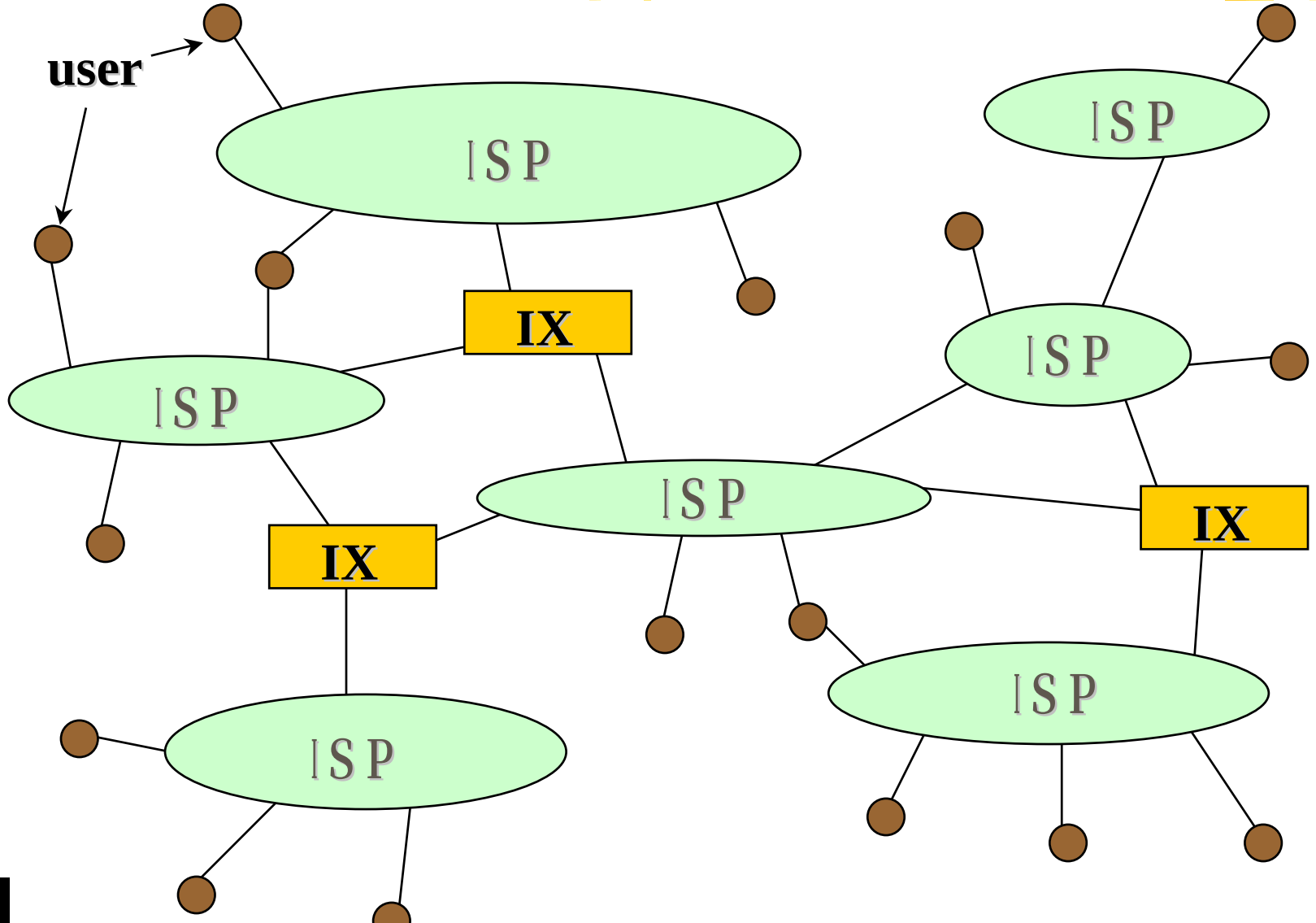
ＢＧＰ４
- プロトコルの概要と運用 -

（株）インターネットイニシアティブ
浅羽登志也
asaba@iij.ad.jp

グローバルな経路制御の概要

現状、問題点、解決策

インターネット全体の構造



ISPとは？

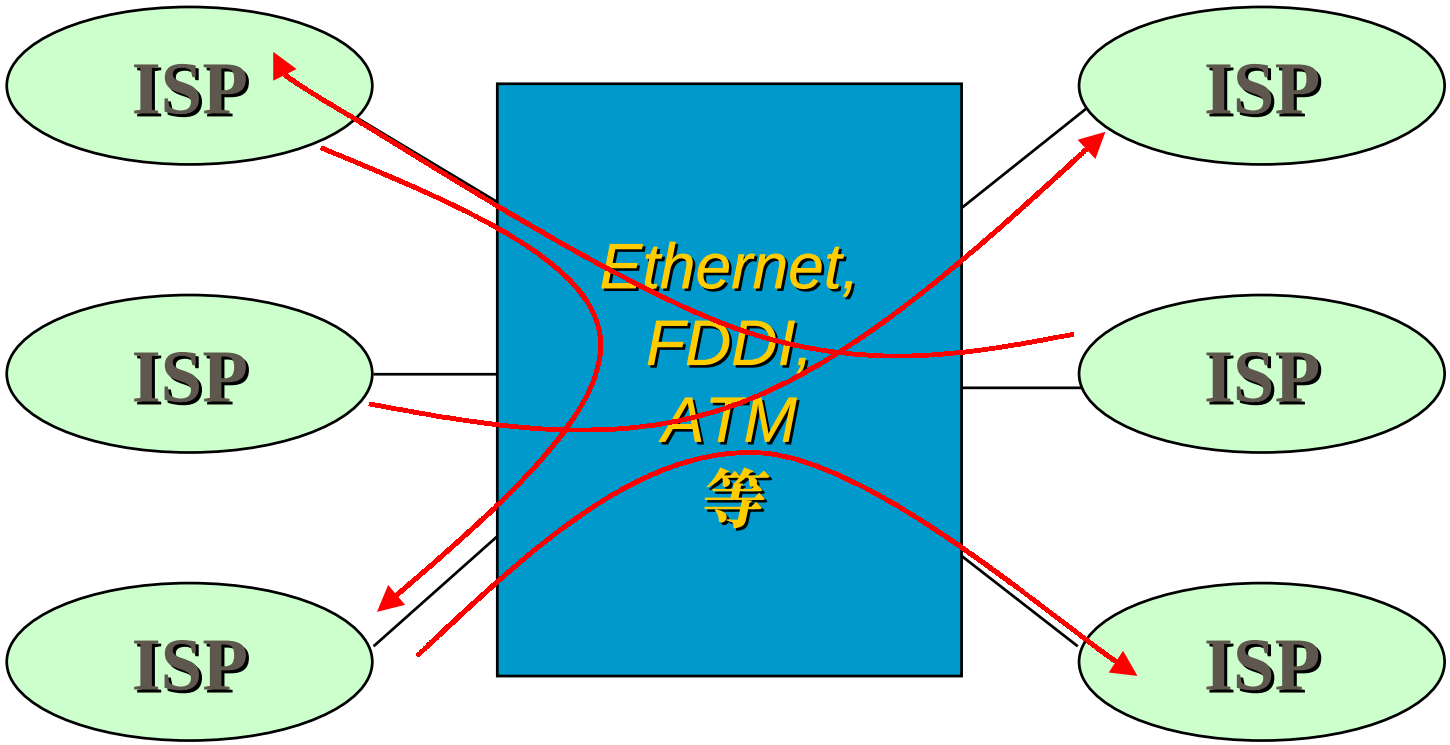
- インターネットへのコネクティビティーを提供
- 複数のISP同士が相互接続して全体を構成
  - “インターネット”
  - 相互接続形態
    - IX経由の接続
    - 直接接続
- ユーザはいずれかのISP経由でコネクティビティーを得る

IIJ

インターネットエクスチェンジ

- 複数ISP間の相互接続を提供するサービス
  - IX (Internet eXchange)
  - ISP同士がトラフィックを交換する場
  - イーサネット、FDDI、ATMなどのマルチアクセス型のデータリンク接続サービス
  - 同じデータリンクメディアを経由して複数ISPと接続が可能
  - 例
    - Network Access Point (NAP)
    - Metropolitan Area Exchange (MAE)
    - LINX, NSPIXP,JPIX, MEX,HKIX, etc.

IXの概念図



経路制御とは？

- インターネットに接続された任意の2ユーザ間の、ネットワーク層での接続性の確立
  - アドレッシング
  - 経路情報の交換
- インターネット上のトラフィックの流れの制御
  - ロードバランス
  - 代替経路の選択
  - ボトルネックの解消

経路制御の階層

- 2階層の経路制御
  - ISPの内部、ISP間
- Interior Gateway (or Routing) Protocol (IGP)
  - コストに基づく経路選択
    - OSPF, RIP2
- Exterior Gateway (or Routing) Protocol (EGP)
  - ポリシーに基づく経路選択
    - BGP4

IIJ

スケーラビリティの問題

- 2つの問題
  - アドレス空間の枯渇
  - 経路表（ルーティングテーブル）の爆発
- 短期的解決策
  - CIDR（Class-less Inter-Domain Routing）の推進
  - プライベートアドレスの活用 (RFC1918)
- 長期的解決策
  - IPv6 (RFC1883)
    - アドレス空間の拡張（32ビット→128ビット）
  - 階層的なアドレス割当と経路制御の推進

Classless Inter-Domain Routing

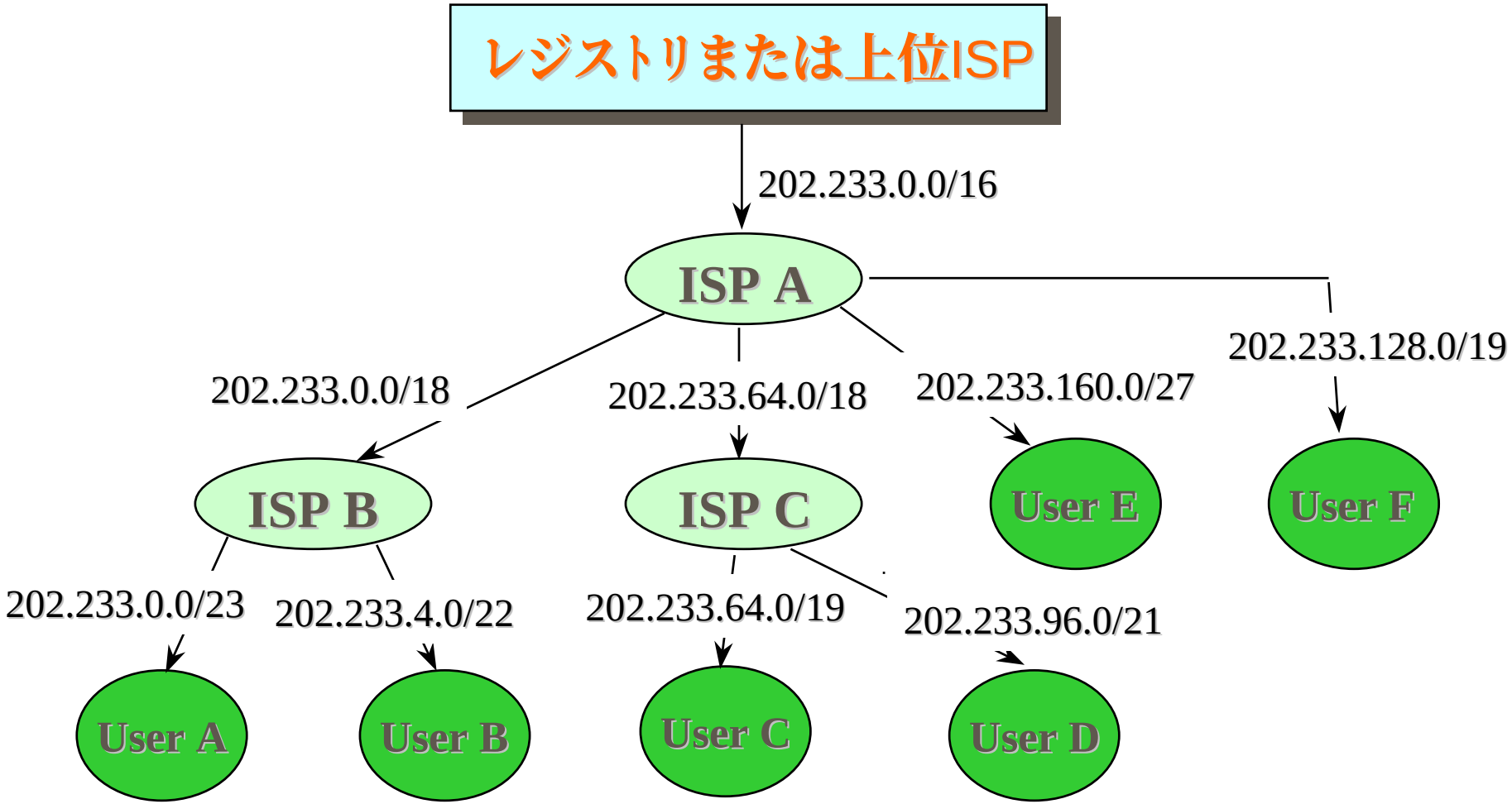
- 目的
  - クラスの概念による弊害の払拭
  - IPv4のアドレススペースの有効利用
  - 経路表のエントリ数の縮少
- 階層的アドレス割当
  - ビット境界に促したアドレス割当
- 経路情報の集成
- アドレスプレフィックス表記
  - 202.232.68.0 - 202.232.68.63 = 202.232.68.0/26

IIJ

Classlessな経路制御

- VLSMのサポート
  - インターフェース / 経路表 / 経路制御プロトコル
- Supernetのサポート
  - アドレス / 経路情報の集成
- “*Classfull*”なアドレス割当と経路制御の概念の排除
  - all-0サブネット, all-1サブネット等
- Classlessな経路情報
  - ネットマスク長の伝播

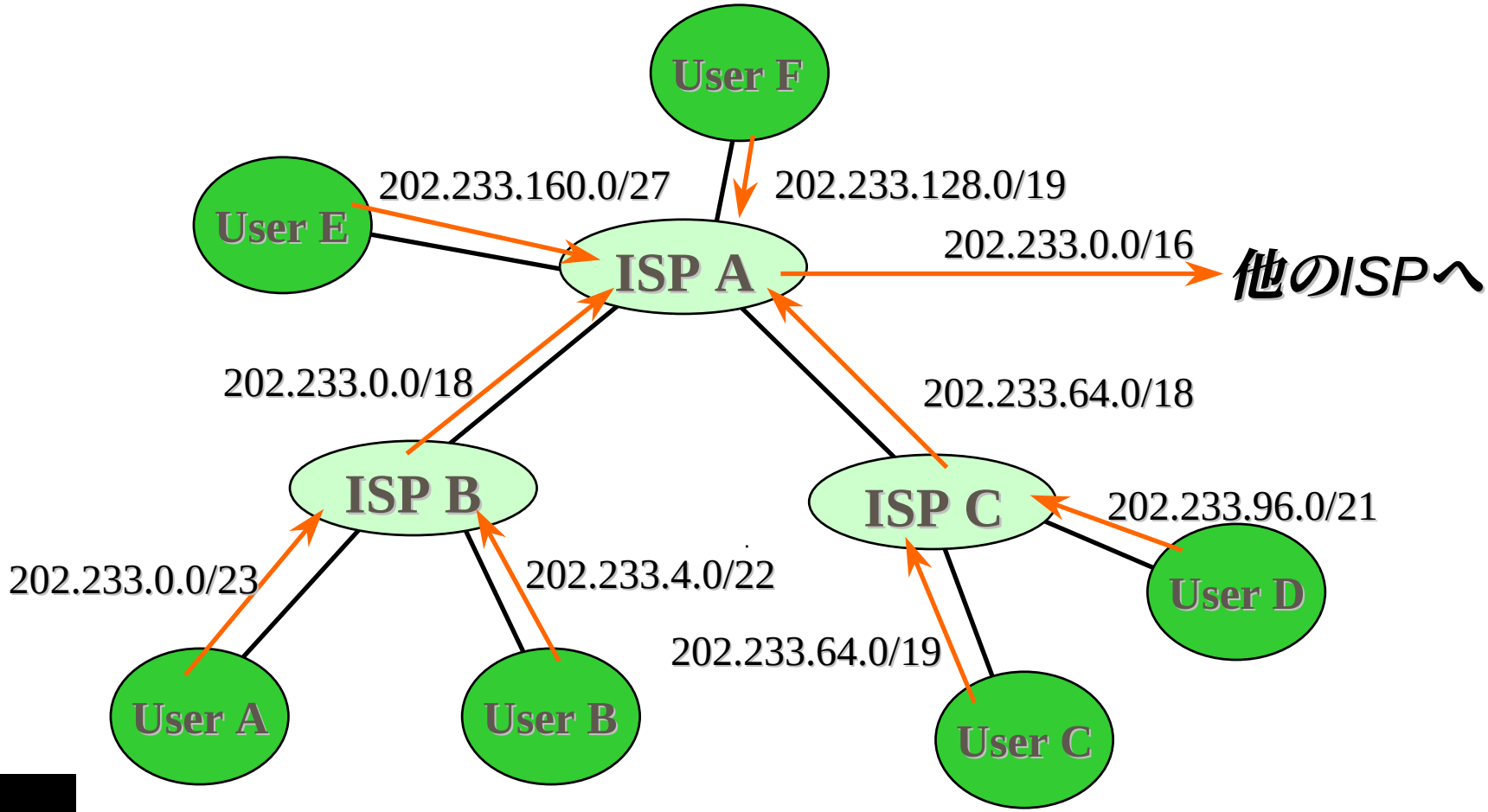
階層的なアドレス割当



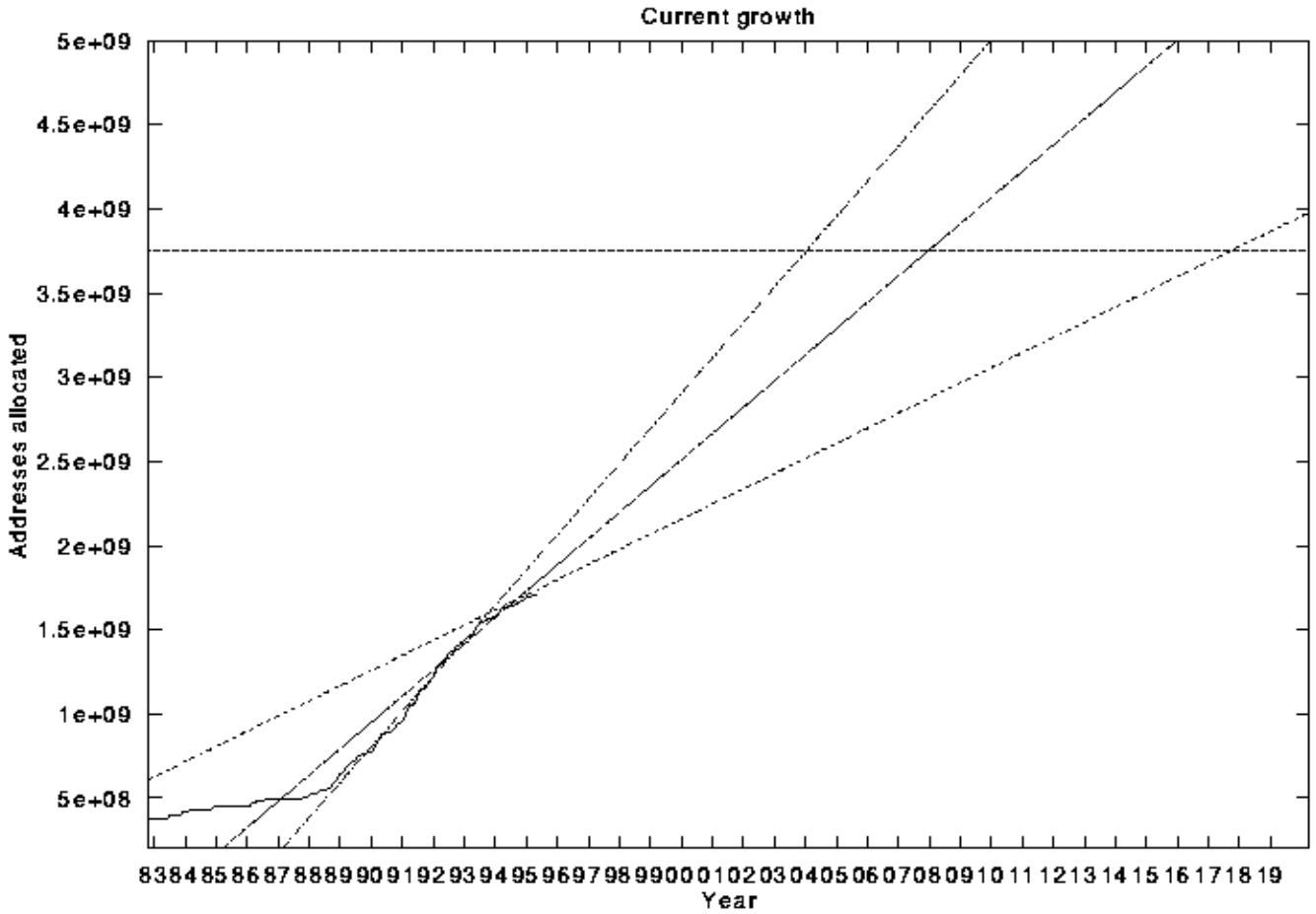
IIJ

経路情報の集成

- ネットワークトポロジに応じた階層的な集成

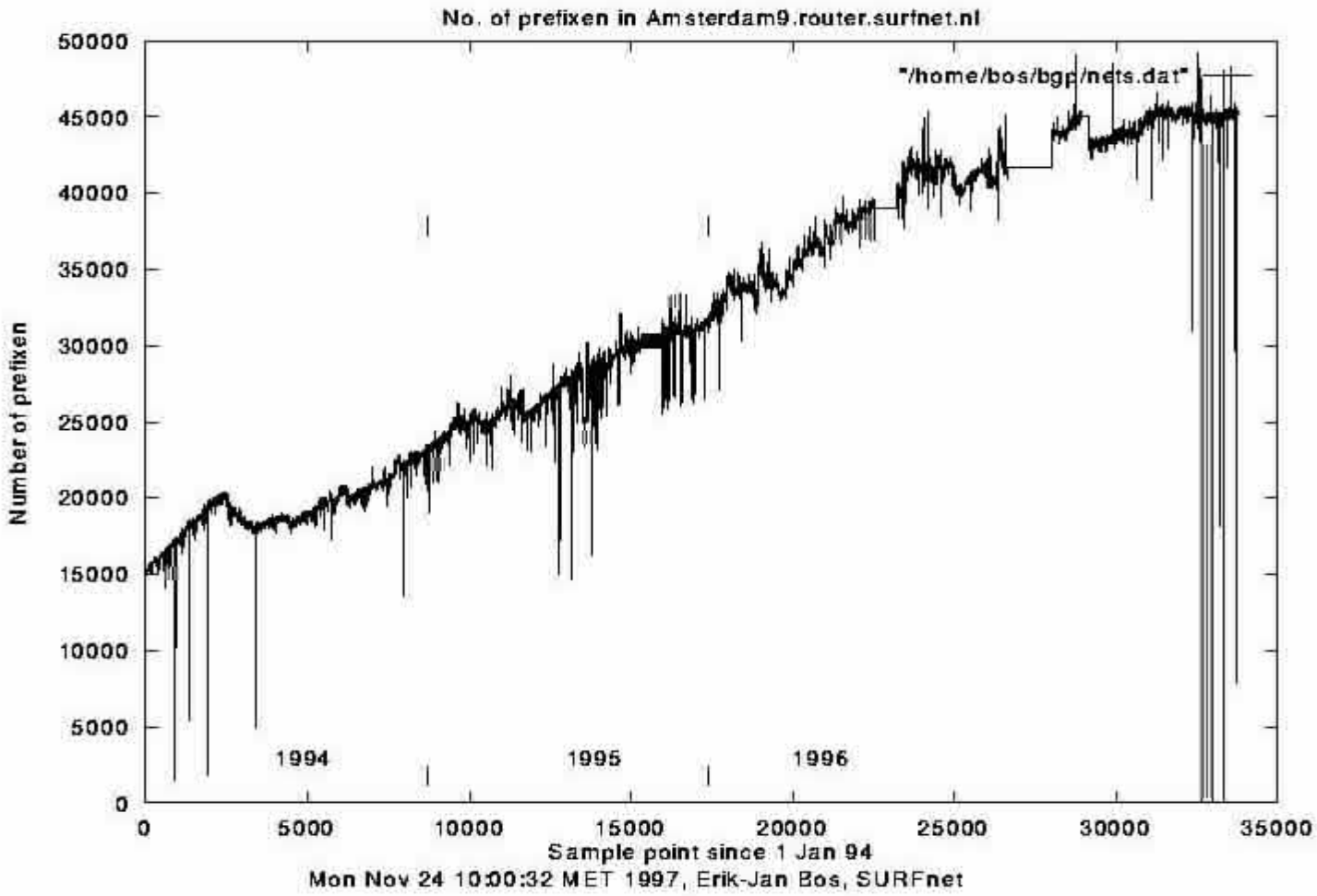


アドレス利用状況



経路表の増大状況

<SURFNET, Eric-Jan Boss氏作成>



BGP4の概要

BGP4(Border Gateway Protocol)

- RFC1771
- AS間経路制御のde-facto標準プロトコル
  - Autonomous System (AS)
    - 単一の管理主体により、単一の経路制御ポリシーにのもとで管理･運用される範囲
    - ISP　AS
    - 現在のインターネットはASの集合体とみなすことが可能
- CIDRのサポート
  - CIDRの実現に不可欠

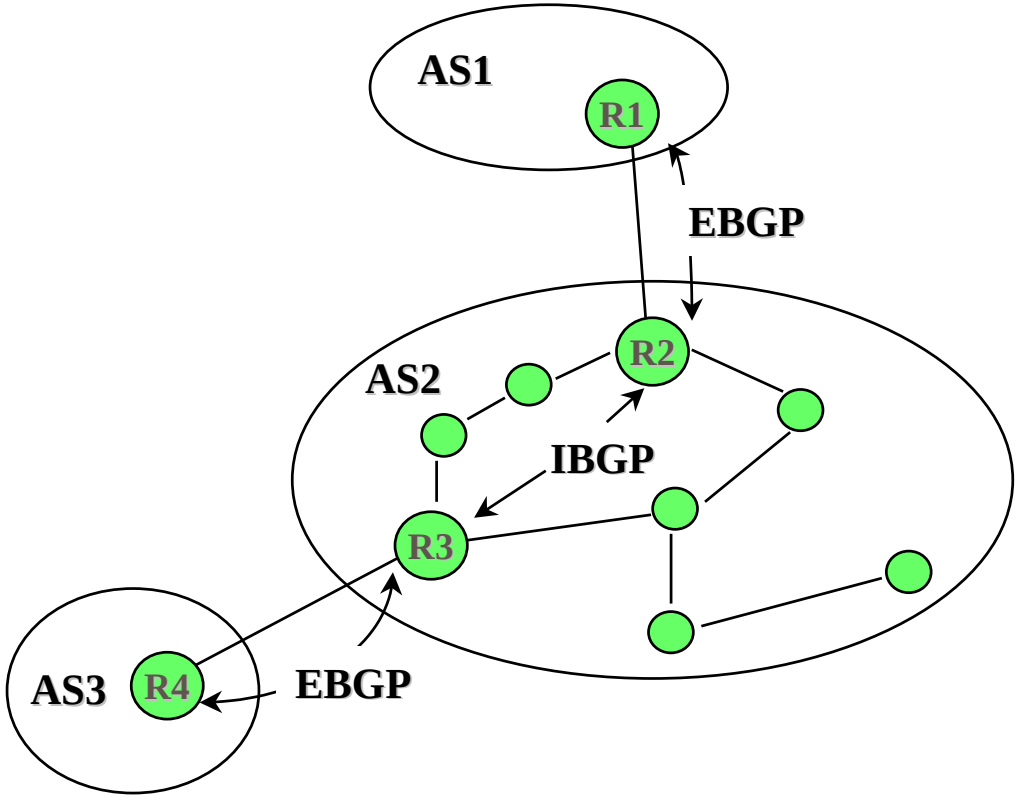
特徴

- TCP (ポート179) を用いる
  - コネクションを張ったルータ間(peer)で1対1の経路情報の交換
  - 経路情報の交換に信頼性を保証
  - RIP等と異なり、Incrementalな情報交換
- 16ビットのAS番号 (例：IIJはAS2497)
- Path Vector方式の経路制御プロトコル
  - 経路情報に付加されたパス属性 (Path Attribute)に基づく経路選択
  - AS Path、Origin、Next Hop、Multi-Exit-Discriminator(MED)、Local Preference、etc.

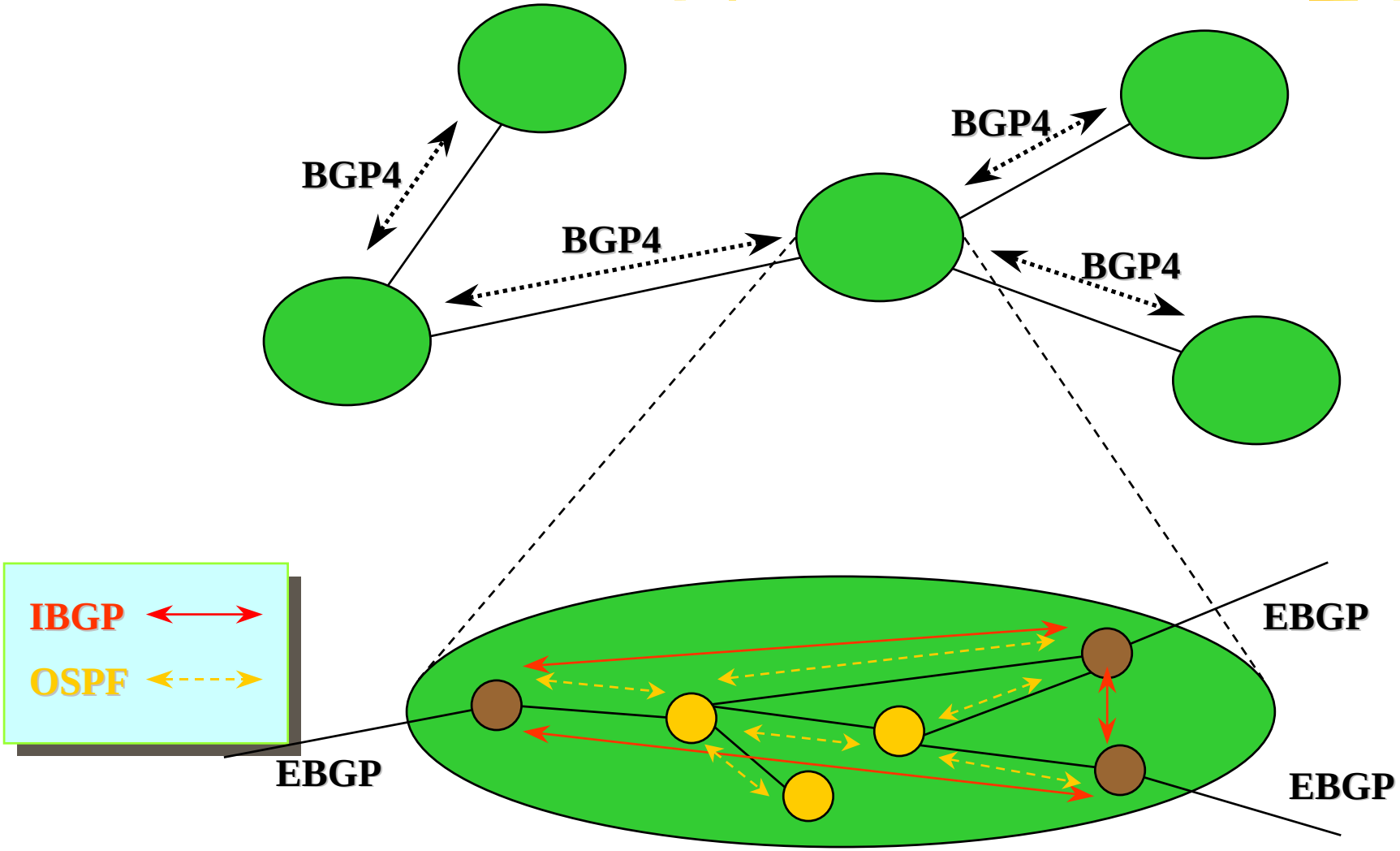
EBGPとIBGP

- BGPスピーカ (ボーダールータ)
  - BGPを用いて経路交換をするルータ等
- EBGP (External BGP)
  - 異なるASに属するBGPスピーカ間のBGPセッション
- IBGP (Internal BGP)
  - 同一AS内部のBGPスピーカ間のBGPセッション
    - full mesh
    - BGPスピーカ間で学んだ経路情報を交換する
    - 他のIBGPスピーカから学んだ経路は伝播しない

BGPを用いたAS間の経路交換



AS外部とAS内部の経路制御



パス属性

- 伝播された各経路の属性を示す
  - 複数経路からの経路選択に用いる
  - ポリシーを表す
- 通過型(Transitive)属性と非通過型(Non-Transitive)属性
- 必須(Mandatory)属性と任意(Optional)属性

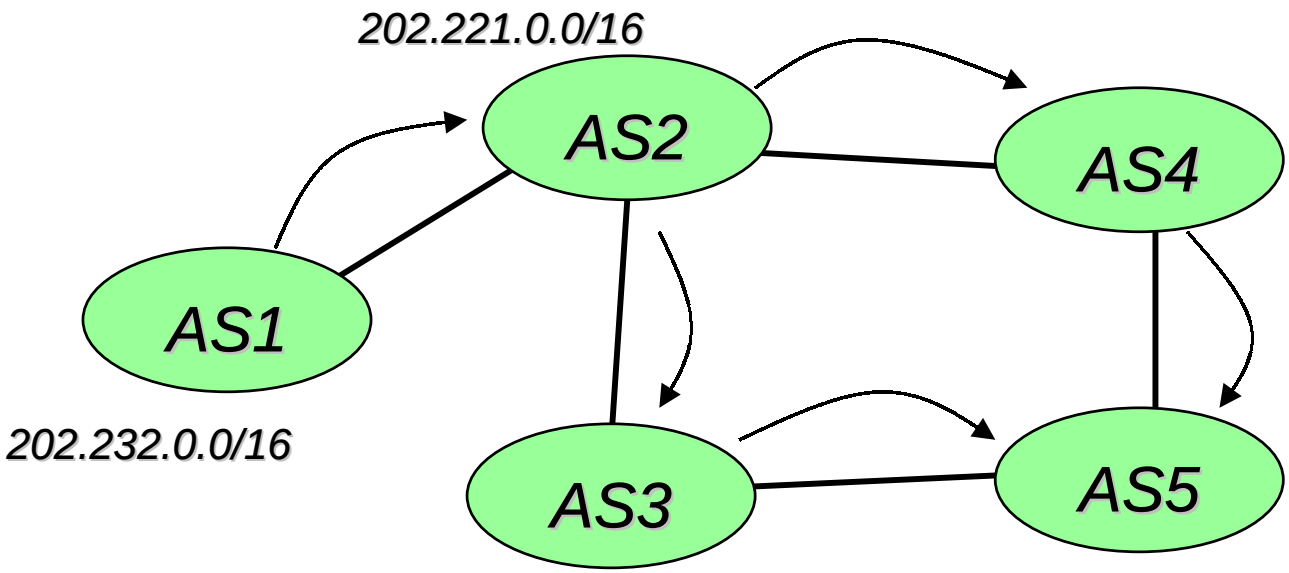
Origin属性

- その経路情報をどこから持ってきたかを表す
- 最初にBGPでアナウンスする時に設定される
- 可能な値
  - IGP
  - EGP
  - Incomplete
- 必須属性

AS Path属性

- 経路情報が伝播する際に経由したASの列/組
- ループの検出
- 一般的にはAS Pathの長さが短いほうが選ばれる
  - ポリシーによる
  - prepend, stuffing等の技巧
- 必須属性

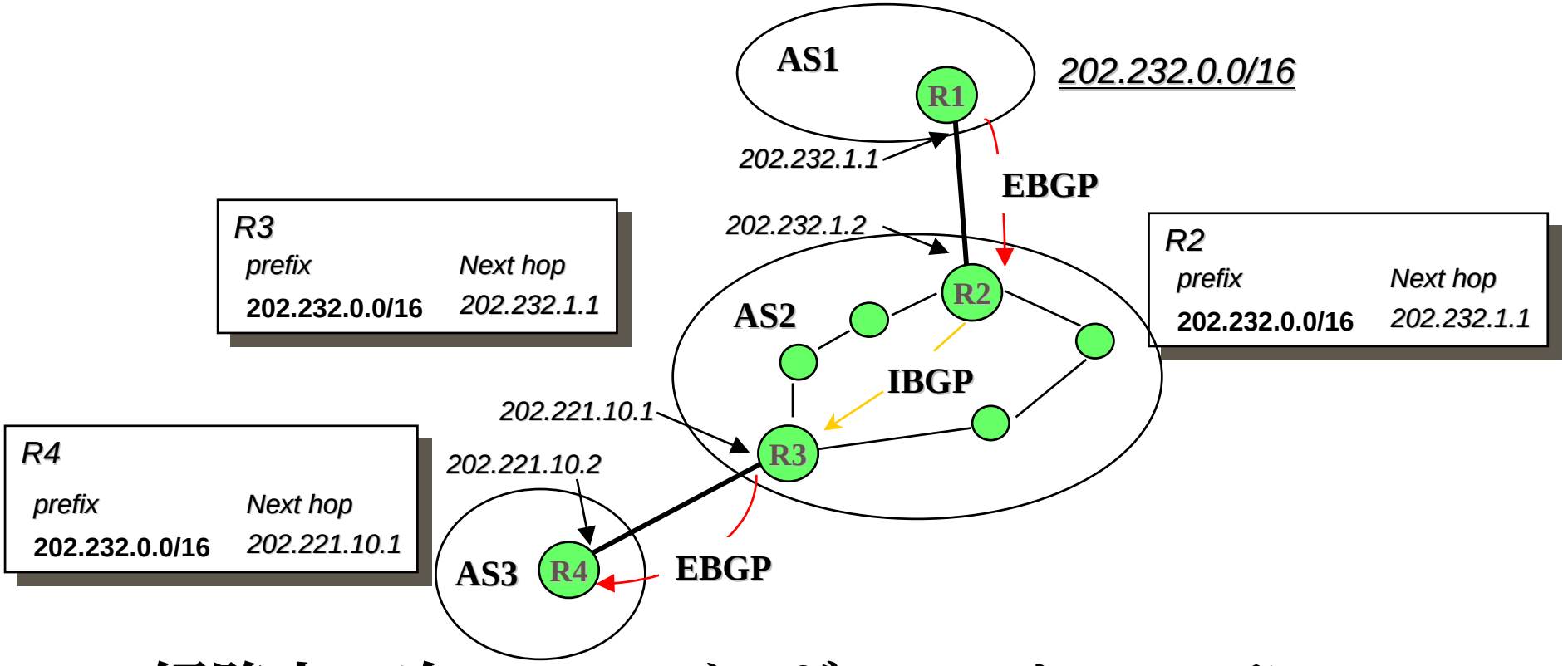
AS Path属性の例



202.221.0.0/16	3 2
202.221.0.0/16	4 2
202.232.0.0/16	3 2 1
202.232.0.0/16	4 2 1

- AS1が202.232.0.0/16を、AS2が202.221.0.0/16をアナウンス

Next Hop属性

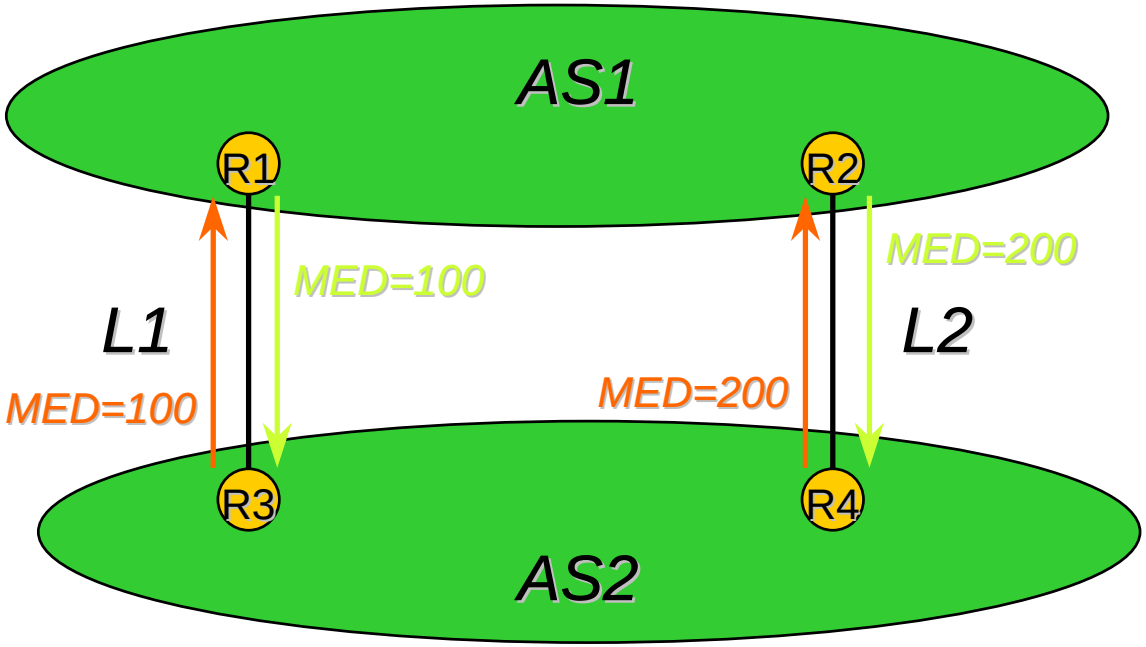


- 経路上の次のASのボーダールータのIPアドレス
- IBGPで伝播するときには値は変わらない
- R3からR1への経路はIGPで解決

Multi-Exit Discriminator(MED)

- 同一隣接ASからの複数経路を区別する
- 値が小さいほうを優先
  - IGPのコストを反映させるも可
  - ロードバランスを考えて設定するも可
- 非通過型属性

MEDの例

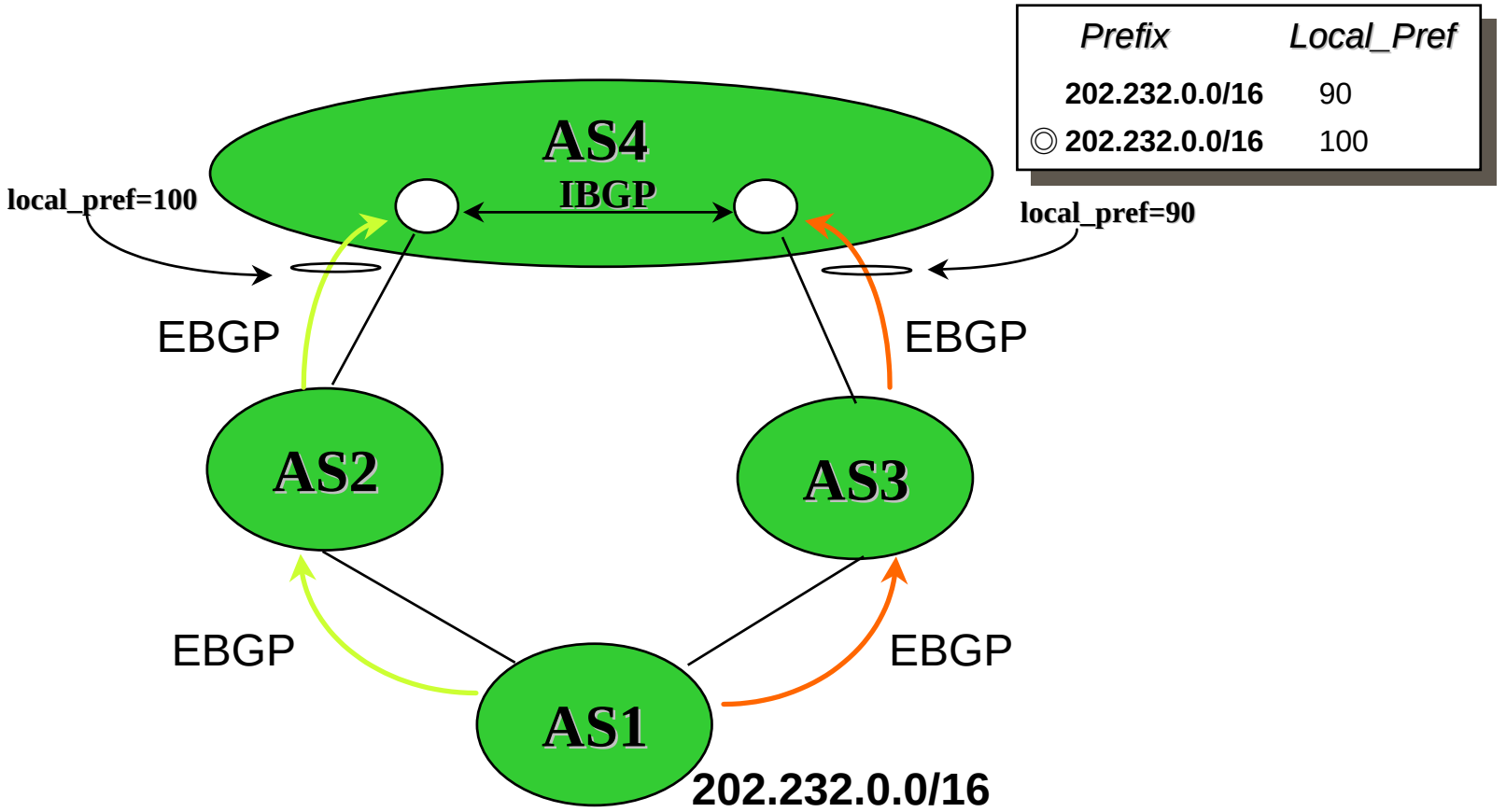


- L1を優先
- AS1とAS2で各リンクのMEDは独立に設定

Local Preference属性

- 同一AS内部で複数経路の優先度を表すために用いられる
- 値が大きいほど優先される
- 非通過型属性

Local Preferenceの例



- AS4ではAS2 AS1という経路が選択される

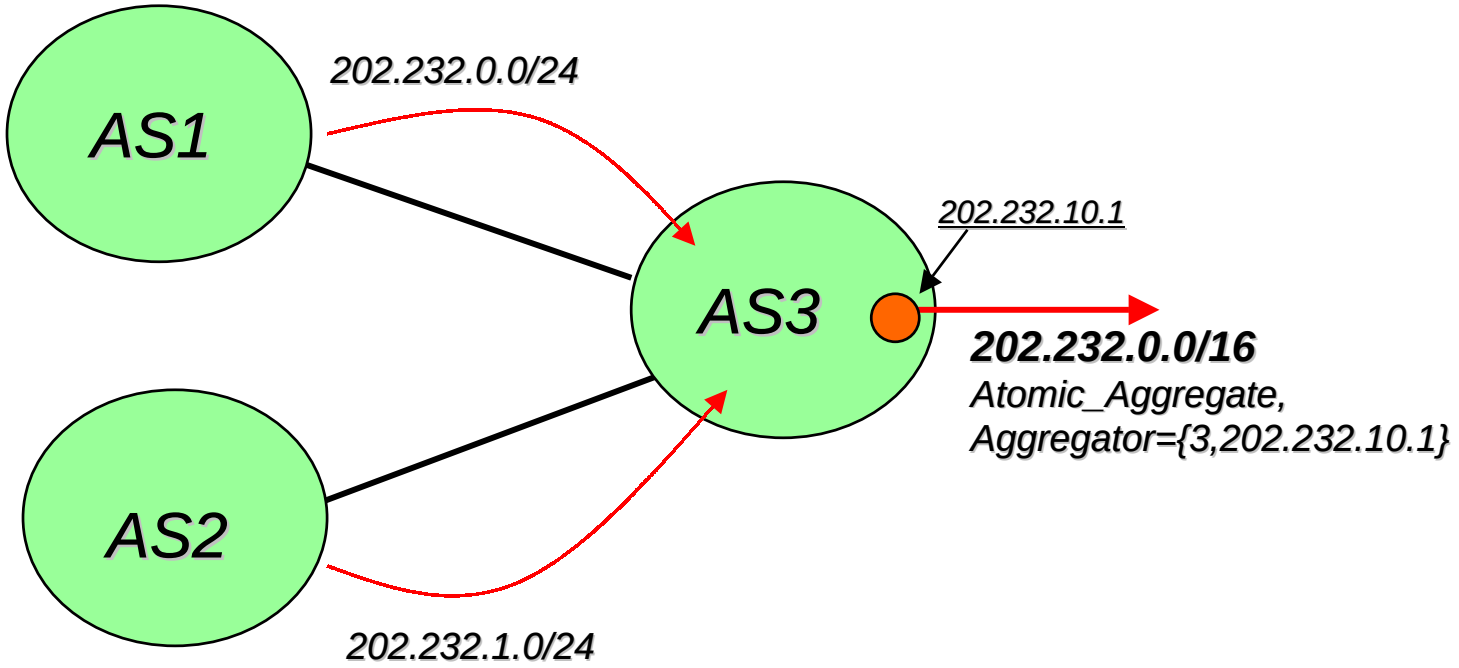
Atomic Aggregate属性

- 経路の集成(Aggregate)を行ったときに付加される属性
- 集成の際に細かい経路に付加されていた情報が欠落したことを示す
- 再び細かい経路に分けることはできない

Aggregator属性

- 経路の集成を行ったBGPスピーカのIPアドレスと、それが属するAS番号を示す属性

経路の集成



- Atomic Aggregate属性とAggregator属性が設定される

Community属性

- RFC1997
- 経路に色をつける
  - ポリシーに応じて経路をグループ分けする
  - 一つの経路が複数のグループに属することも可
- 32ビットの整数値

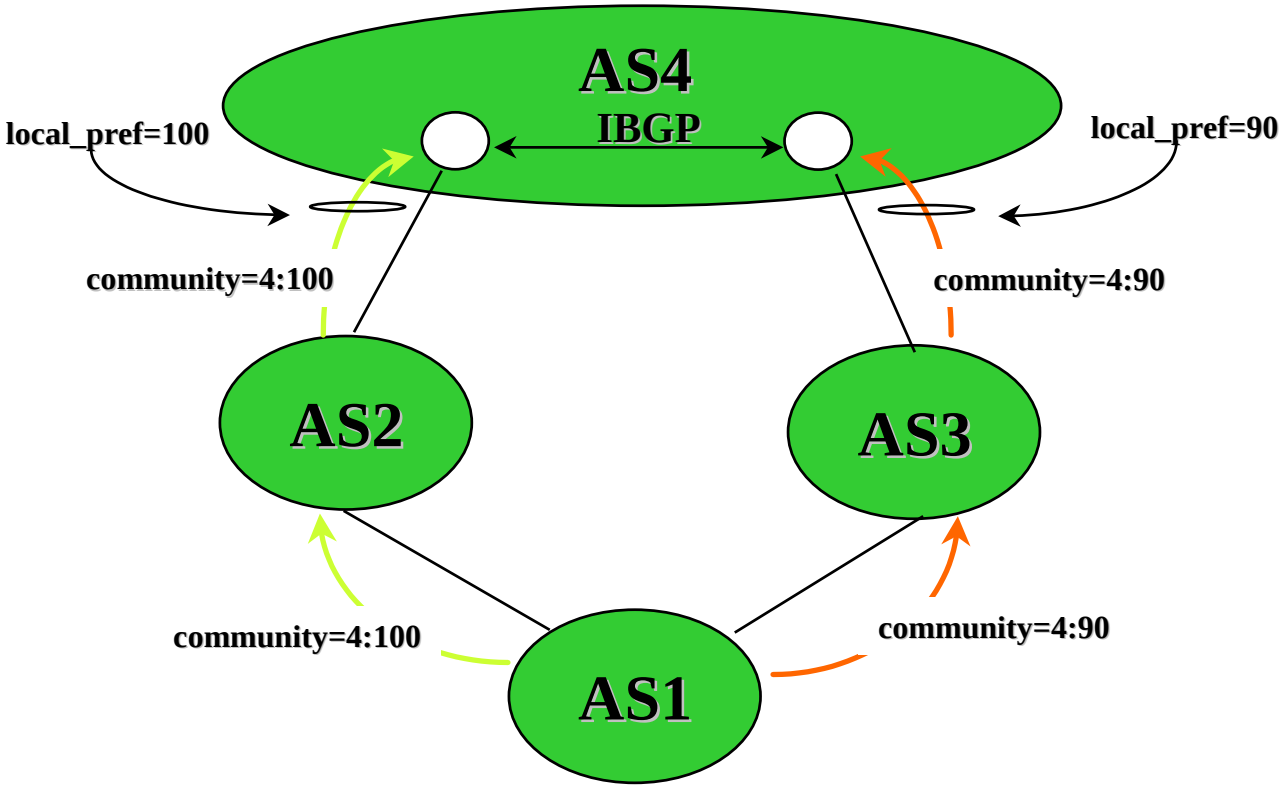
Community属性の値(共通)

- 予約領域
  - 0x00000000 - 0x0000FFFF
  - 0xFFFF0000 - 0xFFFFFFFF
- Well-Known Community
  - NO_EXPORT(0xFFFFFF01)
    - AS外部に出さない
  - NO_ADVERTISE(0xFFFFF02)
    - 他のルータに出さない
  - NO_EXPORT_SUBCONFED(0xFFFFFF03)
    - 同盟中の他のメンバーASに出さない

Communityの値(ユーザ定義)

- 予約されていない値は、AS毎に独自のCommunityを定義できる
  - 上位16ビット: Communityを定義したAS番号
  - 下位16ビット: そのAS内部で用いるCommunity番号
  - 表記法: AS番号:Community番号

Community属性の例1



- RFC1998
- **離れたASでの経路選択を制御**

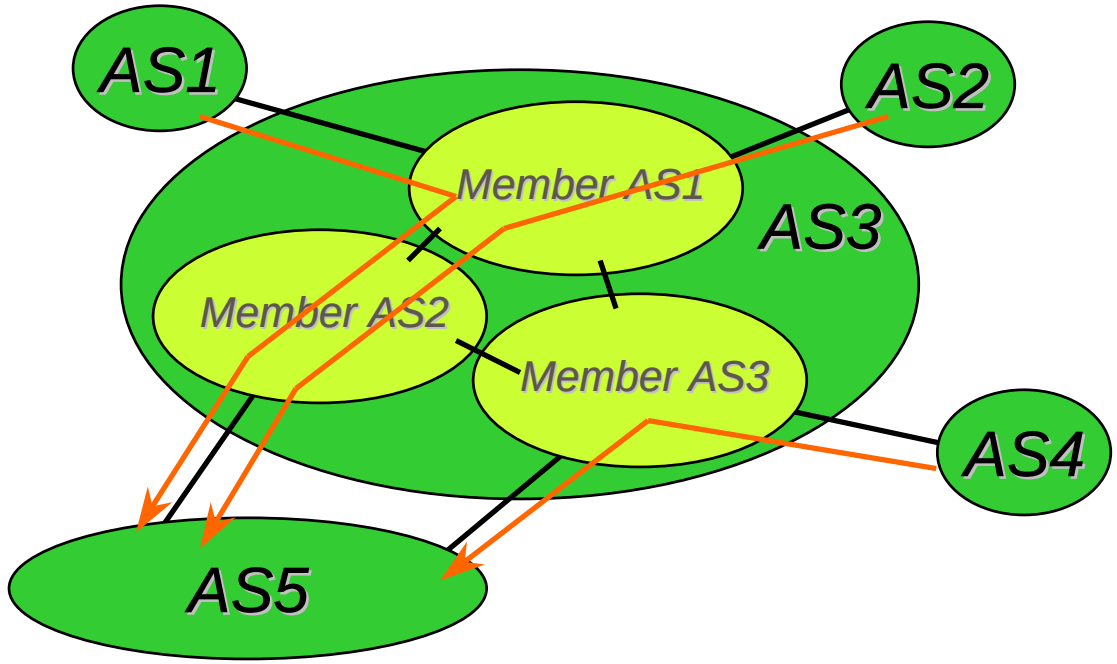
Community属性の例2

- AS内部での経路のグループ分け
  - 例
    - 外部への経路アナウンスのポリシーに応じてCommunityを定義する
      - 2497:10　顧客の経路
      - 2497:20　peerの経路
    - 個別の経路情報ではなくCommunityの値のみに着目してアナウンスの仕方を決められる

AS同盟(Confederation)

- RFC1965
- AS内部を、サブASに分割
  - サブAS間の階層関係、包含関係は無い
  - サブASではAS番号にプライベートAS (64512-65535)を用いる
- 外部からは一つのASに見える
- サブAS間は、IBGPに近いEBGP
  - サブAS間で経路を渡すときには、Next Hop, MED, Local Preference等の値は保存される
- 大きなASで、IBGPのメッシュを減らすのに役立つ

AS同盟(Confederation)

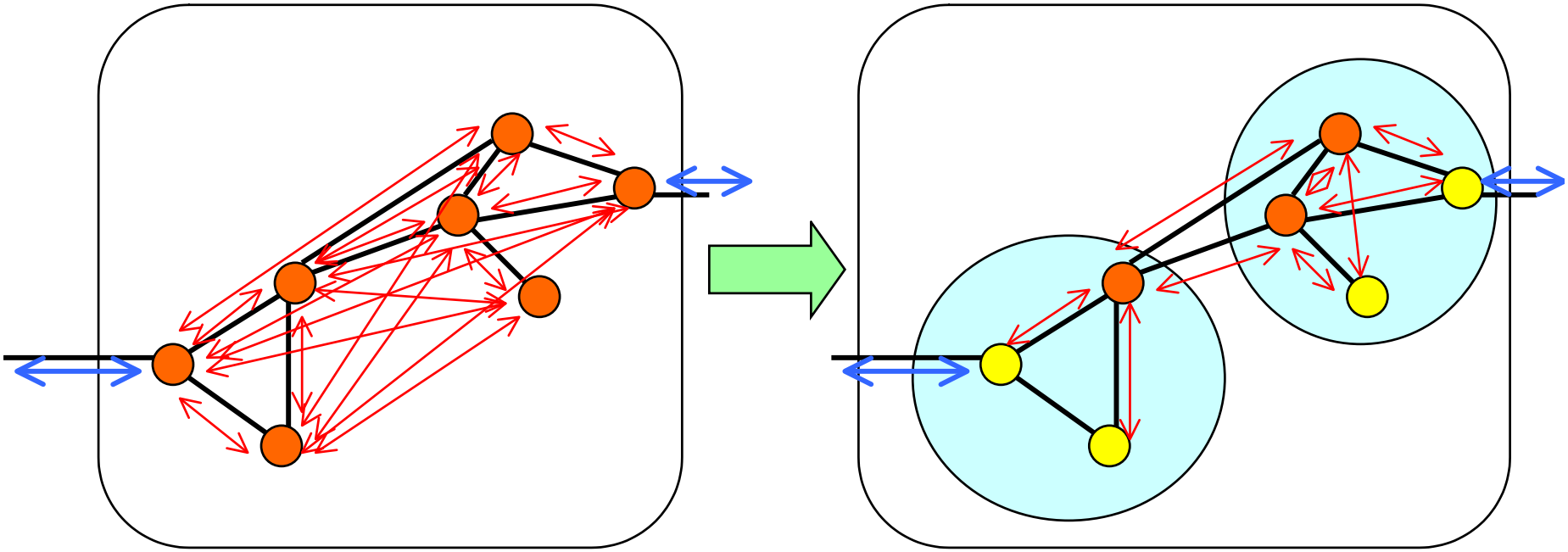


- 単一AS内部での細かいポリシーの違いを実装
- IGPのみを用いて実現するのは難しい

Route Reflector

- IBGPフルメッシュの問題
  - N x NのIBGPセッション
- AS内部で用いるルートサーバ的イメージ
- BGPスピーカをグループ(クラスタ)に分ける
  - リフレクタ
    - AS内の他のクラスタのリフレクタと経路情報を交換
    - クラスタ内のBGPスピーカに経路情報を供給
  - クライアント
    - リフレクタからBGPの経路情報をもらう

Route Reflectorの例

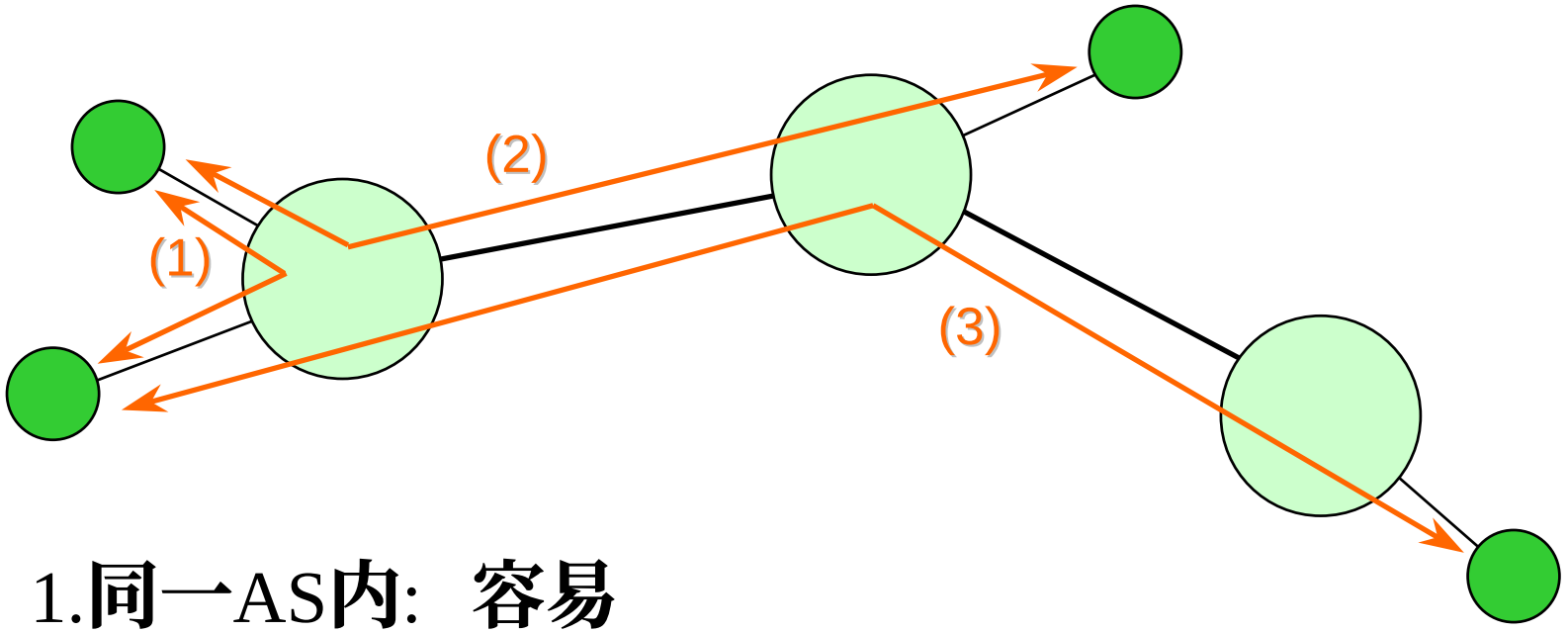


ポリシールーティング

ポリシールーティングとは？

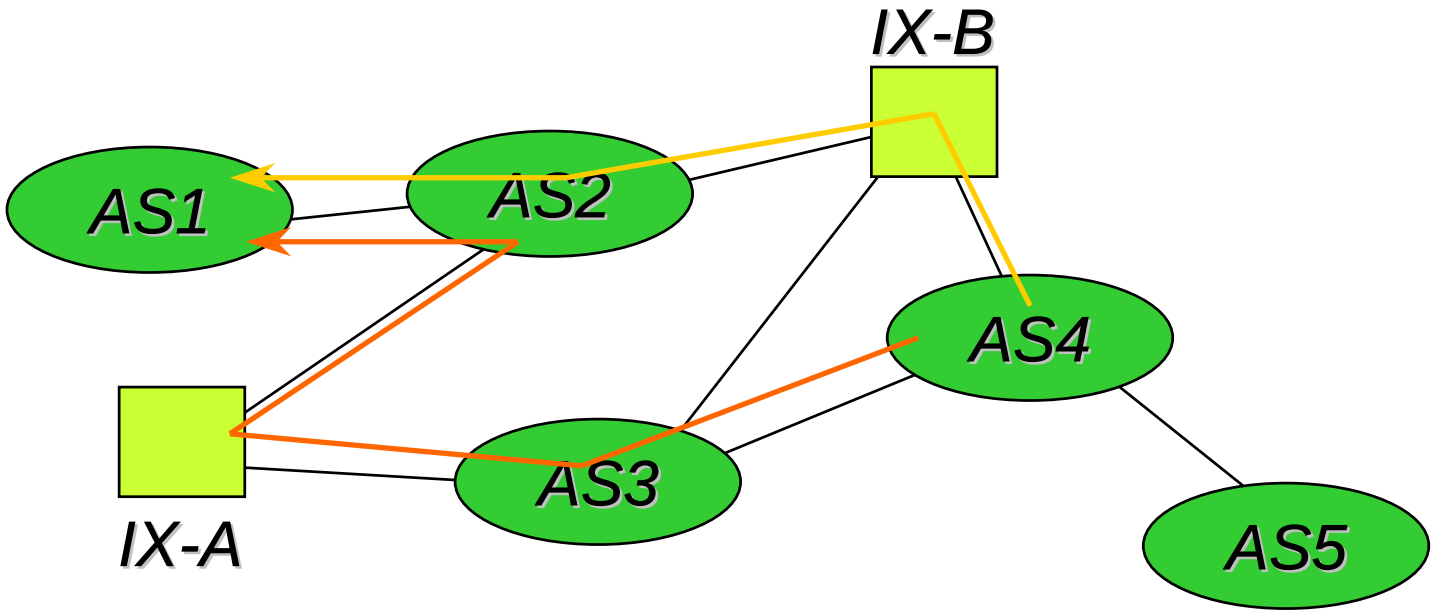
- ポリシーに基づく経路選択
  - 他のISP(AS)とどのようにトラフィックをやりとりしたいか
    - 単に近さやコストをもとにした選択ではない
    - 他のISPとの間でどのように経路情報をやり取りするか
    - 個々の目的地ごとに経路を選択する
    - BGPのパス属性を用いる
  - 経路情報のやり取りの制御だけでは実現できないポリシーもある
    - ネットワークトポロジの再考などが必要

AS間経路制御



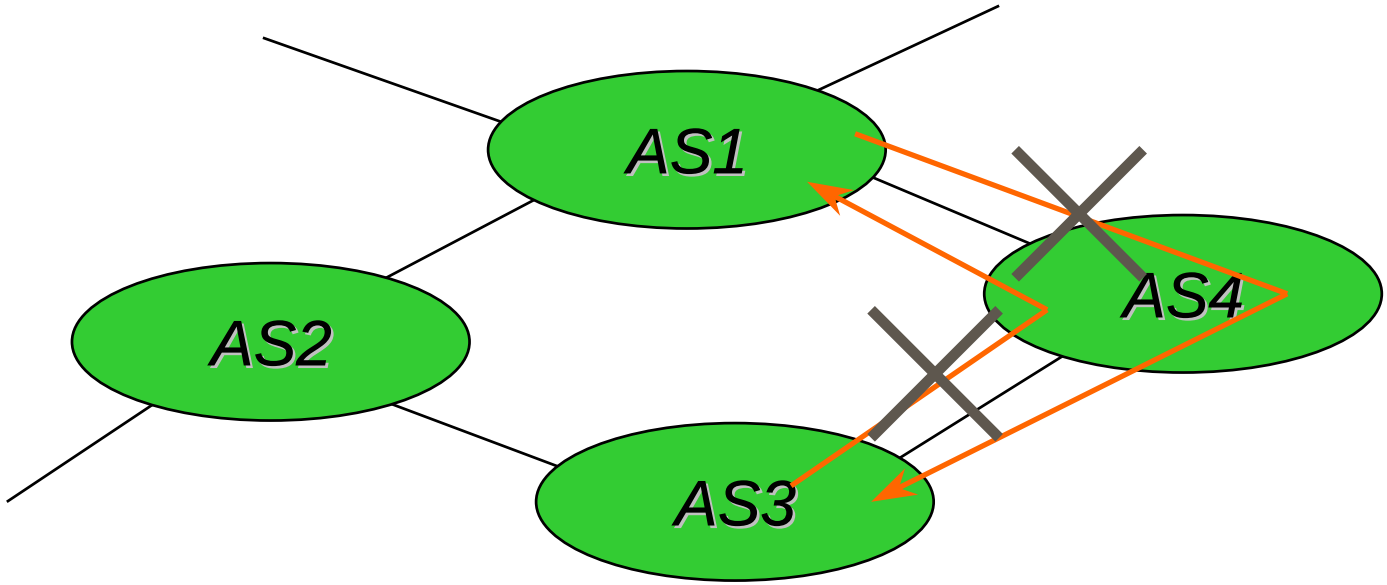
- 1.同一AS内: 容易
- 2.隣接AS間: やや難
- 3.離れたAS間: 難

AS_PATHをもとにしたポリシー



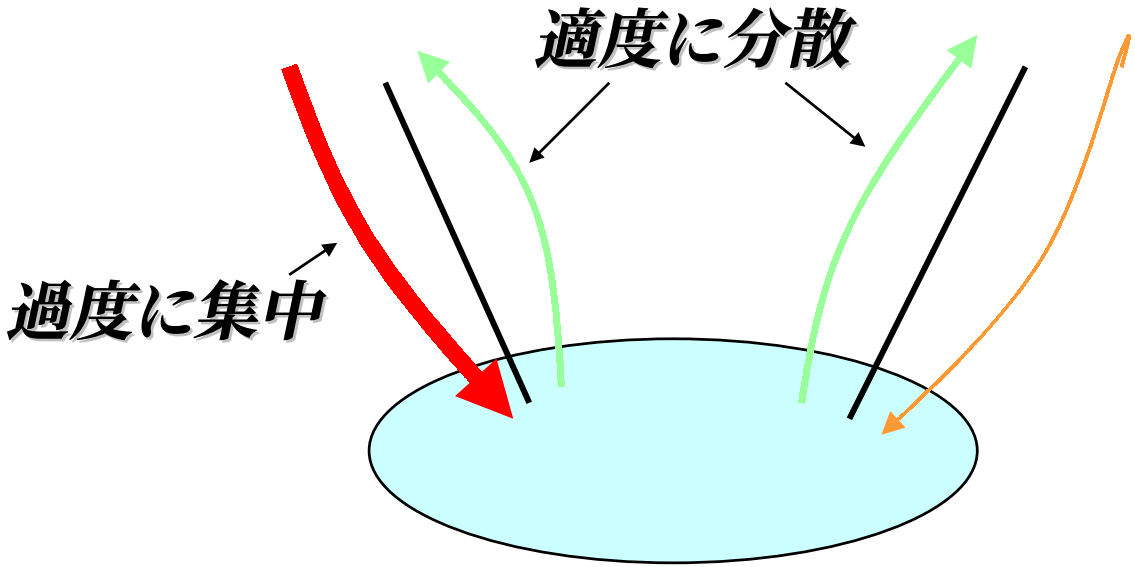
- AS4からAS1への二つの経路からの選択
  - AS2 AS1
  - AS3 AS2 AS1
- AS4は、AS_PATHパス属性などを用いて経路を選択する
- AS5はAS4と異なるポリシーをもてない

通過ポリシー



- AS4は、AS1とAS3の通信を中継したくない
- AS_PATHパス属性を用いた経路情報のフィルタリングなどにより実現

マルチホーム下でのロードバランス

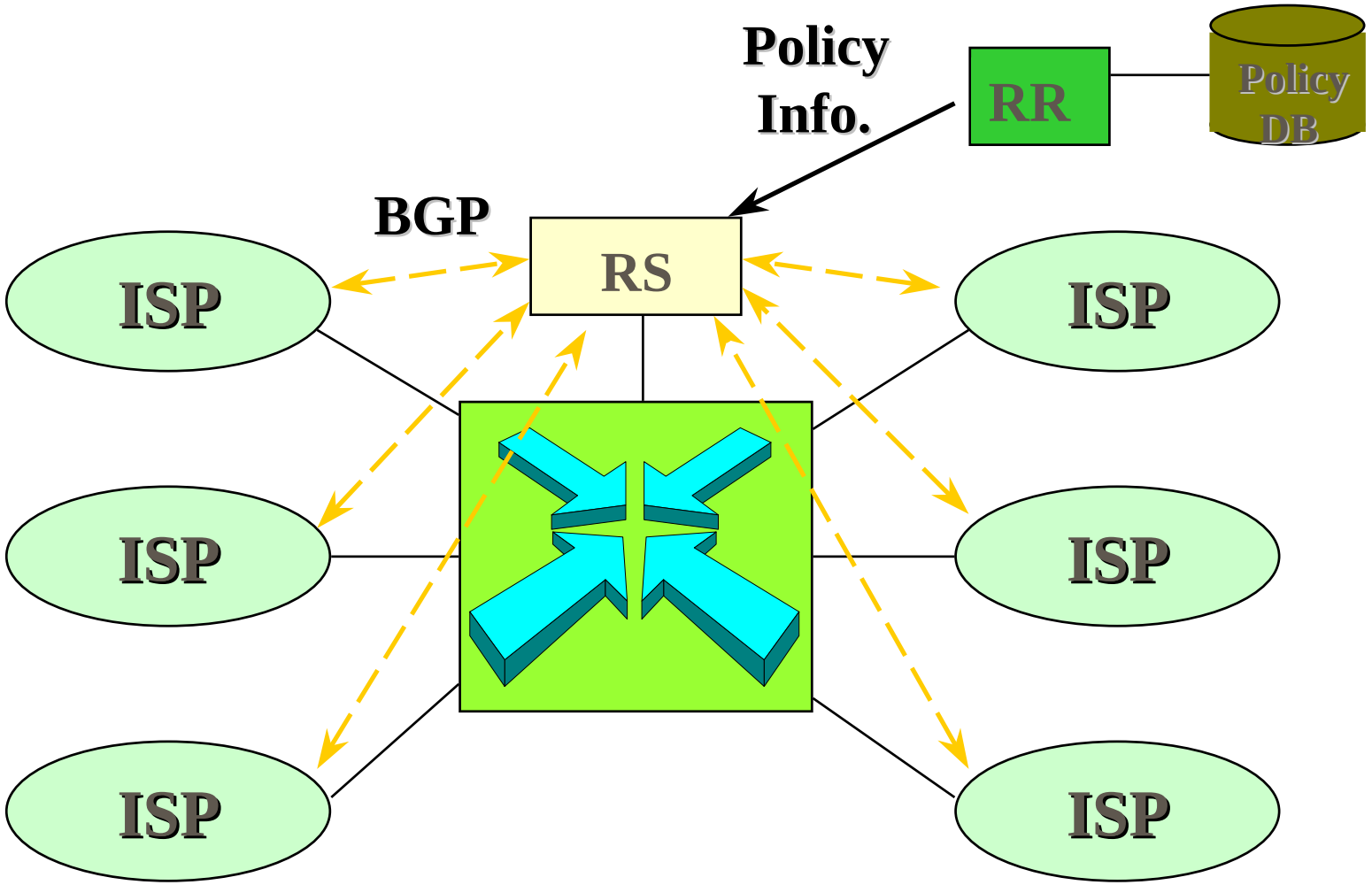


- 外に出るトラフィックは調整がしやすいが、外から入ってくるトラフィックの調整は難しい
- 手探りで調整
  - MED,AS PATH Prepend, Community等を駆使

ルーティングレジストリとルートサーバ

- パケット転送と経路選択のプロセスの分離
- ルーティングレジストリ (RR)
  - 各ASの経路制御ポリシーのデータベース
- ルートサーバ(RS)
  - 第2層エクスチェンジに接続するISPとBGPで通信する
  - RRに登録されたポリシーをもとに、各ISPのボーダルータの経路表を計算する
- いまだ発展途上

ルーティングレジストリとルートサーバ



結論

- BGP4はAS間の経路制御の標準プロトコルである
  - 経路の選択にはパス属性を用いる
  - 実現できるポリシーは限られている
  - 細かなポリシーの実現にはネットワークトポロジの再考なども必要
- BGP4を使えば良いというわけではない
  - 使わなくても良い場合もあれば、使わないほうが良い場合もある
  - 使ったがために運用管理が煩雑になることもある

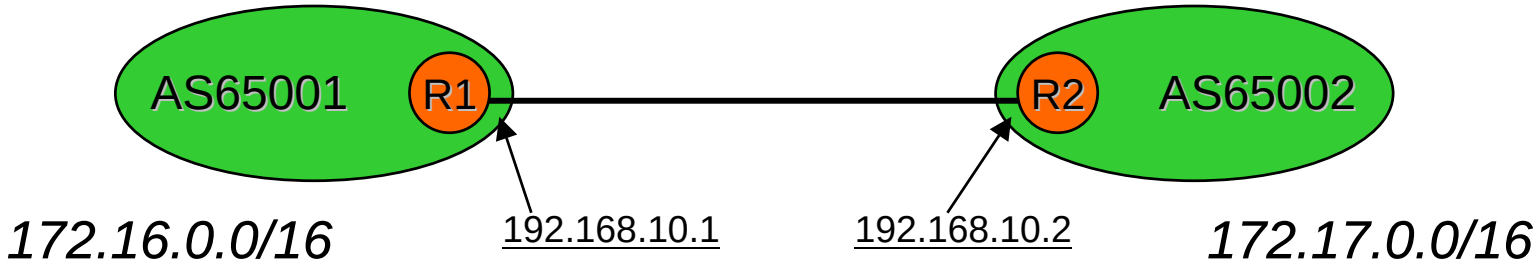
結論

- ポリシールーティングは難しい
  - 単なるマルチホームでもトラフィックをうまく複数のリンクに分散することは難しい
  - できる限り、シンプルなネットワーク構成が望ましい
- 今後も技術開発が必要
  - RR, RS等の管理技術
  - 運用技術の確立
    - route flapping等の問題

付録

サンプルコンフィギュレーション

Ciscoでの基本設定



R1での設定例

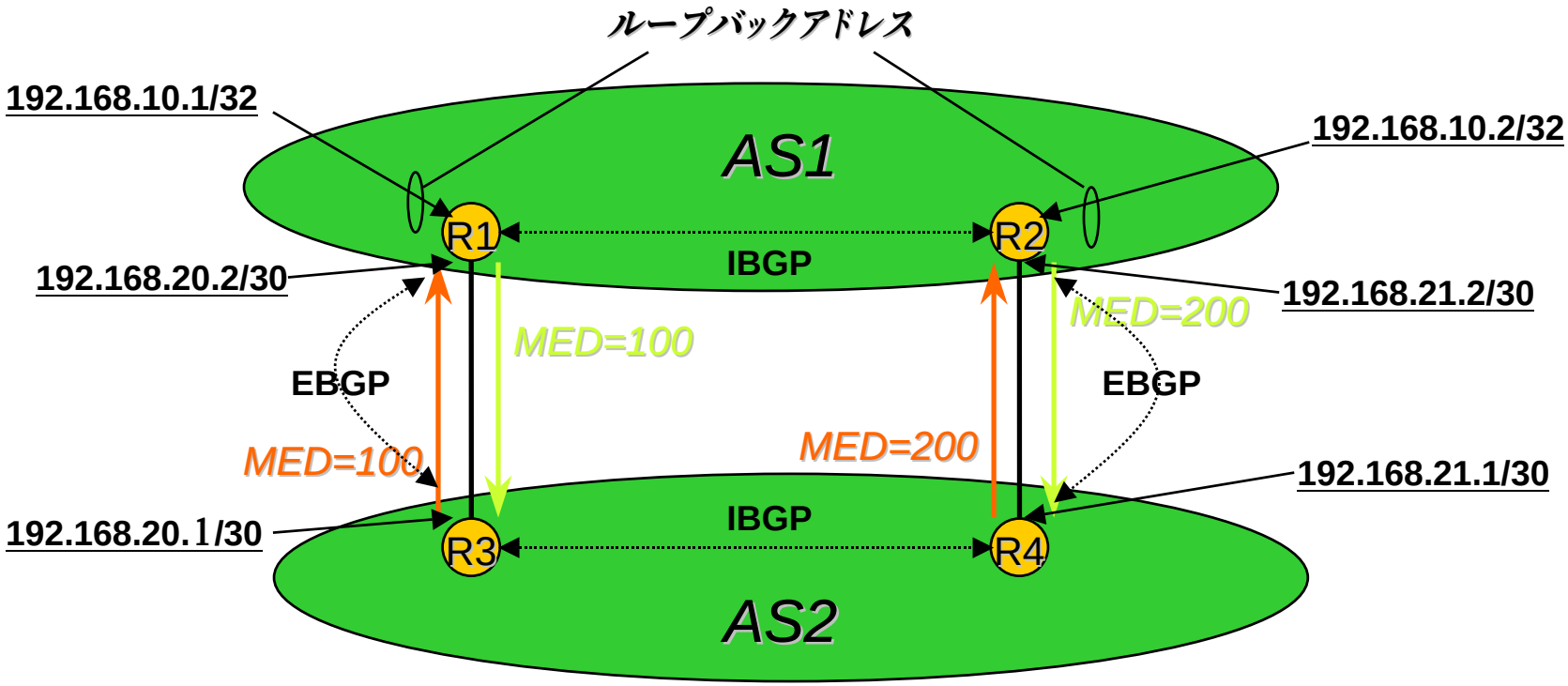
router bgp 65001
network 172.16.0.0
neighbor 192.168.10.2 remote-as 65002

R２での設定例

router bgp 65002
network 172.17.0.0
neighbor 192.168.10.1 remote-as 65001

MED

スライド28の例



MED

R1での設定例

```
interface loopback 0
ip address 192.168.10.1 255.255.255.255

router bgp 1
no synchronization
neighbor 192.168.10.2 remote-as 1
neighbor 192.168.10.2 update-source loopback0
neighbor 192.168.20.1 remote-as 2
neighbor 192.168.20.1 route-map MED-OUT out

route-map MED-OUT permit 10
match as-path 10
set metric 100

ip as-path access-list 10 permit ^$
```

R2での設定例

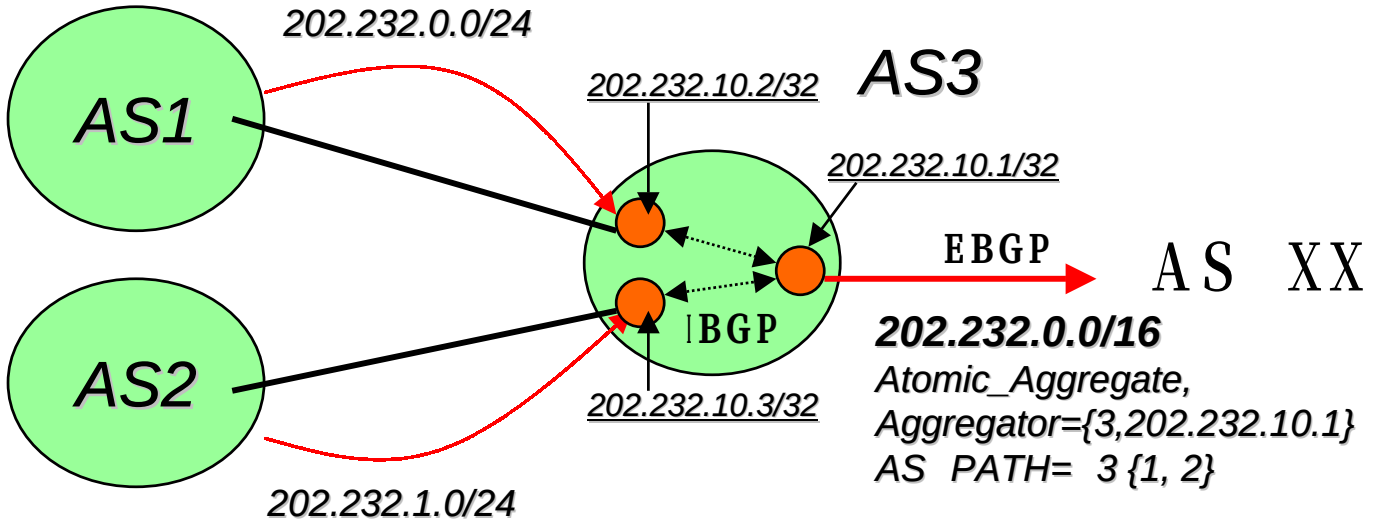
```
interface loopback 0
ip address 192.168.10.2 255.255.255.255

router bgp 1
no synchronization
neighbor 192.168.10.1 remote-as 1
neighbor 192.168.10.1 update-source loopback0
neighbor 192.168.21.1 remote-as 2
neighbor 192.168.21.1 route-map MED-OUT out

route-map MED-OUT permit 10
match as-path 10
set metric 200

ip as-path access-list 10 permit ^$
```

Aggregate

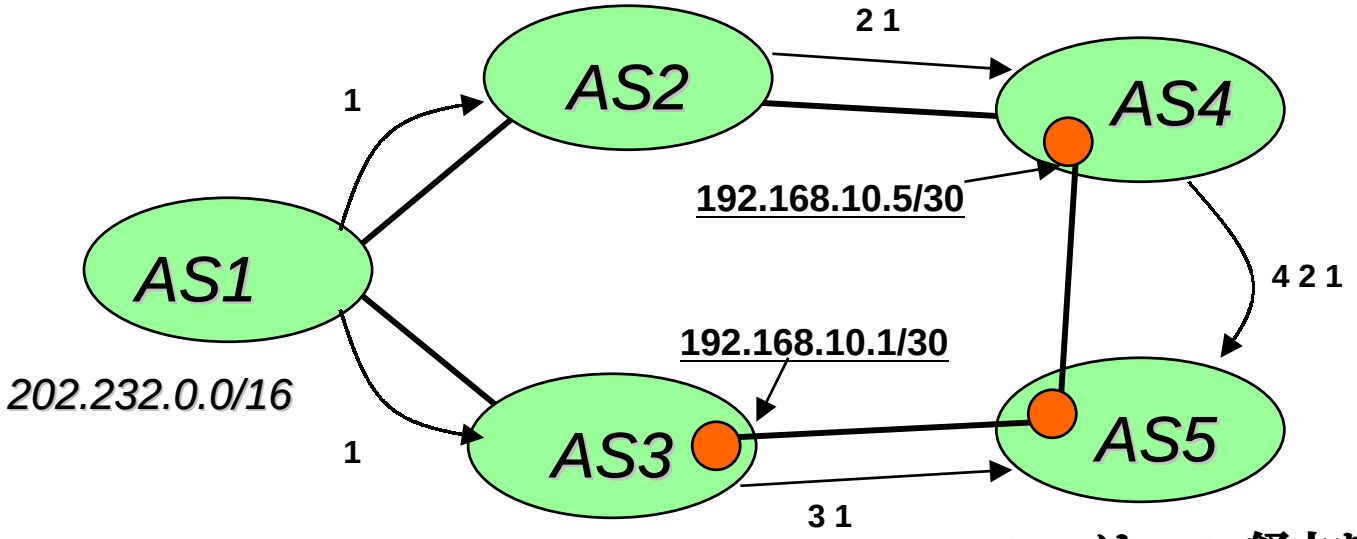


設定例

interface loopback 0
ip address 202.232.10.1 255.255.255.255

router bgp 3
no synchronization
network 202.232.10.0
aggregate-address 202.232.0.0 255.255.0.0 as-set summary-only
neighbor 202.232.10.2 remote-as 3
neighbor 202.232.10.2 update-source loopback0
neighbor 202.232.10.3 remote-as 3
neighbor 202.232.10.3 update-source loopback0
neighbor X.X.X.X remote-as XX

Local-preference



AS1へは、AS4経由を優先したい

注：ciscoのlocal-preferenceのデフォルト値は100

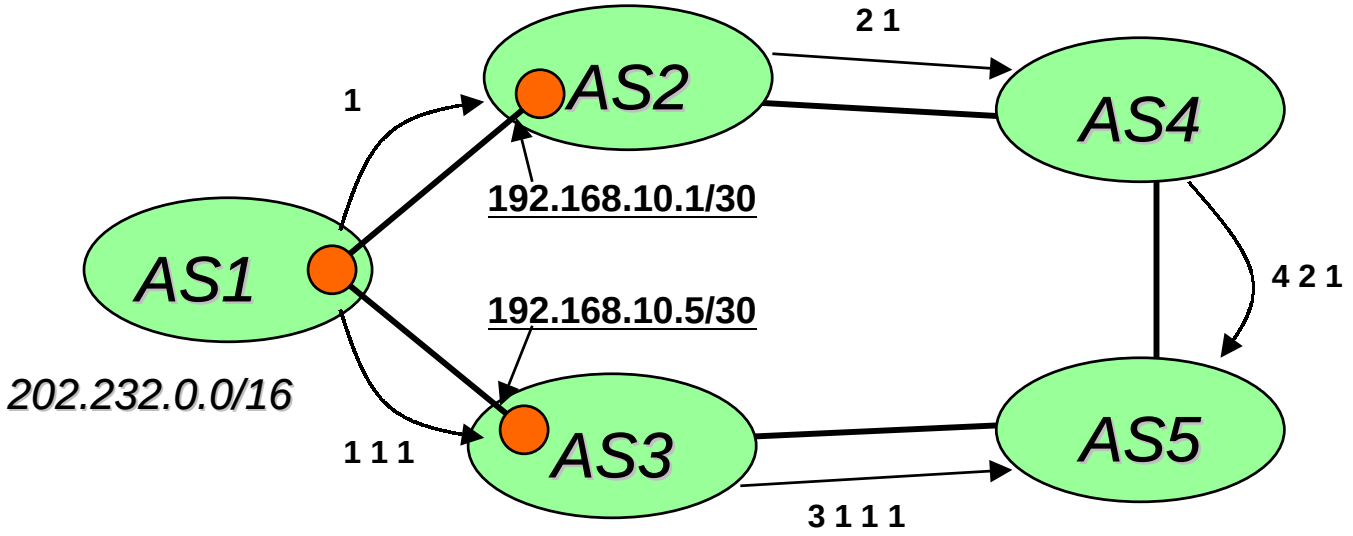
AS5のボーダルータでの設定例

router bgp 5
neighbor 192.168.10.1 remote-as 3
neighbor 192.168.10.1 fromAS3 in

ip as-path access-list 10 permit ^3 1$

route-map fromAS3 permit 10
match as-path 10
set local-preference 90

AS PREPEND

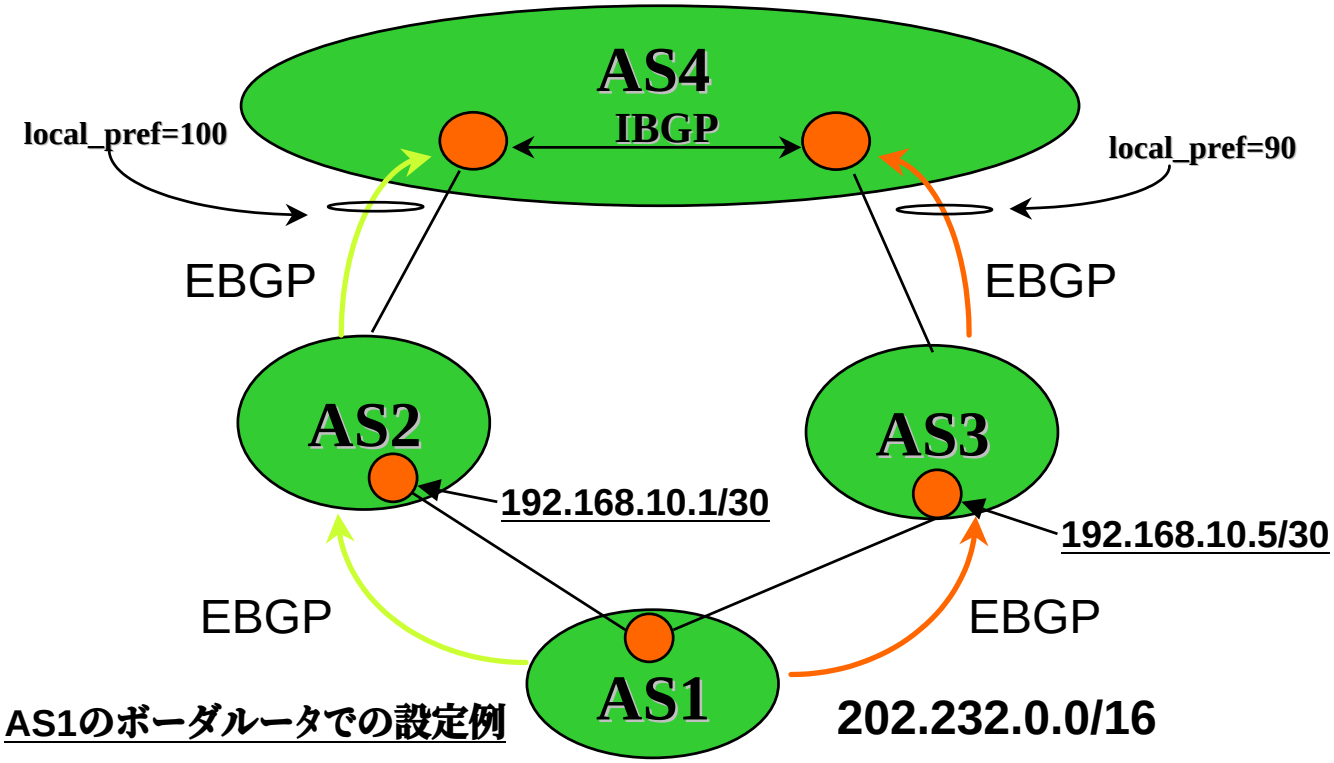


AS1のボーダルータでの設定例

router bgp 1
network 202.232.0.0 mask 255.255.0.0
neighbor 192.168.10.1 remote-as 2
neighbor 192.168.10.5 remote-as 3
neighbor 192.168.10.5 route-map PREPEND out

route-map PREPEND permit 10
set as-path prepend 1 1

Community



AS1のボーダルータでの設定例

ip bgp new-format
access-list 10 202.232.0.0 0.0.255.255

router bgp 1
neighbor 192.168.10.1 remote-as 2
neighbor 192.168.10.1 send-community
neighbor 192.168.10.1 route-map toAS2 out
neighbor 192.168.10.5 remote-as 3
neighbor 192.168.10.5 send-community
neighbor 192.168.10.5 route-map toAS3 out

route-map toAS2 permit 10
match ip address 10
set community 4:100

route-map toAS3 permit 10
match ip address 10
set community 4:90