BGPのシステム設計論

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

概要

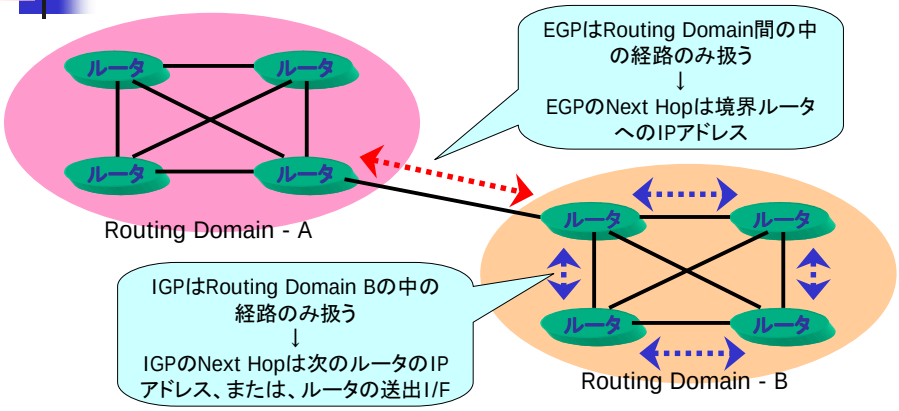
- 関連事項の整理
- BGP・プロトコル概説
- ISPネットワーク拡大に沿った規模対応
- ポリシルーティング
- ポリシルーティングの実際

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

関連事項の整理

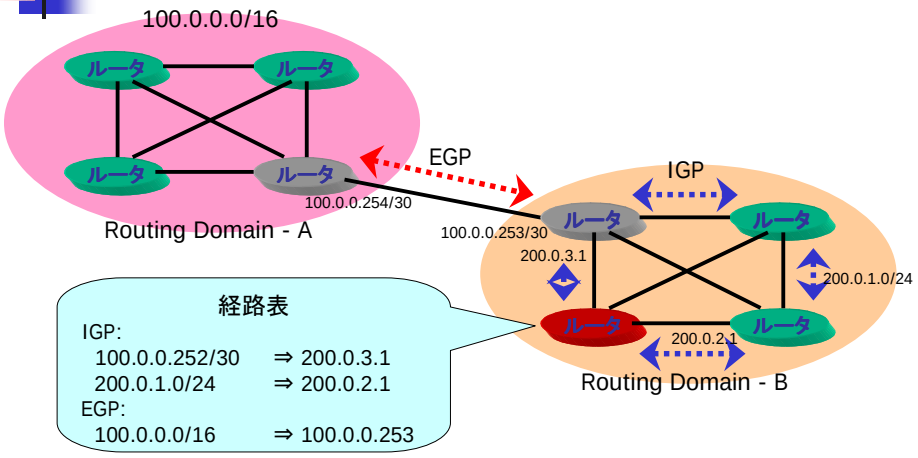
Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

IGPとEGPの違い



Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

IGPとEGPでの経路解決



Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

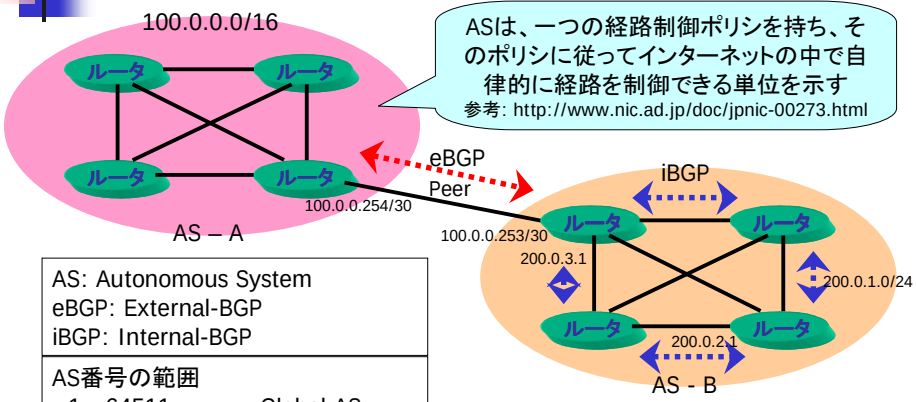
BGP – Border Gateway Protocol

- EGPとして標準であるBGP
  - パスベクタ型(Path Vector)
  - RFC1771
- バージョン
  - BGP-4が標準
  - IPv6はBGP-4を拡張して利用(BGP4+)
- 特徴
  - 様々な経路制御パラメータがある
    - MED, Local Preference, AS Path Length(*), Community…
  - マルチホーム、冗長構成が可能
  - 高いスケーラビリティ
  - 高い拡張性 (Capabilityの利用)

(*)RFC1771にはAS Path長によって最適経路を判断する基準は記述されていない。しかし、現在は一般的に行われている。

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

BGP - Terminology



AS: Autonomous System
eBGP: External-BGP
iBGP: Internal-BGP

AS番号の範囲
1～64511 : Global AS
64512～65534 : Private AS
0, 65535 : IANA Reserve

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

CIDRの復習(1)

- CIDR – Classless Inter-Domain Routing
- クラスレスなAS間の経路制御
  - クラスレスとは、、
    - classA, classB, classCなどのクラスの考え方を除いたもの
    - 対義語==クラスフル(classful)

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

CIDRの復習(2)

- クラスフル(classful) という考え方
  - IPアドレスの先頭オクテットの値でネットワークアドレスの範囲を判断する
    - class A = 1～126— 第一オクテットだけがネットワーク
    - class B = 128～191— 第二オクテットまでネットワーク
    - class C = 192～223— 第三オクテットまでネットワーク
  - ネットワークアドレス単位でしか扱わない（扱えない，伝えない，伝えるすべがない）
  - その中を更に分割したものをサブネットと言う
    - 分割する大きさも自分にしか定義できず、伝えるすべがない
    - クラスフルネットワークの中は統一したサブネットのサイズにしないと扱えない

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

CIDRの復習(3)

- クラスレス(classless)という考え方
  - どこまでがネットワークを示すのかを明示して扱う
  - ネットワークを示すものをプリフィクス(Prefix)と呼ぶ
  - プリフィクスの長さは一般的にビット数で表される
    - Class Cの 202.216.40.0 – 202.216.40/24 (202.216.40.0/24)
- つまりクラスレスだと、
  - 連続するclass Cアドレスを任意の大きさでひとかたまりで扱える
  - Class Aのサブネットも全く同様に扱える
  - Class Cより小さいアドレスブロックも全く同様に、任意の大きさで扱える
    - これがいわゆるVLSM(Variable Length Subnet Mask)

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

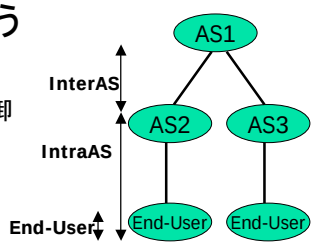
CIDRの復習(4)

- CIDR—クラスレスなAS間経路制御
  - プリフィクス＋プリフィクス長で経路情報を扱う
  - 複数のClassC(=/24)アドレスも（あらゆるアドレスが）、任意の大きさでひとかたまりに扱える
  - AS内の小さなネットワークセグメント，ユーザネットワークをひとかたまりにして他のASに広告できる
    - 経路集成—aggregation

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

The Internetにおける階層的経路制御(1)

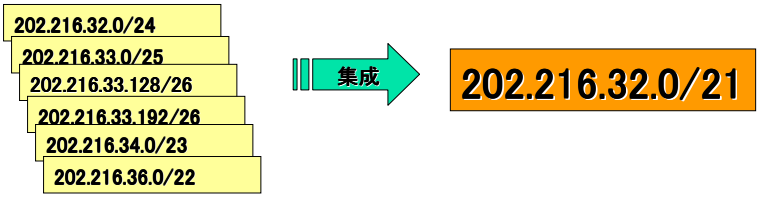
- 全インターネットを3つに階層化して、それぞれ独立して経路制御を扱う
  - InterAS
    - AS間，Default-Freeゾーン，EGPで制御
  - IntraAS
    - AS内，AS内の全経路，IGPで制御
  - End-User
    - ユーザサイト内。StaticやIGPで制御



Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

The Internetにおける階層的経路制御(2)

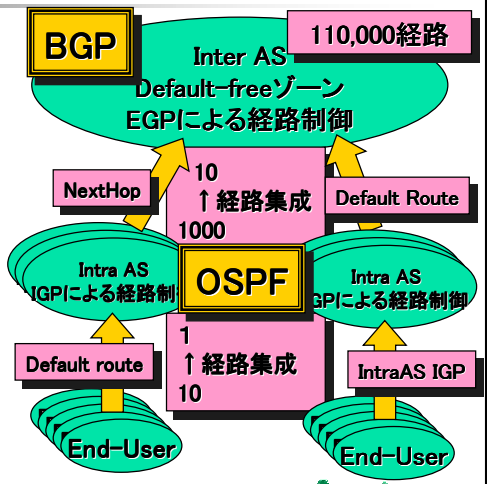
- 経路集成 – Aggregation
  - 複数の経路情報をひとかたまりにして、より大きなサイズの（より短いプリフィクスの）単一の経路情報にすること
  - 現在IPアドレスの割り振りはISP毎に行われているので、そこからユーザに割り当てるIPアドレスは割り振りブロックで集成することができる。



Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

The Internetにおける階層的経路制御(3)

- それぞれの境界で経路集成=情報量の縮退
- 上流の経路は全てdefault route で制御する
- 下流の詳細構成は気にせず、ひとかたまりの経路で制御する



Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

The Internetにおける階層的経路制御(4)

- その内在的矛盾？
  - CIDRは非階層的アドレス形態であったIPアドレスに階層構造を持ち込んだ
  - 階層構造を厳格に推し進めようとすると…
    - 電話番号のように局番固定割り当てのような構造が望ましい
  - 末端に近くなるほどマルチホームがしにくい
    - 小さいアドレスブロックでマルチホームをするのは難しい
      - **AggregationとCIDRの内在的矛盾**
  - 実際問題としては、小さいアドレスブロックでマルチホームすることも容認されつつある
    - 階層的経路制御の崩壊の兆し。。。。
      - Punching Hole Routesの出現(経路の半分は/24…何故？)

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

BGPの動作の仕組み

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

BGPメッセージの種類

- OPEN
  - BGPセッションを開始するときに発行される
- KEEPALIVE
  - BGPセッションが開通していることを確認するために利用される
- NOTIFICATION
  - エラーなどの情報を伝えるために使われる
- UPDATE
  - 実際の経路情報などを伝えるために使われる

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

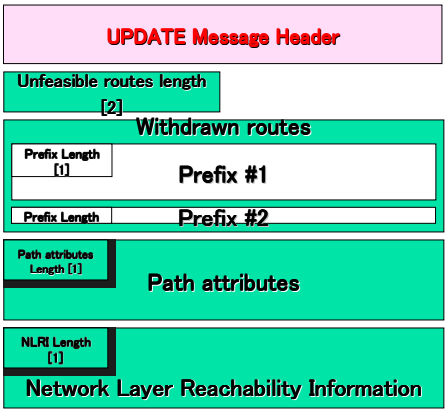
伝わってくる経路情報(IPv4)

Withdrawをする経路(経路表から削除する経路)を格納してある領域のオクテット数

Withdrawをする経路のリスト

新しくアナウンスされる経路、または既にアナウンスされている経路の更新によって新しく設定される経路に対する属性

新しくアナウンスされる経路、または既にアナウンスされている経路のリスト(到達性情報)で、これらの経路すべてに上記の属性が適用される



Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

Path Attributes（パス属性）

- プリフィクスに括りつけられた経路選択制御用の属性値群
- 必須，任意，透過性，非透過性の4つに分類
  - 必須　　– Well-known mandatory
    - 全てのBGPルータで解釈可能で、全ての経路レコードに必要
  - 任意　　– Well-known discretionary
    - 全てのBGPルータで解釈可能で、必ずしもつけなくても良い
  - 透過性　– Optional transitive
    - 一部のBGPルータで解釈されない可能性があり、次のASへも伝播される
  - 非透過性 – Optional non-transitive
    - 一部のBGPルータで解釈されない可能性があり、次のASへ伝播されない

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

崩れ始めたMandatory属性 ～こぼれ話～

- BGPのMandatoryは、BGPで経路を伝播するために必須のものとして定義された属性である。以降に説明するがIPv4の経路だけを想定すればMandatory属性は必須であるのはすごく自然である。
- しかし、BGP4+(俗称であるが..)の登場、BGPを利用したMPLSラベル配送技術の登場とBGPを取り巻く環境はBGPを経路制御プロトコルから、ネットワーク情報伝達プロトコルへと変えていった。
- この流れの中で、既にMandatoryなのにも関わらず既に現実的に省略されている属性も出始めている。

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

79

Path Attributes（パス属性）
Well-known mandatory

- ORIGIN
  - 生成元のASでどういう形でBGP上に生成されたか
    - IGP, EGP, INCOMPLETE の3値
      - 経路のほとんどはIGPがOriginとなる。
      - EGPは他のEGPからBGPに移された経路をいい、現在ほとんど見ることはない。
      - INCOMPLETEは、どこからredistributeされてBGPに移されたわからないけ路をさす。
- AS_PATH
  - 生成元ASまでの経過ASのリスト
- NEXT_HOP
  - そのプリフィクスへの次のホップとなるIPアドレス
    - 一般的に隣接するルータではなく、ASの出口のIPアドレスとなる。

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

80

Path Attributes（パス属性）
ポリシ制御のプレイヤーたち

- LOCAL_PREF – Well-Known Discretionary
  - Local Preference
  - AS内で他ASから受け取った経路に関する優先度をつけるのに用いる
- MULTI_EXIT_DISC – Non-Transitive
  - Multi Exit Discriminator
  - 複数相互接続点を持つ隣接ASに対してそれぞれの優先度を伝える
- COMMUNITY – Transitive
  - 任意の32ビットの情報を伝達する

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

eBGPとiBGP

- eBGP – External BGP
  - 他のASとの間でセッションを張り経路情報の交換を行う
- iBGP – Internal BGP
  - 同じASの複数のBGPルータの間で、それぞれがeBGPを介して入手した（あるいは自AS内から生成した）外部経路を交換し、AS内の経路情報の同期を取る
  - 基本的には、iBGPで入手した経路情報はiBGPで以遠伝播しない
    - 全てのBGPルータとiBGPセッションを確立する必要がある（回避方法は後ほど）

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

ISPネットワーク拡大に沿った規模対応設計

～BGPの導入～

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

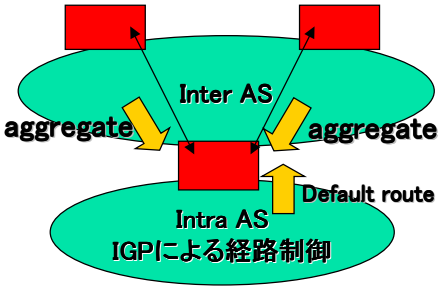
AS番号はどうやって割り当てを受けるのか

- JPNICが割り当てを行う
  - http://www.nic.ad.jp/ja/ip/asnumber.html
- AS割り当ての条件
  - RFC1930
    - 日本語訳も一応ある
    - ftp://ftp.nic.ad.jp/jpnic/ipaddress/rfc1930-jp.txt
    - あくまでガイドラインであって、実際の細かい条件はRIRs(APNICなど)によって決定され運用されている。
  - マルチホームの条件はなくなりました。

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

最も単純なBGPの導入

- IGPでデフォルトルートが指されるルータが単一のボーダルータ
- BGP→AS→独自の経路制御ポリシだから、2つ以上のASに接続



問題点：single point of failure
複数箇所で他のASと接続したい

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

BGP導入の実際

- 2つ以上の国内大手ISPを上流としてマルチホーム接続
- NSPIXP, JPIX, JPNAPなどのインターネットエクスチェンジに加入して、国内到達性を確保。別途国際ゲートウェイISP（あるいは国内大手ISP）に加入して海外到達性を確保
  - アドレスブロックは、JPNICなどから割り当てをうける
  - 現在では、IX経路で国際トランジットをもらうケースもある。

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

BGPの基本的コンフィグレーション(1)

```
router bgp 20000
 no synchronization
 no auto-summary
 network 172.16.0.0 mask 255.255.0.0
 network 192.0.1.0 mask 255.255.255.0
 aggregate-address 223.224.0.0 255.255.0.0 summary-only
 neighbor 202.249.2.60 remote-as 4689
 neighbor 202.249.2.60 route-map AS4689in in
 neighbor 202.249.2.60 route-map ixp-out out
```

BGP起動

BGPグローバルコマンド

IGPで経路があればBGPで広告

含まれるプリフィクスがIGPにあれば集成経路を広告

集成経路以外を抑制

Peer確立

Route-mapでポリシを記述

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

BGPの基本的コンフィグレーション(2)

- Inbound方向のルートマップの例

```
route-map AS4689in permit 10
 match as-path 10
 set local-preference 110
!
route-map AS4689in permit 20
 match as-path 20
 set local-preference 100
!
```

シーケンス番号

シーケンス番号

シーケンス番号順に構成され、その順番に評価される

それぞれのシーケンスでは適合条件とアクションを定義する

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

BGPの基本的コンフィグレーション(3)

- Outbound方向のルートマップの例

```
route-map ixp-out permit 10
 match as-path 30
 set metric 1000
!
```

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

ISPネットワーク拡大に沿った規模対応設計

～iBGPシステムの構築～

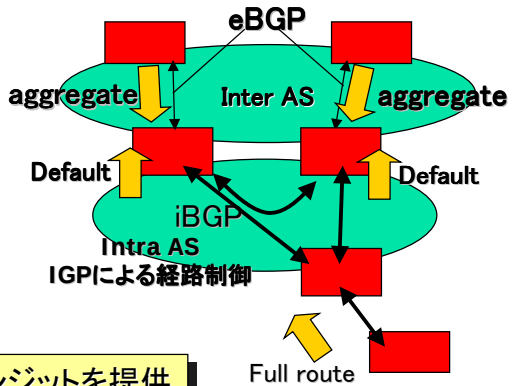
Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

2つのボーダルータを置く

- デフォルトが2つ
  - IGP的に近いほうを選択する
- ボーダルータ間の経路情報の同期？

↓

iBGPの確立



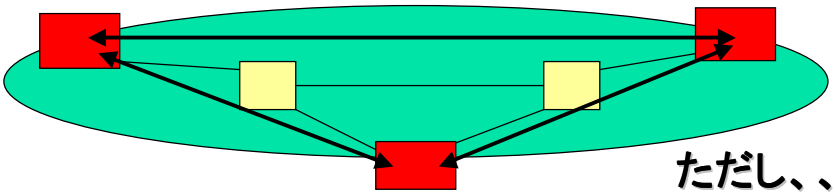
次の課題：BGP加入者にトランジットを提供

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

91

iBGPの注意点

- eBGPは直接隣接を必要とするが、iBGPはAS内での同期が目的なので離れていても確立可能
- iBGPは全てのボーダルータとセッションを張る必要がある
  - ボーダルータでなくてはならないという制限はない

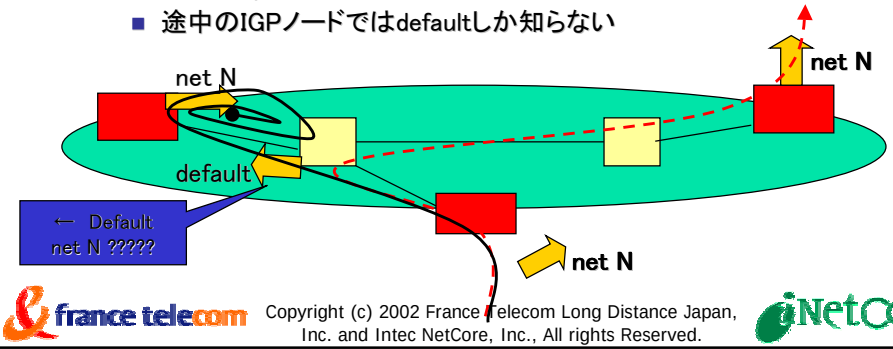


Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

92

iBGP・仕様上の問題点

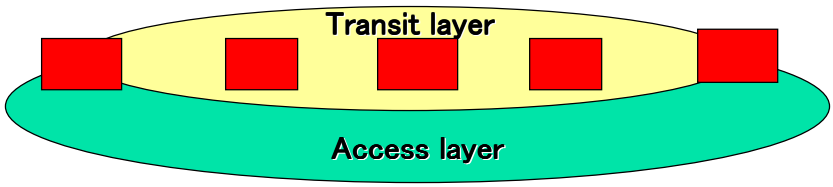
- Synchronization問題
  - トランジットしようとする経路はIGPで観測されていなければならない
  - Next-hopが別のボーダルータだった場合
  - 途中のIGPノードではdefaultしか知らない



Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

iBGPシステムの解

- No synchronization
  - IGP synchronizationの縛りを解くコマンド(c社)
  - IGPで経路観測されない経路も利用可能
    - つまり、BGPルータ間に非BGPルータがあると矛盾が発生
- トランジット層の総BGPノード化
  - トランジット層とアクセス層の二層構造へ
  - BGPユーザが多い場合、「総トランジット層」に近づく



Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

iBGP問題のまとめ

- iBGPは隣接していなくても確立可能
- 仕様では、中間ノードが経路制御できないと問題があるので、IGPでBGP経路を知っている必要があった
- がしかし、それでは経路制御階層化の意味がないので、IGPとの同期を外すほうがよい
- IGP同期を外す結果、全てのBGPルータは隣接する必要がある
- BGPルータ(トランジット)層と非BGPルータ(アクセス)層の二層に階層化
- 総トランジット層へ

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

iBGPシステムの基本(1) NEXT_HOPをIGPで観測する

- iBGPで伝播される外部経路では、基本的にNEXT_HOPの値は変わらない
  - eBGPの隣接ルータのIPアドレス
- BGP経路は、NEXT_HOPがIGPでreachableでなければ有効とならない。 そこで、、
  - IXやプライベートピアリングのセグメントをIGPで認識させる
    - 例えばpassive-interfaceでOSPFプロセスに定義する
  - eBGPルータで、iBGPピアに対してnexthop-selfを設定して、自分のIPアドレスをNEXT_HOPとして使う

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

iBGPシステムの基本(2) loopbackをピア設定に利用する

- iBGPピアの設定では、Loopbackアドレスを利用するのが「基本」
  - Loopbackインターフェースはダウンしない
    - 隣接ルータと対面するインターフェースが落ちても迂回して到達することが可能
    - LoopbackインターフェースにもIGPを起動することを忘れずに
  - 全BGPルータで同じIPアドレスで対象ルータを認識することが可能
    - IXなどに接続するボーダルータで、且つそこに2台以上のルータを接続した場合には注意が必要。
      - iBGPがIX越しに接続される可能性があり、思わぬ経路トラブルを起こす可能性が高い

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

iBGPの基本的コンフィグレーション

```
Interface Loopback 0
 ip address 202.216.41.1 255.255.255.255
!
Interface FastEthernet 2/0
 description NSPIXP2 Segment
 ip address 202.249.2.41 255.255.255.0
!
Router ospf 4689
 network 202.216.41.1 0.0.0.0 area 0
 network 202.249.2.0 0.0.0.255 area 0
 passive-interface loopback 0
 passive-interface FastEtheret 2/0
!
Router bgp 4689
 neighbor IBGP peer-group
 neighbor IBGP remote-as 4689
 neighbor IBGP update-source Loopback 0
 neighbor 202.216.41.2 peer-group IBGP
 neighbor 202.216.41.3 peer-group IBGP
 neighbor 202.216.41.4 peer-group IBGP
```

- Loopback 0 の設定　/32で構わない
- FastE2/0 がIXセグメントだったとする
- LoopbackとIXセグメントをOSPF上で定義、かつ非活性とする。これによって他のBGPルータでもそれぞれがIGP上で認識される
- peer-groupを利用してみる。等質なコンフィグには非常に有効
- Update-source で、ピアリングに利用するIPアドレスを定義する
- iBGPにloopbackアドレスを利用すると、BGPルータをIPアドレスで認識できるので運用上非常に便利

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

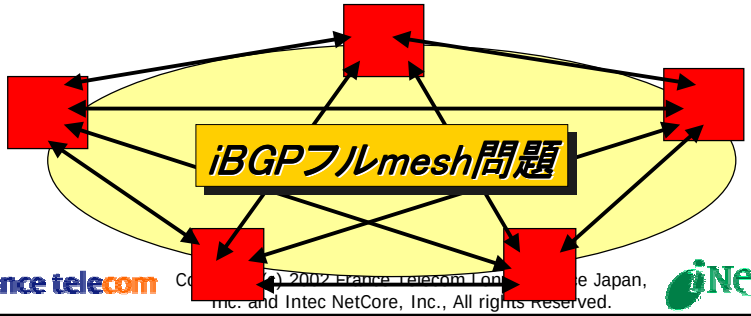
ISPネットワーク拡大に沿った規模対応設計

iBGPシステムのスケーラビリティ

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

iBGPシステムのスケーラビリティ

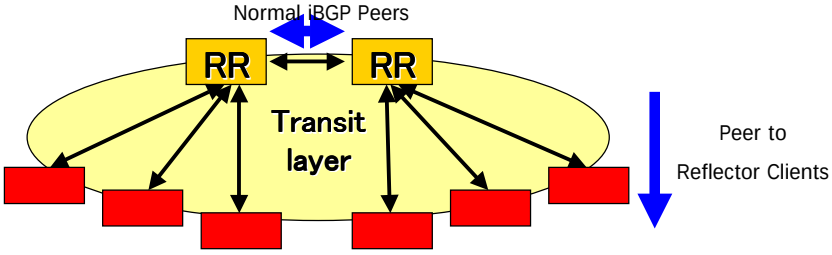
- iBGPで得た経路は他のiBGPpeerに再伝播しないため、全ノードをmesh状にpeerする
  - ボーダルータ5ノードで既に10peer
  - 10ノードでは? $_{10}C_2 = 45$
    - 11ノード目の増設にあたって10peerの追加



Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

iBGPフルmesh問題解決策 iBGPルートリフレクタ(1)

- リフレクタとリフレクタクライアントの2階層化
- リフレクタからクライアントにはiBGPで得た経路を再分配する



Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

iBGPフルmesh問題解決策 iBGPルートリフレクタ(2)

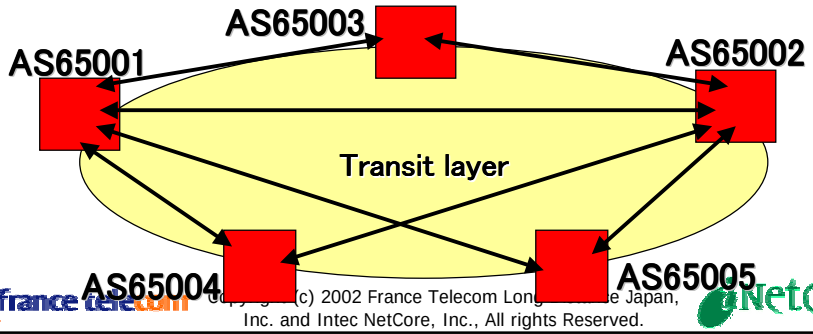
- コンフィグレーション
  - リフレクタ側で以下のように設定
  - クライアント側では設定不要
    - 階層化可能
  - 階層化しない場合はリフレクタ同士は以前Full Meshな構成が必要
    - Non-Client BGP Peers

```
router bgp 14186
 bgp cluster-id FOUR-BYTE-CLUSTER-ID
 neighbor CLI.ENT.IPA.DDR remote-as 14186
 neighbor CLI.ENT.IPA.DDR route-reflector-client
```

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

iBGPフルmesh問題解決策 BGPコンフェデレーション(1)

- BGPコンフェデレーション(confederation)
  - ASの中を更に小さい単位でsubASに分け、その間をeBGPで結ぶ
  - フルmeshにはる必要はなくなる
    - Sub-AS内でのFull Meshは依然必要



Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

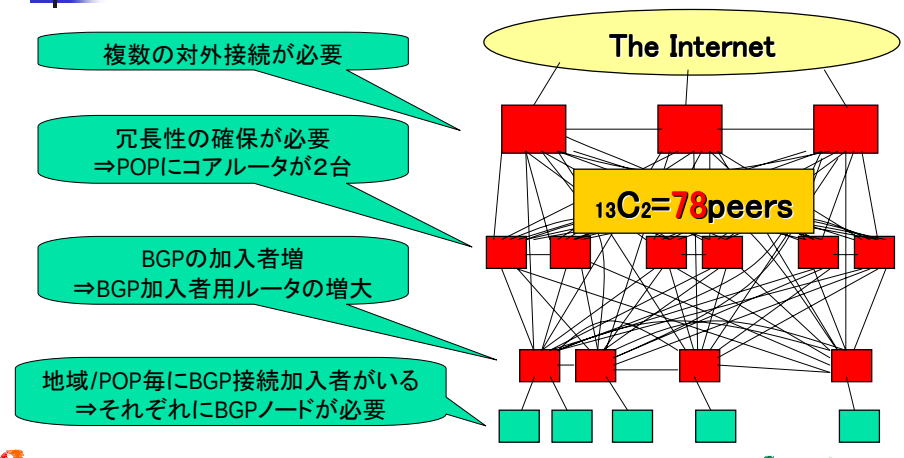
iBGPフルmesh問題解決策 BGPコンフェデレーション(2)

- コンフィグレーション
  - プライベートASを利用するのが普通
  - Confed内部となるAS番号をconfed peersで定義

```
router bgp 65000
  bgp confederation identifier 4689
  bgp confederation peers 65001 65002 65003 65004
  network .....
```

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

AS内BGPスケーラビリティ問題の実際



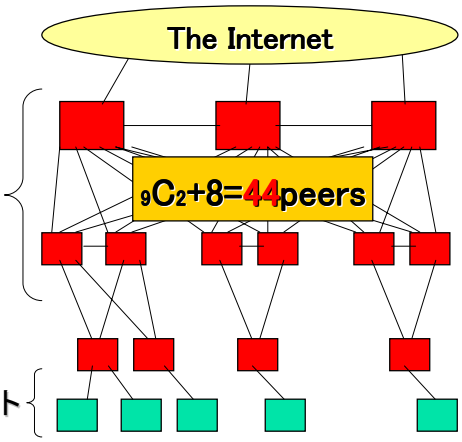
Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

105

AS内BGPスケーラビリティ問題の実際 ～RRによる解法～

- RRの導入

POPコアルータと対外接続ルータをフルメッシュ



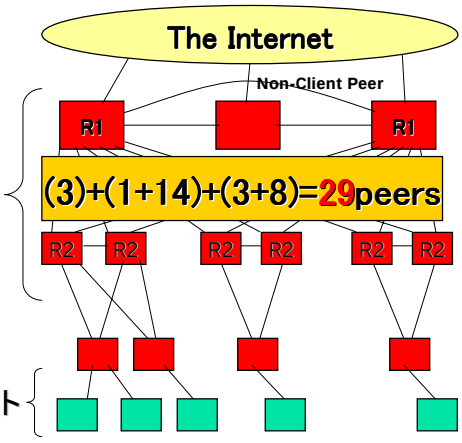
Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

106

AS内BGPスケーラビリティ問題の実際 ～RRによる解法～

- RRの導入

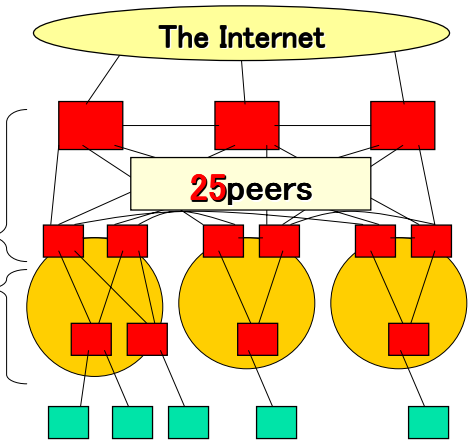
設計上この階層をさらにReflector-Clientの構成にすればさらにPeerの数は減る



Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

AS内BGPスケーラビリティ問題の実際 ～コンフェデレーションによる解法～

- 地域・POPごとにsubASを設定
- BGP加入者収容ルータとの間にiBGPを設定
- IGPは分割，単一どちらでもOK



Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

ISPネットワーク拡大に沿った規模対応設計

スケーラビリティとトラブル回避

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

eBGPのスケーラビリティ

- 経路数
  - 110,000経路(**) ⇒ 最近の伸びはほとんどない。
  - 所要メモリサイズに影響
    - 256MB必要
- Peerの数
  - IXで多数のpeerを張るとメモリ所要に影響
  - 50peer程度+upsteamで10MB程度余分に消費

**http://www.apnic.net/stats/bgp/TOTAL/totalann.html

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

eBGPの問題回避技術(1)-1

- 誤広告対策
  - 隣接ASが広告する経路は完全にいつも正しいとは限らない
    - 誤った経路受領は障害の原因となる
  - AS-pathによるフィルタリング
    - 隣接ASから広告する旨を予め知らせてもらったAS-pathの経路しか受け取らない
    - Prefixが間違っている場合には防げない。

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

eBGPの問題回避技術(1)-2

- プリフィクスフィルタリング
  - (主に顧客の場合)顧客が広告するプリフィクスを予め知らせてもらい、フィルタする
    - Prefixが変わった際にいちいち設定が必要で、ISP-顧客間のやり取りが煩雑。
  - Dynamic Route Filtering
    - IRRの登録情報を元にFilterを自動生成して運用する。
    - システムに問題が発生した場合にすべてがFilterされる可能性がある
- Maximum-prefix を絞る
  - Neighbor NE.IG.HB.OR maximum-prefix 1000 900
    - C社コマンド。1000経路までしか受けず、900でアラーム
  - 当該ASからの設定ミスなどによる大量アナウンスを防止する。特定のPrefixをブロックすることはできない。

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

eBGPの問題回避技術(2)

- Route flapping
  - リンク不安定などによる経路広告のばたつき
  - 経路更新，消去の連続でCPUリソースを浪費
  - 対処策：Flap Dampening
    - ..(config-router)# bgp dampening　c社コマンド
    - ばたつく経路に一定時間のペナルティを課して、経路テーブルから消す
    - 一方でデフォルトの設定では、メンテナンスなどで正常なアナウンスをした経路までブロックしてしまう可能性がある。
      - Non-Transitiveの副作用 / 実装依存
      - http://www.nanog.org/mtg-0210/flap.html

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

eBGPの問題回避技術(3)

- ポリシ変更の負担軽減
  - ポリシ変更を反映には、peerのクリアが必要
    - Upstreamの場合、full route を受けるため負担
  - 対処策： soft-reconfiguration　c社機能
    - クリアなしに経路に対するポリシ反映
    - Outbound はコンフィグそのままで実行可能
      - Clear ip bgp PEER soft out
      - 一旦広告していた経路を取り消して、再広告
    - Inbound はneighbor定義が必要
      - Neighbor ADDRESS soft-reconfiguration inbound
      - ネイバから受けたそのものを蓄えておき、それに対して新たなポリシを適用
      - メモリが余分に必要なので注意。Full routeで10MB程度
  - 対処策： Graceful Restart
    - Peerのクリアを行っても、Peerが再度張りなおされた後、Peerが切れている間にUpdateされた経路のみが、広報されUpdateの負荷を最小限にする仕組み
      - http://www.ietf.org/proceedings/02mar/I-D/draft-ietf-idr-restart-02.txt

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

ポリシルーティング

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

ポリシルーティング

- BGPにおける経路情報の扱い
  - プリフィクス(NLRI)＋パス属性
  - パス属性値の調整，パス属性値に基づく経路選択を行うことができる
- ルーティングポリシ
  - 複数peerを持つASとの間でどのようにトラヒックを交換するか
  - セキュリティのために経路をフィルタする
  - 複数のupstreamに対するトラヒックバランス

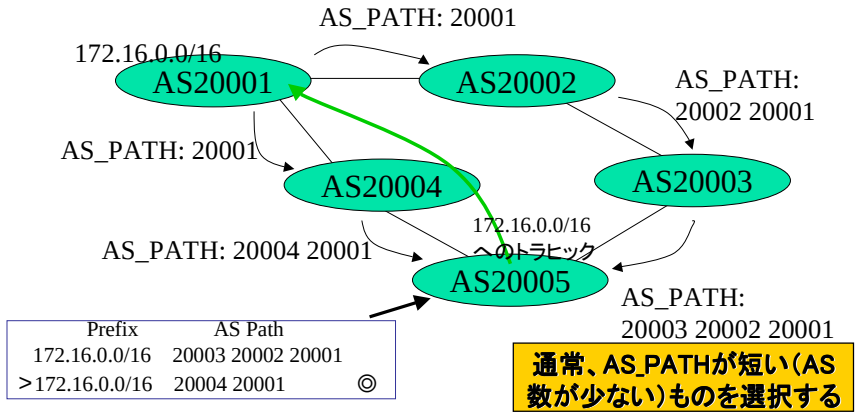
Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

ポリシルーティングを可能にするパス属性値

- AS_PATH
  - 経過AS列，短いほうが優先。
    - ただし,RFC1771には規定なし
  - AS-path prependでAS列長の調整が可能
- LOCAL_PREF – Local Preference
  - 設計者意図の優先順位付け
- MULTI_EXIT_DISC – Multi Exit Discriminator (MED)
  - 隣接するAS間で複数peerがある場合の優先度
- COMMUNITY – Community Attribute
  - 32ビットの値を付加できる。プロトコル上、値に意味はないが、有効な利用法がカレントプラクティスに存在
  - ルータごとにCommunityの値を判断してPrepend, LocalPref, MEDなどを適宜負荷することが可能

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

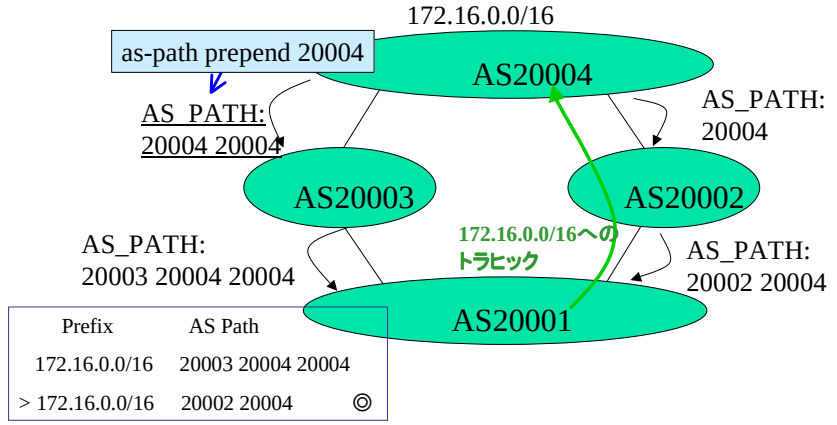
AS_PATH



Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

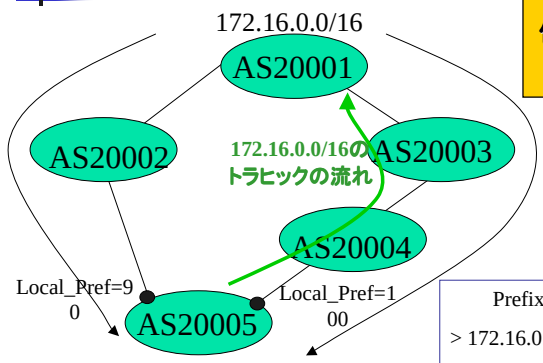
AS Path Prepend

ASを余計につけて、AS_PATH_lengthを長く見せるテクニック



Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

LOCAL_PREF



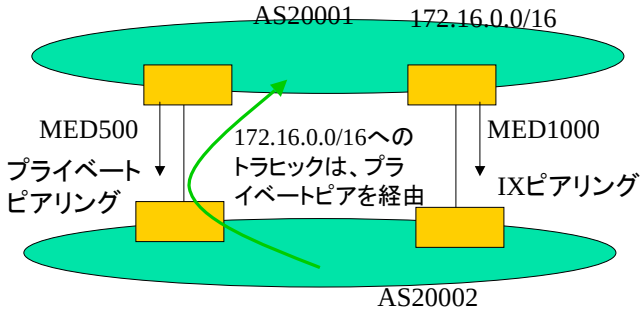
他ASから受領する経路に付加し、大きいほうが優先される

通常 AS_Pathの小さいAS2経由が優先だが、Local_Prefにより、AS4の経路を優先する。

	Prefix	AS Path	LocPref	
>	172.16.0.0/16	20004 20003 20001	100	◎
	172.16.0.0/16	20002 20001	90	

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

MULTI_EXIT_DISC



	Prefix	AS Path	MED	
	172.16..0.0/24	20001	200	
>	172.16..0.0/24	20001	100	◎

広告する経路に付加し隣接ASに伝えられ、隣接ASにおける経路選択で使われる。数値が小さい方が優先

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

COMMUNITY(1)

- 32ビットの整数値，透過性
- Well-known Community
  - No-export:
    - 自AS以外に広告しない
  - No-advertise:
    - 受領したルータ以降に広告しない
- Well-known ではないCommunity
  - 経路情報を受領したAS，ルータで解釈させ、何らかのポリシ付加(Prepend, LocalPrefなど)を発生させる

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

COMMUNITY(2)

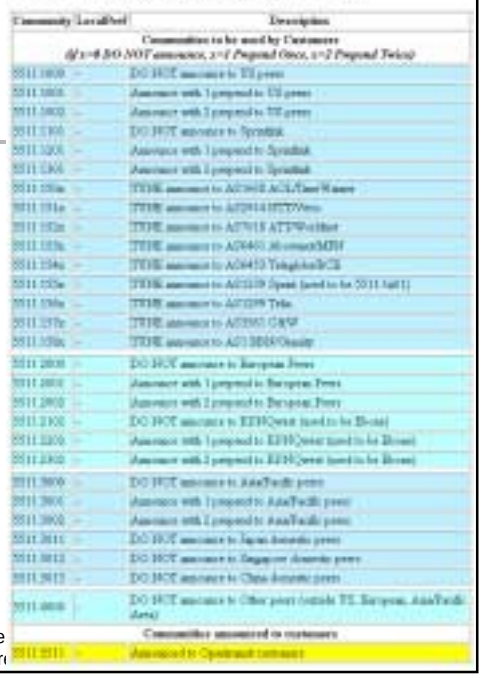
- 一般的な利用法
  - New-format – 32ビットを16ビットずつに二分
    - 5511:1000
  - 上位 – ターゲットAS
  - 下位 – ターゲットASでの動作
- 例1:RFC1998　　MCI（現CWnet）における実装例
  - 3561:70　そのプリフィクスにLocPref=70付与
  - 3561:80　そのプリフィクスにLocPref=80付与
  - . . . .
  - そのASからの戻りトラヒックの制御に便利！

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

COMMUNITY (3)

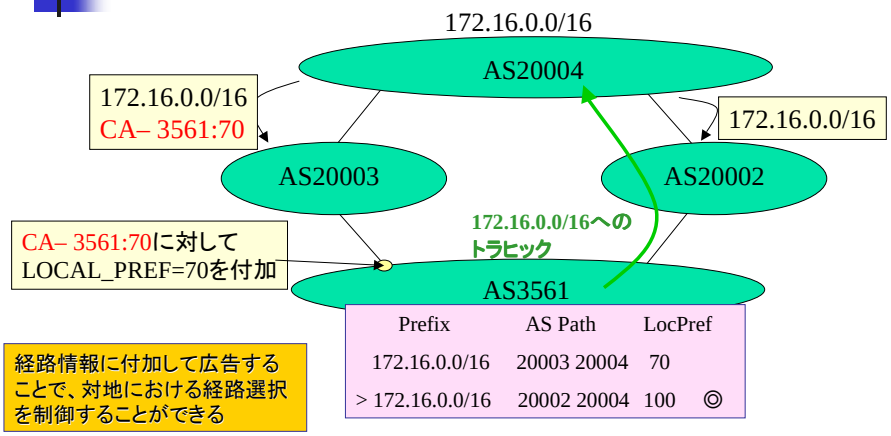
- AS5511の例
  - 地域，ピアごとに、広告経路上のprepend及び非広告を指定可能

BGP Community in Opentransit



Copyright (c) 2002 France Inc. and Intec NetCor

COMMUNITYの利用方法



経路情報に付加して広告することで、対地における経路選択を制御することができる

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

BGPの最適経路の決定プロセス

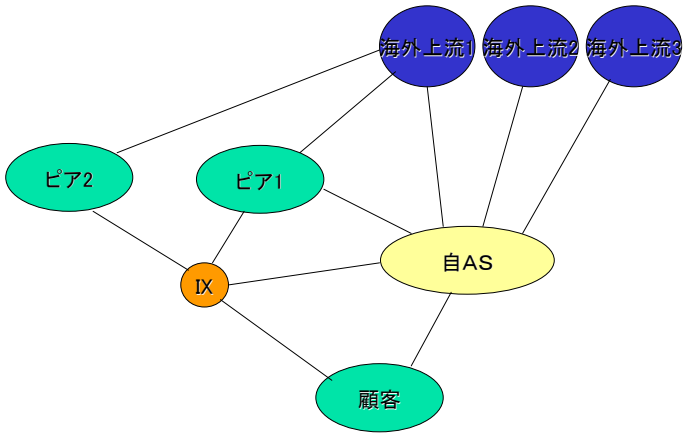
- 同一プリフィクスの経路情報が複数があるとき、パス属性値に拠って最適方路を決定
  - 以下、ciscoの例
  - 1. Local Preferenceが大きい
  - 2. AS_PATHが短い
  - 3. MEDが小さい
  - 4. IGP上でNext-hopが近い(cost/metric)
  - 5. BGPのルータIDが小さい
    - 正確には5つの判断基準ではなく、IP Addressが小さいものを選ぶなど、細かいものを含めると10個の基準がある。

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

ポリシルーティングの実際

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

相互接続の例



Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

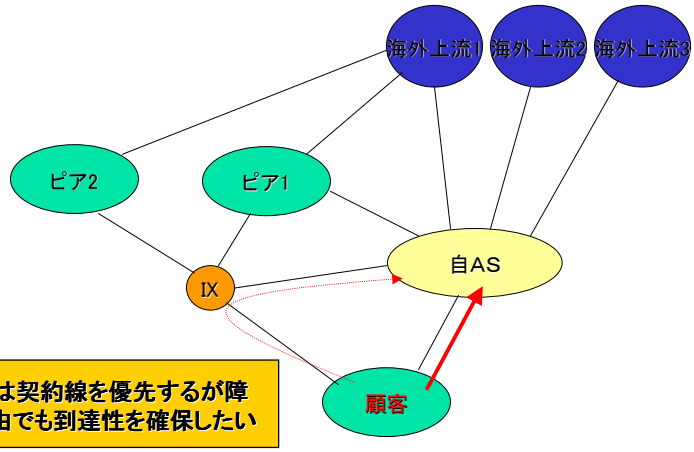
ポリシルーティングの基本検討

- 相互接続別・対地別の基本ポリシ付け
  - Outbound/Inbound を対にして、どういう経路を交換するか
  - 相互接続別
    - 顧客　　フルルート供給，顧客経路のみ受領
    - ピア相手　　自網顧客経路のみを相互に交換
    - 海外上流　　自網顧客経路のみ供給，フルルート受領
  - 対地別（優先する順番にパスを並べる）
    - 顧客　　直接，IX経由，Upstream経由
    - 国内対地　　プライベートピア経由，IXピア経由，Upstream経由
    - 海外対地　　安い順番，品質の良い順番，とりあえず無制御

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

129

受領経路優先順序検討(国内)

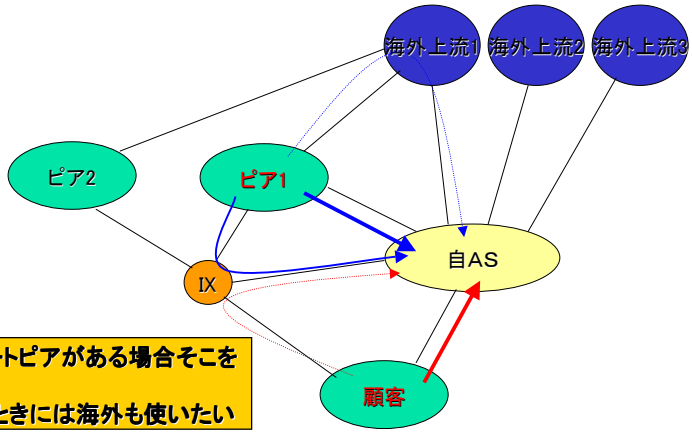


・顧客に対しては契約線を優先するが障害時にはIX経由でも到達性を確保したい

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

130

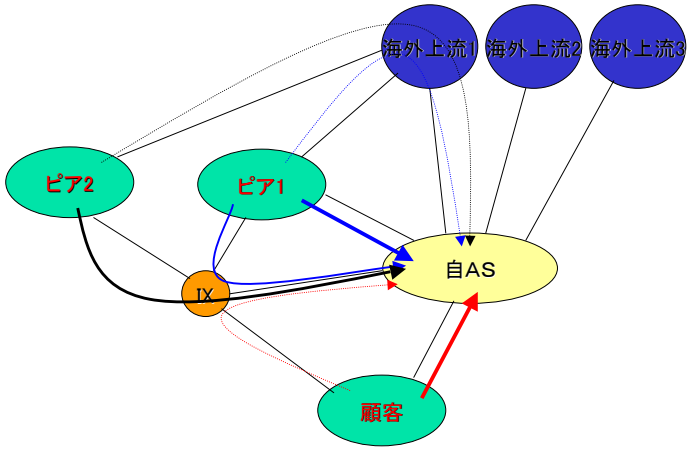
受領経路優先順序検討(国内)



・ピアはプライベートピアがある場合そこを優先にしたい
・国内が全滅したときには海外も使いたい

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

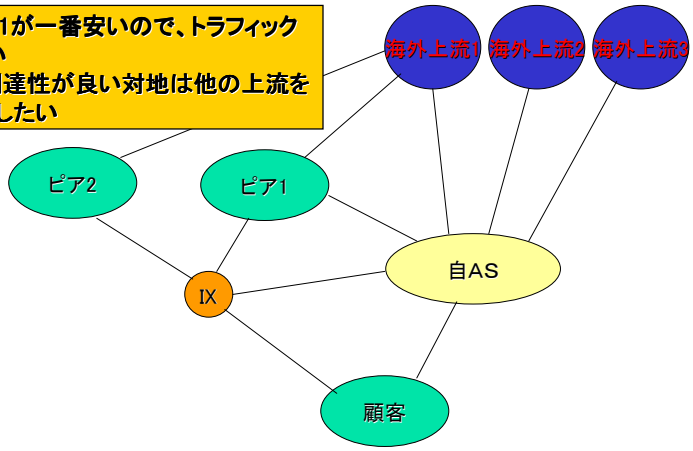
受領経路優先順序検討(国内)



Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

受領経路優先順序検討(海外)

・海外上流1が一番安いので、トラフィックを集めたい
・しかし、到達性が良い対地は他の上流を使うようにしたい



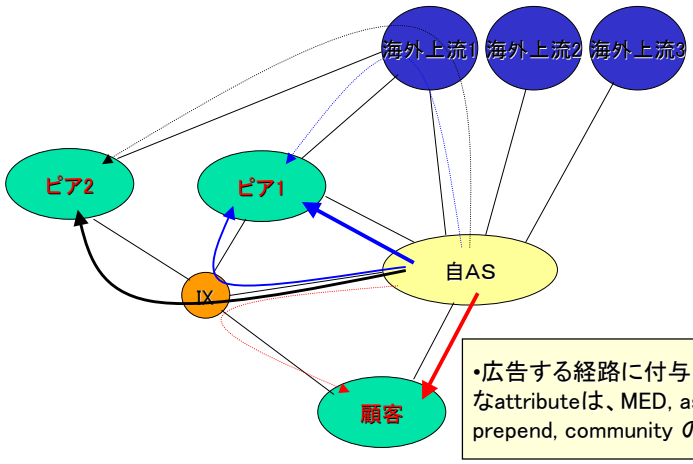
Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

受領経路に関するルーティングポリシ実装案

- 各eBGPピアで、受領経路に対して以下の通りLOCAL_PREFを付与する
  - 顧客 110
  - プライベートピアリング 100
  - IXピアリング 95
  - 海外上流 90
- 海外上流に関して、上流2，上流3から受領する経路にAS-path prepend を1hop掛ける

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

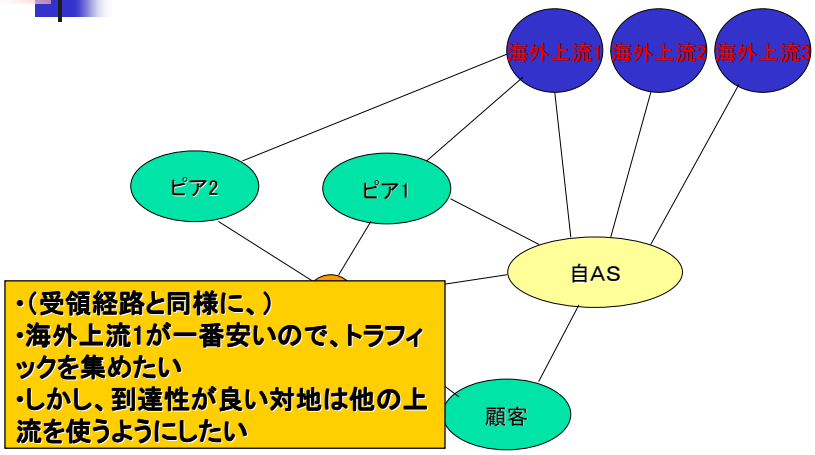
広告経路ポリシ検討（国内）



•広告する経路に付与して有効なattributeは、MED, as-path prepend, community のみ

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

広告経路ポリシ検討（海外）



Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

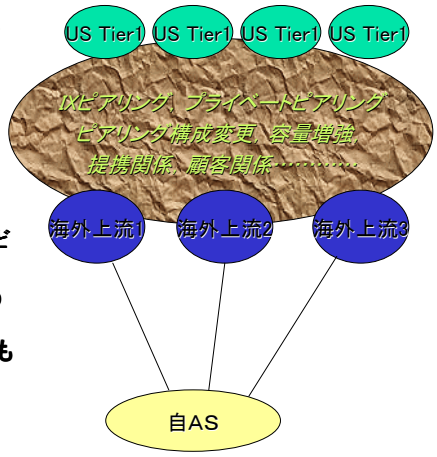
広告経路に関するルーティングポリシ実装案

- 各eBGPピアで、広告経路に対して以下の通りMULTI_EXIT_DISCを付与する
  - 顧客　500
  - プライベートピアリング　900
  - IXピアリング　1000
- 海外上流に関して、上流2，上流3に広告する経路にAS-path prepend を1hop掛ける

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

海外上流のトラフィック制御の難しさ

- 海外上流からのinboundトラフィックのバランスは日々変化する
  - 接続構成は常に更新される
- 調整方法としては、
  - as-path prepend,
  - Community
  - プリフィクスの一部のみ適用
    - 広報Prefixを分割してアナウンスするなど
- 精密な調整が不要な工夫が必要
  - 全部従量課金サービスにしてコストへのインパクトを少なくする
  - 十分なキャパシティを準備して突出しても性能低下にならないようにする
    - キャパシティプランニングの重要性大



Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

スケーラブルな経路制御設計

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

BGPとOSPFの比較(1)

OSPF	BGP
IGP : Interior Gateway Protocol	EGP : Exterior Gateway Protocol
IP上に直接乗るプロトコル Protocol number: 89	TCP上に乗るプロトコル Port number: 179
リンクステート型プロトコル リンクステート情報を伝播 状態変更毎にLSA，連鎖伝播	パスベクター型プロトコル パス情報を伝播 状態変更毎にUPDATE，連鎖伝播

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

BGPとOSPFの比較(2)

OSPF	BGP
基本的に、OSPFを起動した隣接ルータ全てと経路交換 マルチキャストでセグメント上の全OSPFルータとやりとり	明示的に定義した隣接ルータのみと経路交換 隣接ルータ毎にBGPセッションを確立（ピアリング）
あるネットワーク（ルータ）の状態変更は、全ルータのパスツリー再作成を引き起こす 30分でリフレッシュ--flooding	あるネットワークの状態変化は基本的にはそのプリフィクスだけの問題 リフレッシュなし

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

BGPとOSPFの比較(3)

OSPF	BGP
トポロジの管理に主眼を置く	プリフィクス(ネットワーク)のパス属性に着目
エリア内共通のLSDBを全ルータが作成し、LSDBから各ルータそれぞれがパスツリーを作成	受領したUPDATEは各AS，ルータのポリシに基づいて処理，以遠伝播する
経路個別のポリシ付加は不可	経路個別にポリシ付加が可能 →パス属性値としてプリフィクスに付加
精密で敏速な経路制御	ポリシに基づいた経路制御

2002 France Teleco
ntec NetCore, Inc.,

Scalable Routing Design Principles

- RFC2791, July 2000, Jessica Yu
- ドラフトの日本語訳あり
  - http://www.janog.gr.jp/doc/draft-yu-routing-scaling-01-j.txt

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

143

スケーラビリティ確保のための指針

- 階層構造化
  - iBGP RRの階層化，OSPFのバックボーンエリアと他エリア
- 区画化
  - BGP Confederation，OSPFのエリア分割
- 適切なトレードオフの設定
  - BGP flap dampening
- 経路制御処理の負担を軽減
  - 経路集成，集約
- スケーラブルな経路制御ポリシ，実装
  - できるだけシンプルにする，できるだけ自動化する

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

144

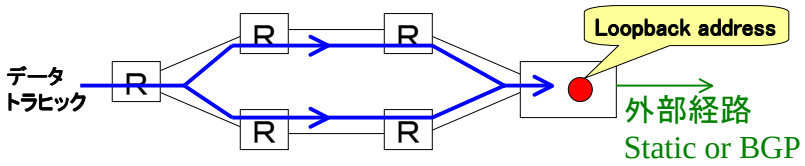
世界規模ISPにおける典型的なネットワーク構成

- 全ルータでBGPが起動される
  - そもそも末端ルータでもメモリフル実装
- BGPはルートリフレクタで階層化
- 加入者ルータ以外は二重化構成
- IGP(IS-ISが多い)によるロードバランシング実現
- Staticは場合によってはBGPにredistribute

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

OSPFによるNEXT_HOPへのロードバランシングの仕組み

- その経路へデータが行くためにはBGP next-hopであるredistributeしたルータのloopbackアドレスへ向かおうとする
- BGP next-hopへ向けてOSPFで作られたルーティングテーブルをrecursive lookupする
  - ロードバランスする



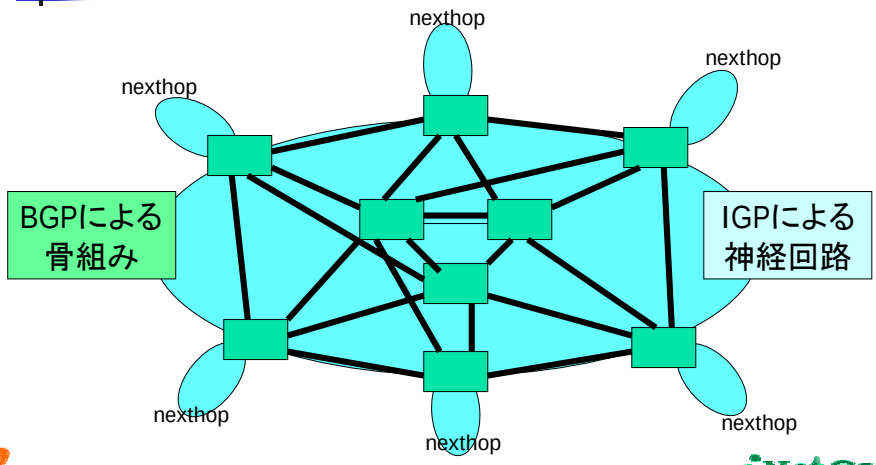
Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

BGPとOSPFの分担

- OSPFはトポロジ管理に関しては精巧だが、外部経路を扱うことは不得手
  - 外部経路のフィルタリングも難しい
  - たとえスタティックルートでも、多くなると安定しない
- BGPはトポロジ管理はできないが、外部経路のコントロールは非常に得意
  - ポリシの付加やフィルタリングも容易

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

骨組みと神経回路



Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

参考文献

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

参考文献

- RFC2791 - Scalable Routing Design Principles
  - Jessica Yu
- インターネットルーティングアーキテクチャ 第2版
  - Sam Halabi / Danny McPherson著, 鈴木 訳
- インターネットルーティング入門
  - 友近・池尻・小早川 著, 翔泳社
- インターネットルーティング
  - C. Huitema 著, 前村 監修・エクストランス 訳, 翔泳社

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.

Question?

Kuniaki KONDO
kuniaki@inetcore.com
Akinori MAEMURA
akinori.maemura@francetelecom.com

Copyright (c) 2002 France Telecom Long Distance Japan, Inc. and Intec NetCore, Inc., All rights Reserved.