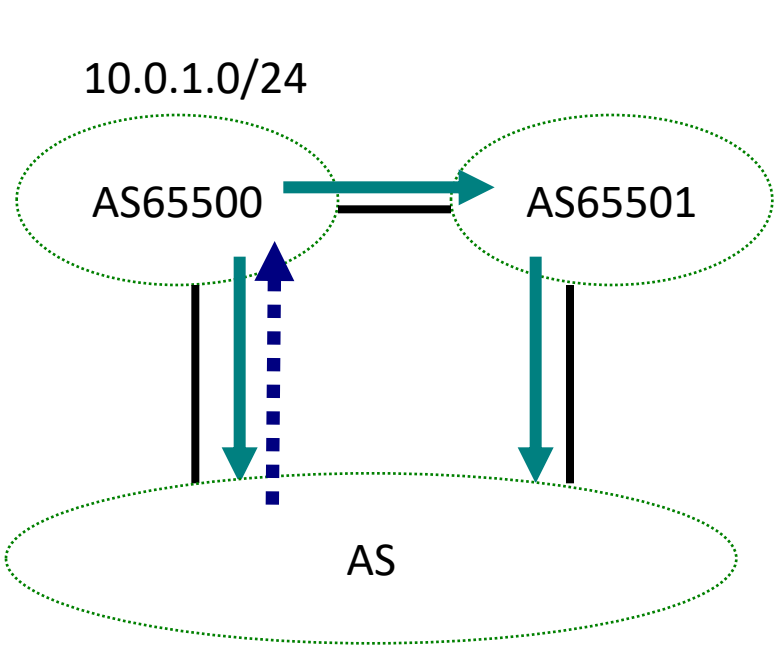


AS's view:

prefix	LocPref	AS_PATH
10.0.1.0/24	100	65500
>10.0.1.0/24	120	65501 65500

- Local Preferenceの大きな値が優先
- あるAS経由の経路を優先したい場合に有効

AS_PATH



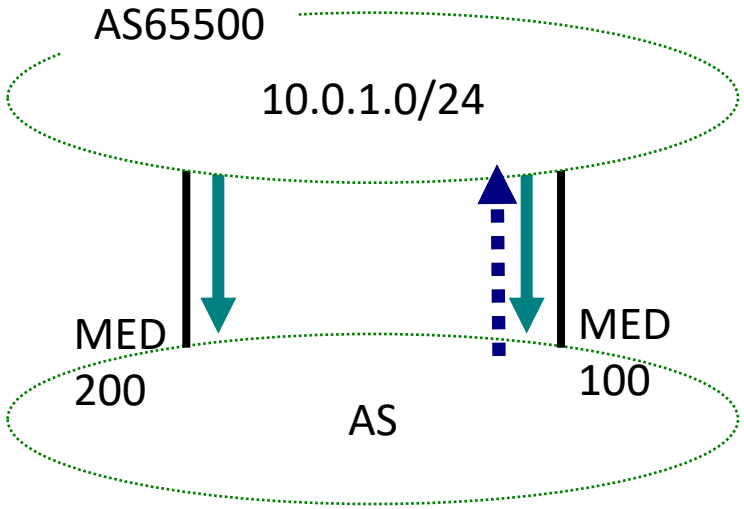
BGP UPDATE
trafficの流れ

AS's view:

prefix	AS_PATH
>10.0.1.0/24	65500
10.0.1.0/24	65501 65500

- AS_PATH長が短い経路が優先

MED(MULTI_EXIT_DISC)

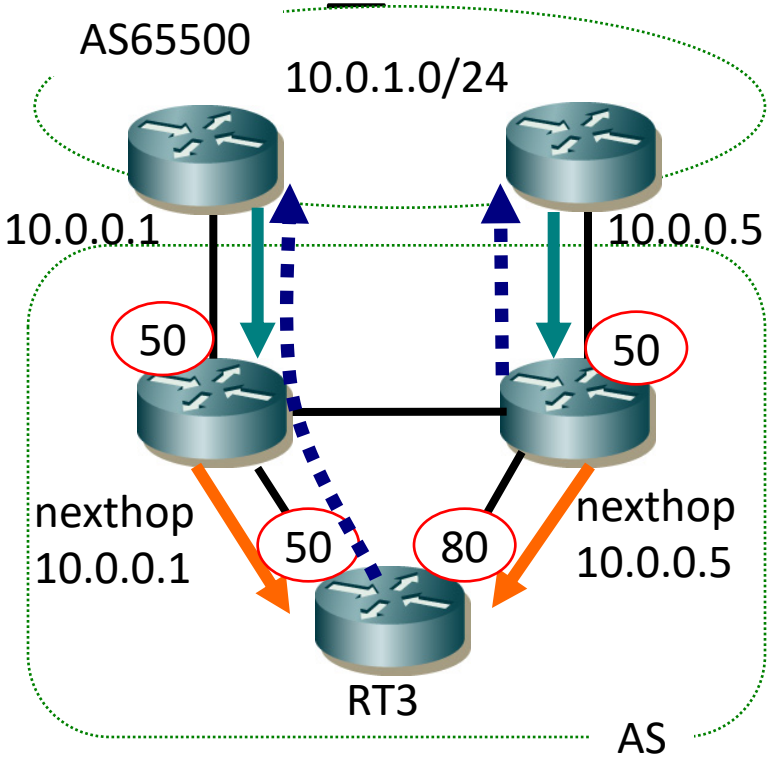


AS's view:

	prefix	MED
>	10.0.1.0/24	100
	10.0.1.0/24	200

- MEDの値が小さい経路が優先
- あるASとの複数接続に優先順位をつけたい場合に有効

NEXT_HOP COST



- NEXT_HOPへのigpコストが小さい経路を優先
- これを利用したのが closest exit

RT3's view:

prefix	nexthop [cost]
>10.0.1.0/24	10.0.0.1 [100]
10.0.1.0/24	10.0.0.5 [130]

他ASへの広報で重要な属性値

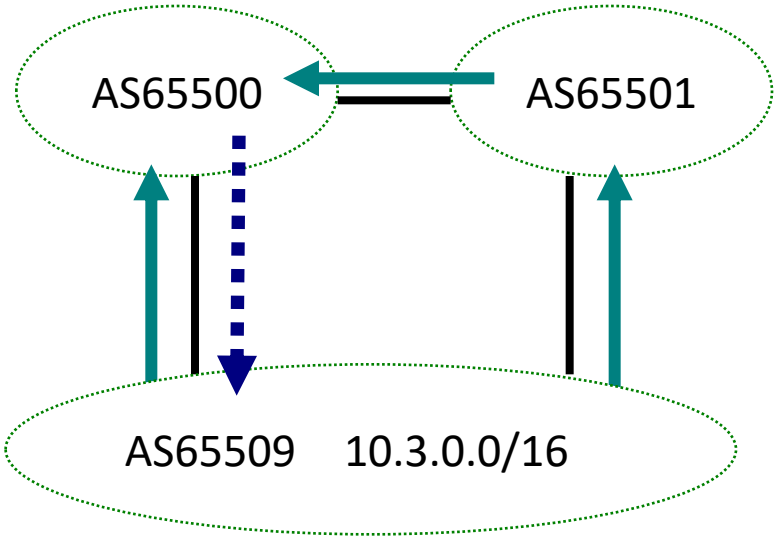
- AS_PATH
  - prependでAS Path長を伸ばす
- MED
  - 複数接続に優先順位をつける
- Community
  - 広報先ASでの処理を期待する

- 相手とのポリシのすり合わせが重要

AS_PATH (広報時)

AS65500

prefix	AS-PATH
>10.3.0.0/16	65509
10.3.0.0/16	65501 65509

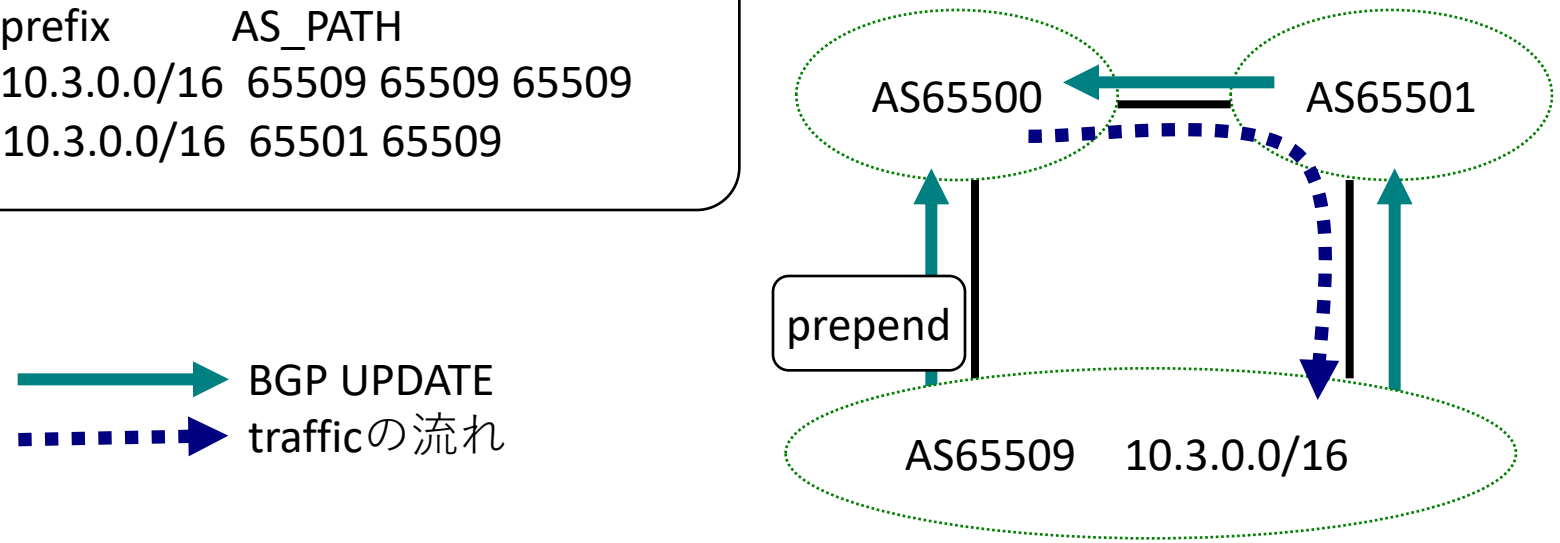


- AS_PATH長が短い経路が優先

AS_PATH prepend

AS65500

prefix	AS_PATH
10.3.0.0/16	65509 65509 65509
>10.3.0.0/16	65501 65509

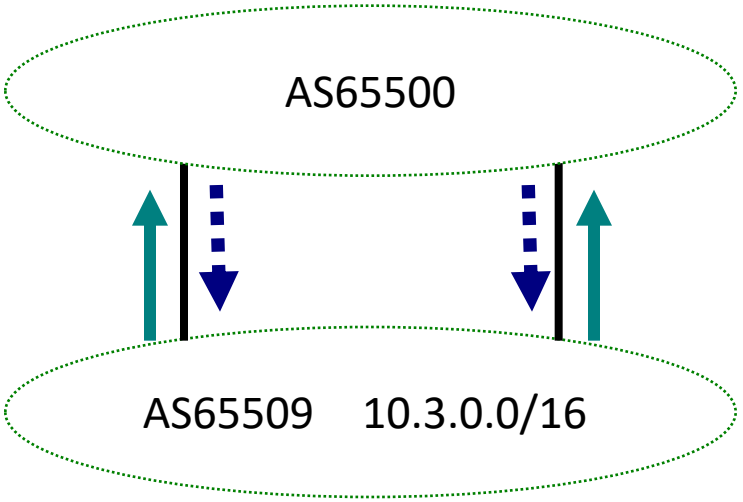


- あるASとの接続リンクを利用したくない場合に、AS_PATHを長くして優先度を下げることが出来る

広報通常時

AS65500

prefix	AS_PATH
10.3.0.0/16	65509
10.3.0.0/16	65509



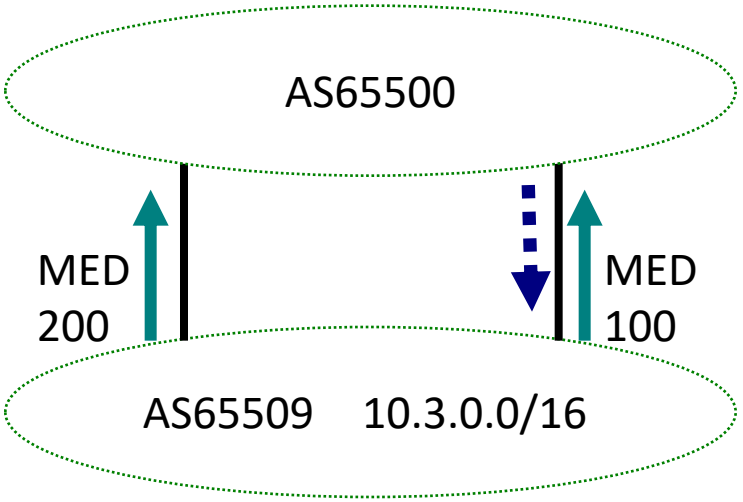
BGP UPDATE
trafficの流れ

- AS65500で特別な制御を行っていなければ、closest exitになるはず
  - トラヒックの分散は相手ASの構成に依存する

MED（広報時）

AS65500

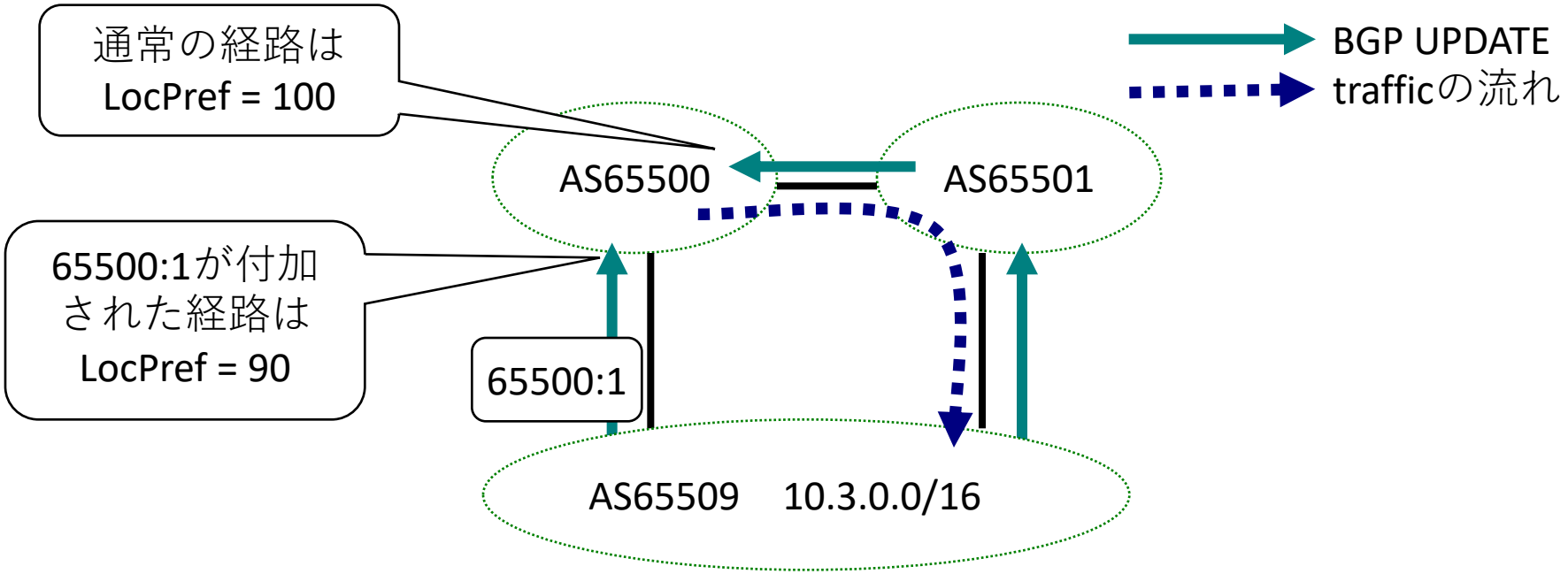
prefix	MED	AS_PATH
>10.3.0.0/16	100	65509
10.3.0.0/16	200	65509



BGP UPDATE
trafficの流れ

- 複数接続に優先順位をつけたい場合
- AS65500でMEDを受け付ける設定になっていれば、小さなMED値の経路が優先される
- MEDを受け付けるかどうかは相手ASのポリシ依存

Community利用例

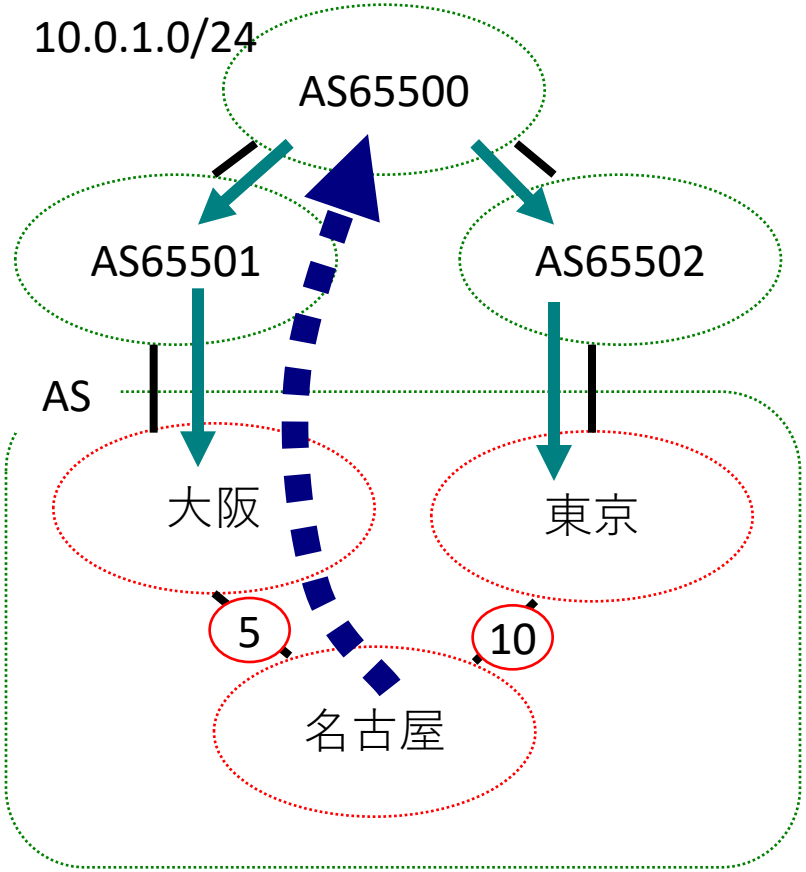


- AS65500がCommunity制御を実装していれば利用できる
- 経路にCommunity情報を付加して、その制御を利用する
- Communityを受け付けるかどうかはASのポリシ依存

BGPのパス選択

OSPFとBGPの関わりなどを
解説する

closest exitとBGP

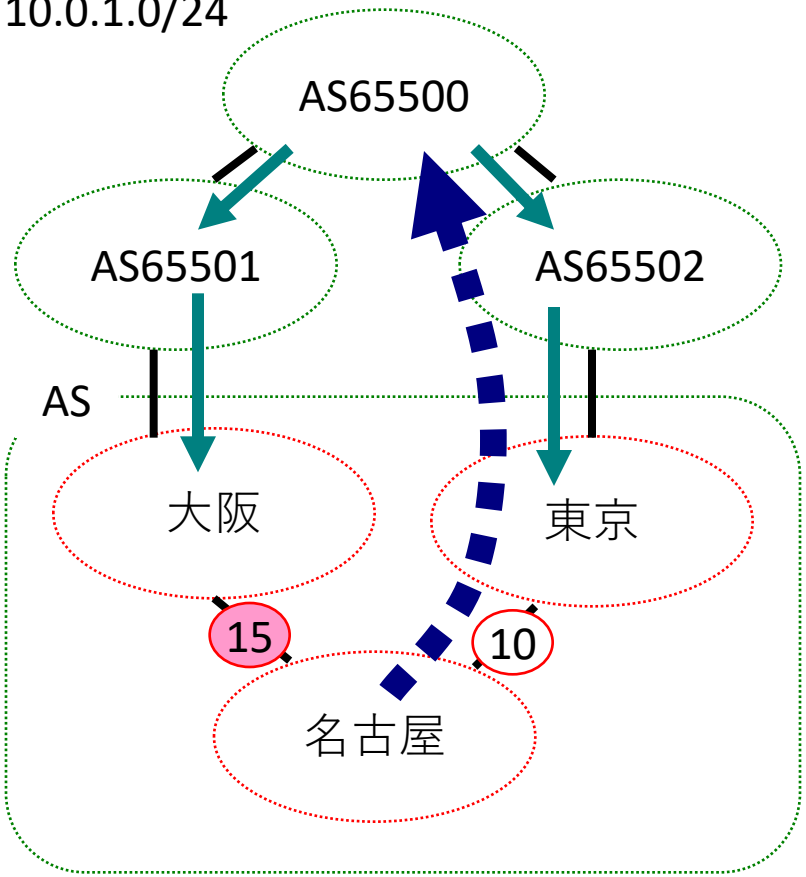


名古屋

prefix	AS_PATH	
>10.0.1.0/24	65501 65500	5
10.0.1.0/24	65502 65500	10

- 名古屋では、65501(大阪)経由を選択中

OSPFのコスト変更



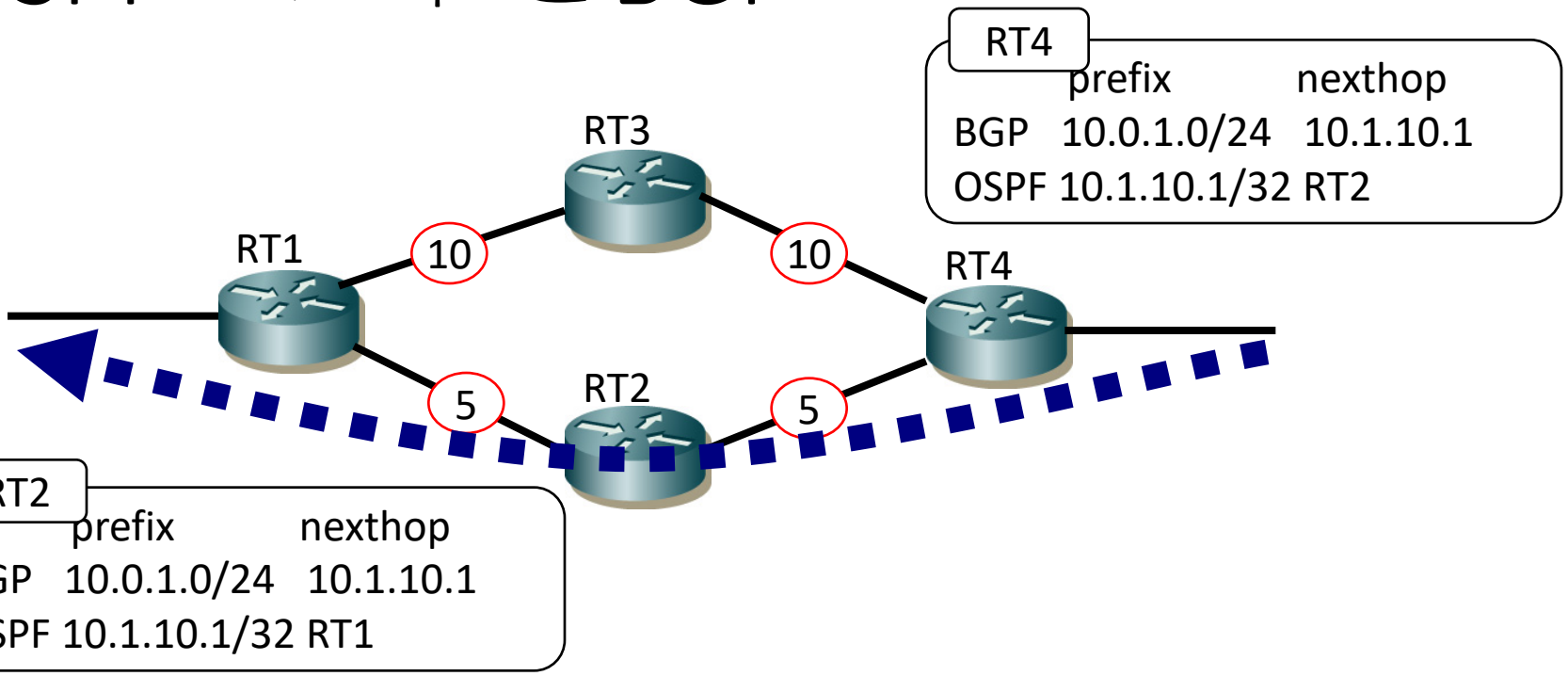
BGP UPDATE
trafficの流れ

名古屋

prefix	AS_PATH	
10.0.1.0/24	65501 65500	15
>10.0.1.0/24	65502 65500	10

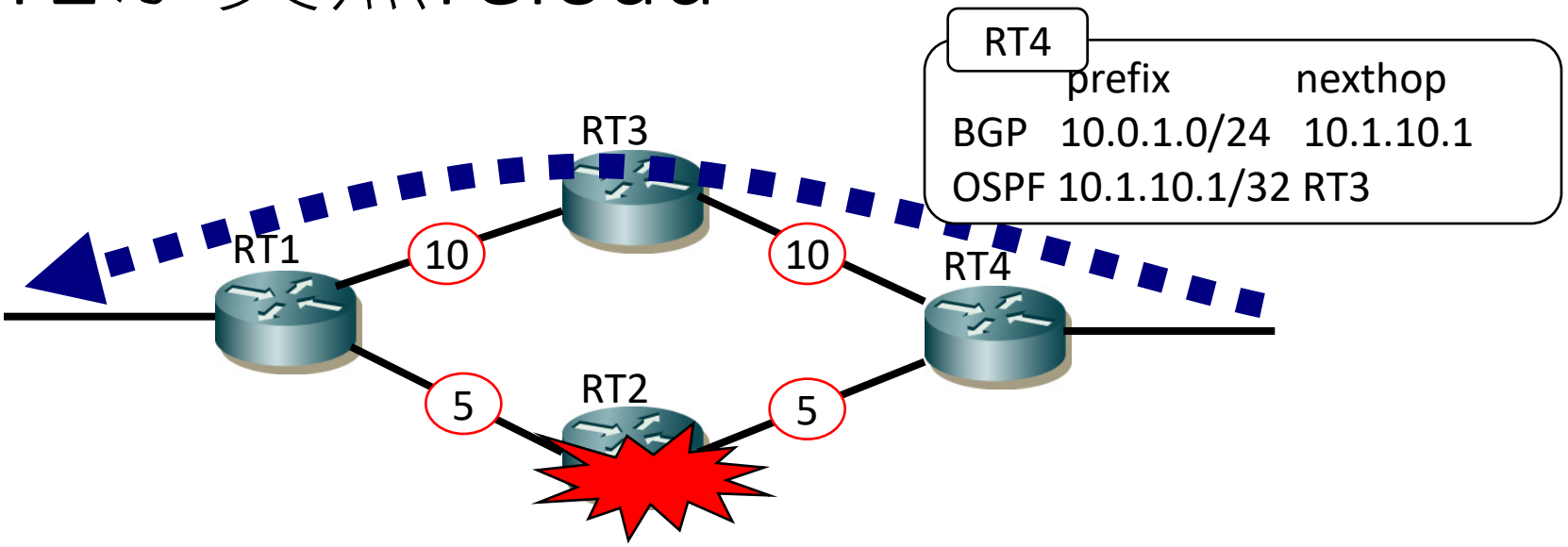
- 名古屋からは65502(東京経由)に更新

OSPFコストとBGP



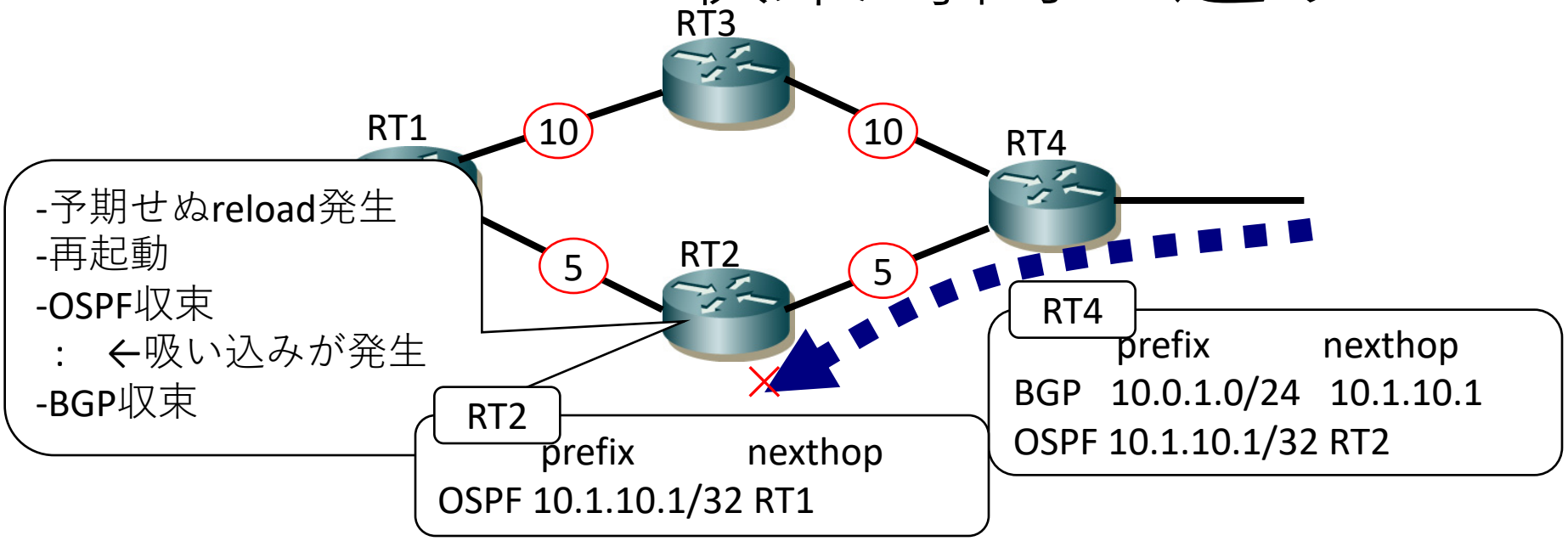
- BGPネクストホップへのOSPFコストが一番小さな経路が選択される

RT2が突然reload



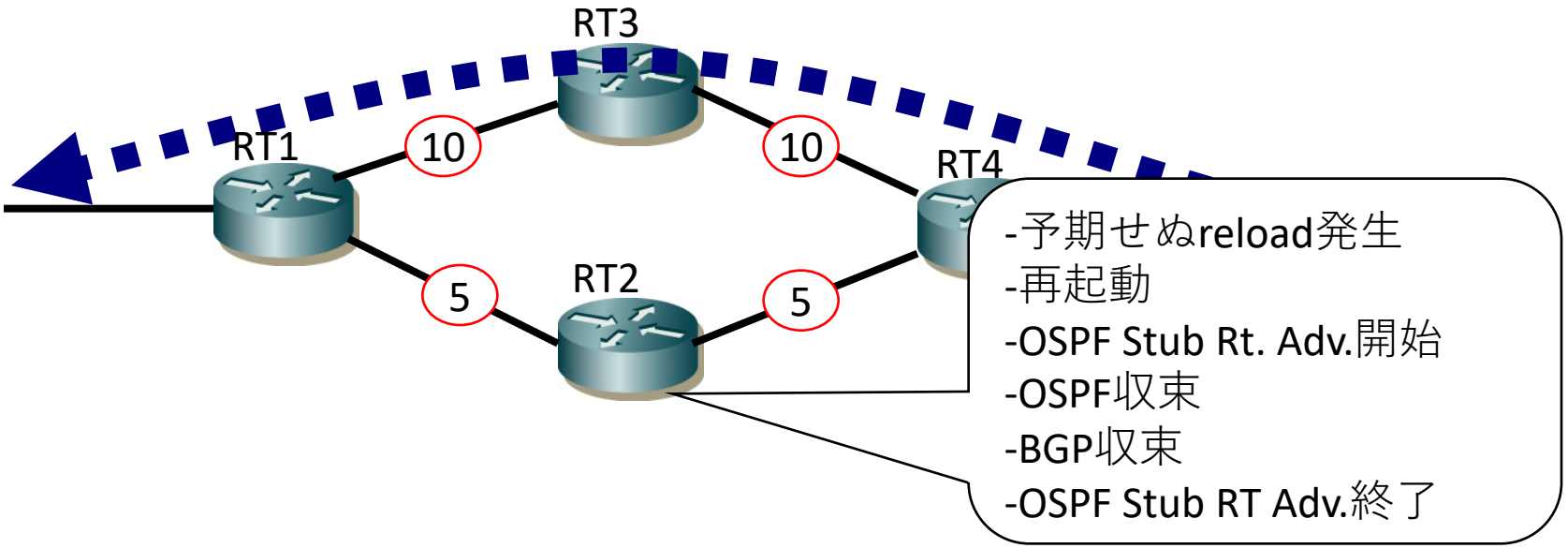
- RT２が再起動・・・
- 他のルータが障害を検出し、OSPF再計算
- トラヒックはRT３を迂回している

OSPFとBGPの収束時間が違う



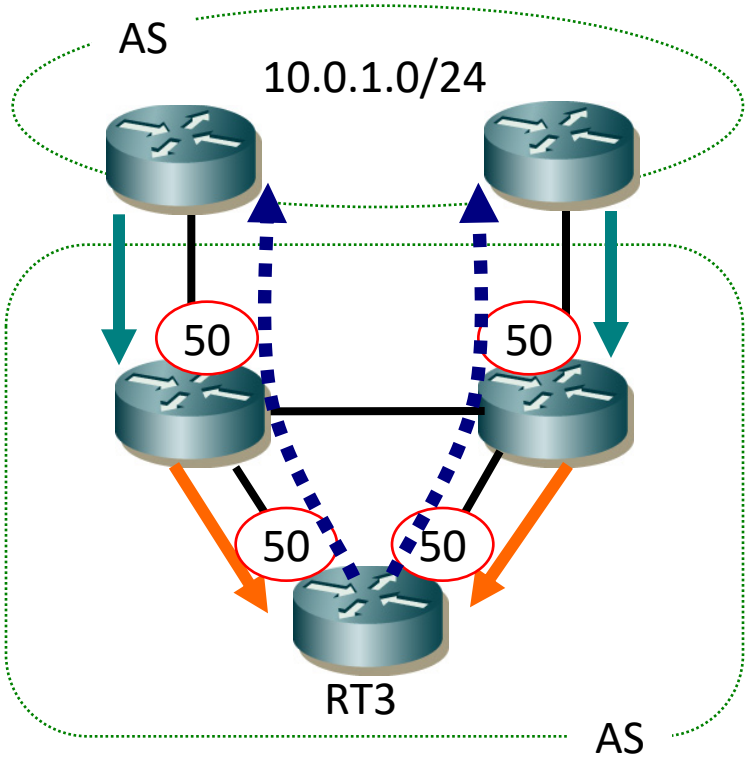
- OSPFは収束したので、RT4ではRT2側を選択
- RT２はまだBGP経路を受信しきっていない
- その間、RT2がトラヒックを破棄してしまう

OSPF StubRouterAdvertisement



- ルータを経由するトラヒックを迂回させる機能
- OSPF起動後に実施して、BGP収束までトラヒックを迂回させる等の利用が考えられる
- 詳しくは[RFC3137]を参照

BGP Multipath



- 複数の経路を有効にできる手法
  - ベンダの実装依存
  - 経路選択で特定の段階まで優先度が一致すればMultipathとして扱う
- RT3でMultipathを使用
  - RT3が他のルータに広報する経路は通常選択される１つのベスト経路のみ

BGP4+

- BGP4のマルチプロトコル(IPv6)対応
  - [RFC2545] [RFC2858]
- OPENメッセージでマルチプロトコル対応を通知
- BGPセッションはIPv4 or IPv6どちらでも可
  - IPv6だとglobal unicast or link-localが選べる
  - IPv6の到達性を保証するには、IPv6でセッションを確立するのがお勧め
- NEXT_HOPは global unicast (+ link-local)
  - プレフィックスと共にMP_REACH_NLRIで運ばれる

BGPの転用

- BGPは、ルータにTCPで情報を通知できる
- パス属性で情報を運ぶ
  - IPv6経路等もパス属性で運ばれる

  ∴パス属性のみでNLRIが無いUPDATEも有効
- 経路を運ぶ以外の目的にも利用されるようになってきた

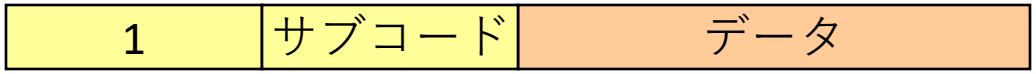
BGP NOTIFICATION メッセージ

BGP NOTIFICATIONメッセージ

1. メッセージヘッダエラー
2. OPENメッセージエラー
3. UPDATEメッセージエラー
4. HoldTime超過
5. 状態遷移エラー
6. Cease
7. ROUTE-REFRESHエラー

コード1: メッセージヘッダエラー

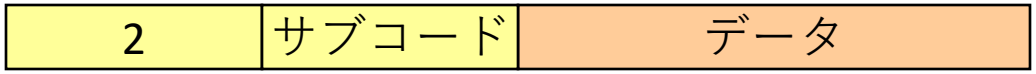
- メッセージヘッダの処理中にエラーを検出



サブコード	エラー内容	データに含まれる内容
0	特定なし	
1	Markerの値が不正	
2	Lengthの値が不正	そのLengthの値
3	解釈できないタイプ	そのタイプの値

コード2: OPENメッセージエラー

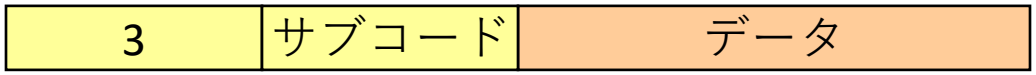
- OPENメッセージの処理中にエラーを検出



サブコード	エラー内容	データに含まれる内容
0	特定なし	
1	バージョン不一致	サポートする最も近いバージョン
2	AS番号でエラー	
3	BGP IDが不正	
4	解釈できないオプション	
5	[Deprecated]	
6	ホールドタイマ値に対応できない	
7	サポートしていないCapability	そのCapabilityコード

コード3: UPDATEメッセージエラー

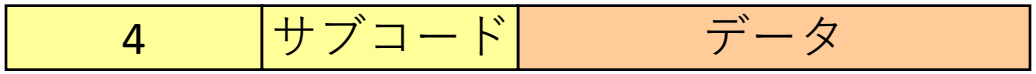
- UPDATEメッセージの処理中にエラーを検出



サブコード	エラー内容	データに含まれる内容
0	特定なし	
1	アトリビュートが不正	
2	周知必須属性が解釈できなかった	エラーを検出した属性値データ
3	周知必須属性が不足している	不足していた属性値のタイプ
4	コードフラグが不正	エラーを検出した属性値データ
5	パス属性値が不正	エラーを検出した属性値データ
6	ORIGIN属性値が不正	エラーを検出した属性値データ
7	[Deprecated]	
8	NEXT_HOP属性値の書式が不正	エラーを検出した属性値データ
9	オプション属性値でエラー	エラーを検出した属性値データ
10	NLRIの書式が不正	
11	AS_PATH属性値が不正	

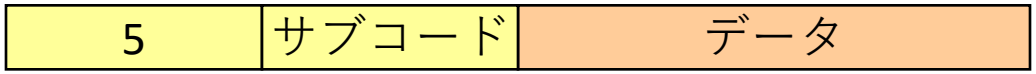
コード4: HoldTimer超過

- HoldTimer期間中に、UPDATEもKEEPALIVEも受信しなかった



コード5: 状態遷移エラー

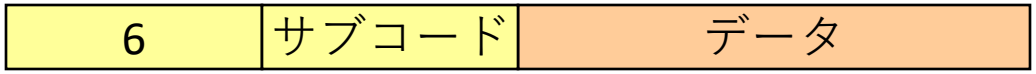
- 予期せぬイベントが発生



サブコード	エラー内容	データに含まれる内容
0	特定なし	
1	Open状態でエラー発生	
2	OpenConfirm状態でエラー発生	
3	Established状態でエラー発生	

コード6: Cease

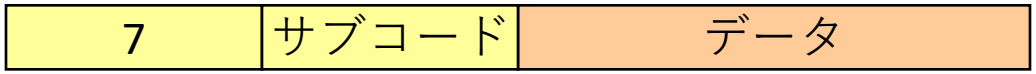
- その他のエラーを検出



サブコード	エラー内容	データに含まれる内容
1	最大受信経路数に到達	AFI, SAFI, prefix上限値
2	Administrative Shutdown	
3	設定削除	
4	Administrative Reset	
5	接続拒否	
6	その他の設定変更	
7	接続競合の解決	
8	リソース不足	

コード7: ROUTE-REFRESHエラー

- Route Refreshでエラーが発生



サブコード	エラー内容	データに含まれる内容
0	特定なし	
1	メッセージ長が不正	

BGPパス属性値
コードタイプ

BGPパス属性値コードタイプ

属性値	タイプ	概要
1 ORIGIN	周知必須	経路の生成情報
2 AS_PATH	周知必須	経路が通過したASの情報
3 NEXT_HOP	周知必須	経路のフォワード先IPアドレス
4 MULTI_EXIT_DISC	オプション非通知	複数出口から経路選定する際の優先度
5 LOCAL_PREF	周知任意	経路のローカル優先度
6 ATOMIC_AGGREGATE	周知任意	BGP経路が途中で集約された情報
7 AGGREGATOR	オプション通知	経路集約を行なったルータ
8 COMMUNITIES	オプション通知	経路に付加するタグ情報

BGPパス属性値コードタイプ続き

属性値	タイプ	概要
9 ORIGINATOR	オプション非通知	クラスタ内での経路生成ルータ
10 CLUSTER_LIST	オプション非通知	経路を反射したクラスタIDのリスト
14 MP_REACH_NLRI	オプション非通知	マルチプロトコルの到達可能経路
15 MP_UNREACH_NLRI	オプション非通知	マルチプロトコルの到達不可能経路
16 EXTENDED COMMUNITIES	オプション通知	拡張されたCOMMUNITIES(主にVPN)
17 AS4_PATH	オプション通知	古い実装で4ByteAS情報を通過させる
18 AS4_AGGREGATOR	オプション通知	古い実装で4ByteAS情報を通過させる
32 LARGE_COMMUNITY	オプション通知	経路に付加するタグ情報

特別なAS番号一覧

AS番号	用途
0	使ってはならない(RFC7606)
64496-64511	文書記載用AS番号
64512-65534	プライベートAS番号
65535	使うべきではない(RFC7300)
65536-65551	文書記載用AS番号
65552-131071	予約
4200000000-4294967294	プライベートAS番号
4294967295	使うべきではない(RFC7300)