ISPバックボーンネットワークにおける経路制御設計
～ 理論編 ～

前村　昌紀

フランステレコム
ネットワーク・アンド・キャリア・ディビジョン
アジア地域IPプロダクト担当

france telecom

目次

- ルーティング基礎事項の整理
  - ルーティングとは，RIPのおさらい，クラスレス
- OSPF – Open Shortest Path First
  - プロセス確立，LSDB，LSA，
  - 方路選択と耐障害性，IS-ISとの比較
- BGP – Border Gateway Protocol
  - ASと階層的経路制御，パケットタイプ，
  - パス属性値，ポリシルーティング
  - iBGPシステムの設計，スケーラビリティなど

ルーティング基礎事項の整理

ISPバックボーンネットワークにおける経路制御設計～ 理論編 ～

france telecom

セクション目次

- ルーティング基礎事項の整理
  - ルーティングとは何か，原型としてのRIP
  - クラスフルとクラスレス
  - IGP,EGP,ASとアルゴリズム

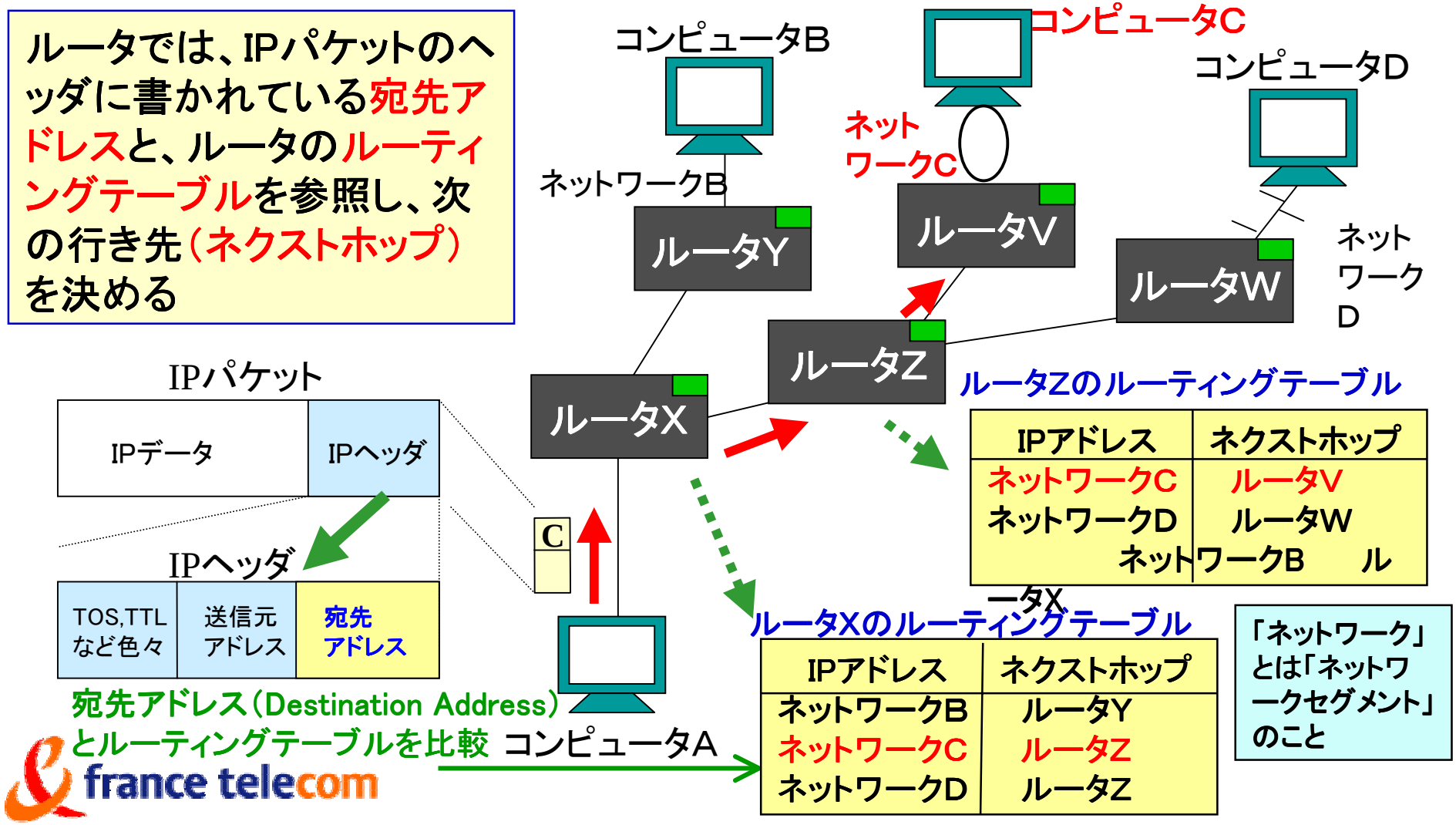
ルーティングとは何か
原型としてのRIP

ISPバックボーンネットワークにおける経路制御設計～ 理論編 ～
ルーティング基礎事項の整理

france telecom

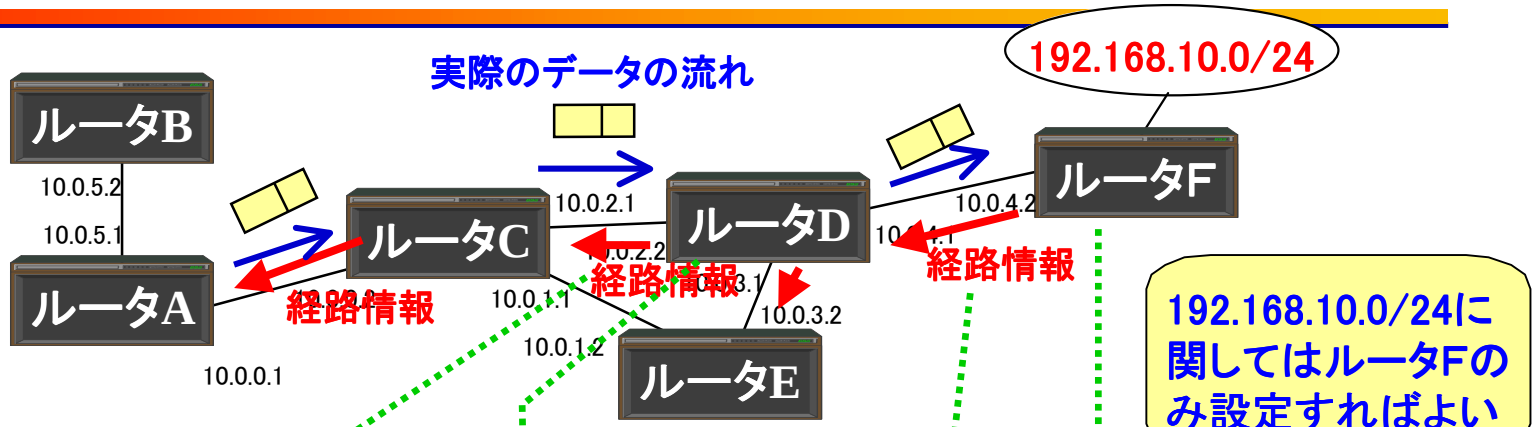
ルータとルーティング

ルータでは、IPパケットのヘッダに書かれている宛先アドレスと、ルータのルーティングテーブルを参照し、次の行き先（ネクストホップ）を決める



ダイナミックルーティング

IPパケットが、192.168.10.0/24に行きたいとき



192.168.10.0/24に関してはルータFのみ設定すればよい

①:ルータFは192.168.10.0/24は自分のところにあるぞ！とみんなに知らせる。

②:①の情報(経路情報)がルータFからルータDに伝わる

③:ルータDは経路情報を受けてルーティングテーブルを作成する。

ルータD のルーティングテーブル

IPアドレス	ネクストホップ
..*.*/*	10.0.2.1
..*.*/*	10.0.3.2
192.168.10.0/24	10.0.4.2

④:192.168.10.0/24はこっちの方にあるという情報がルータDからルータCやルータEに伝わる。

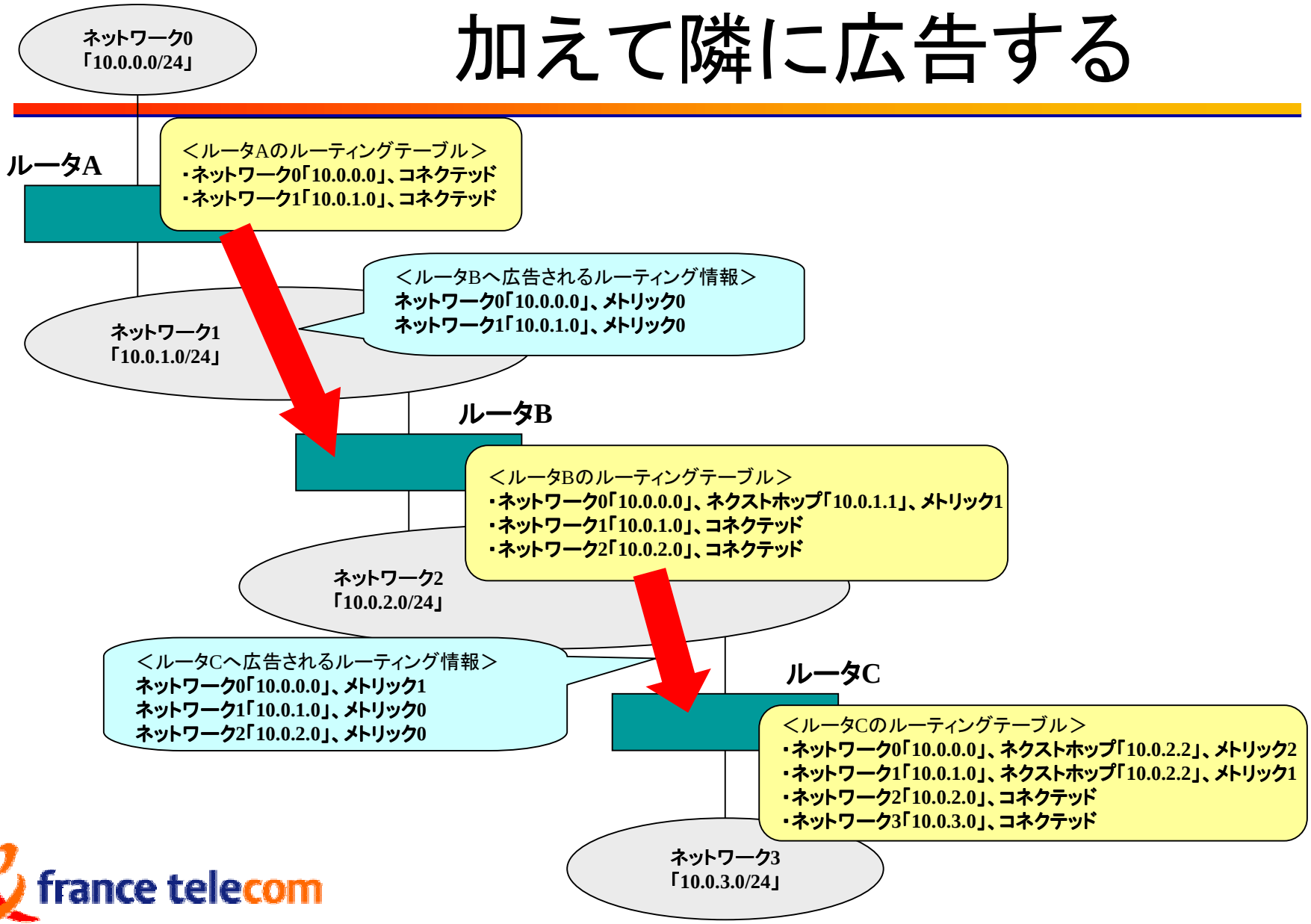
⑤:ルータCやルータEは経路情報を受けてルーティングテーブルを作成する。

⑥:その経路情報がルータAに伝わり、ルータAでルーティングテーブルが作成される。

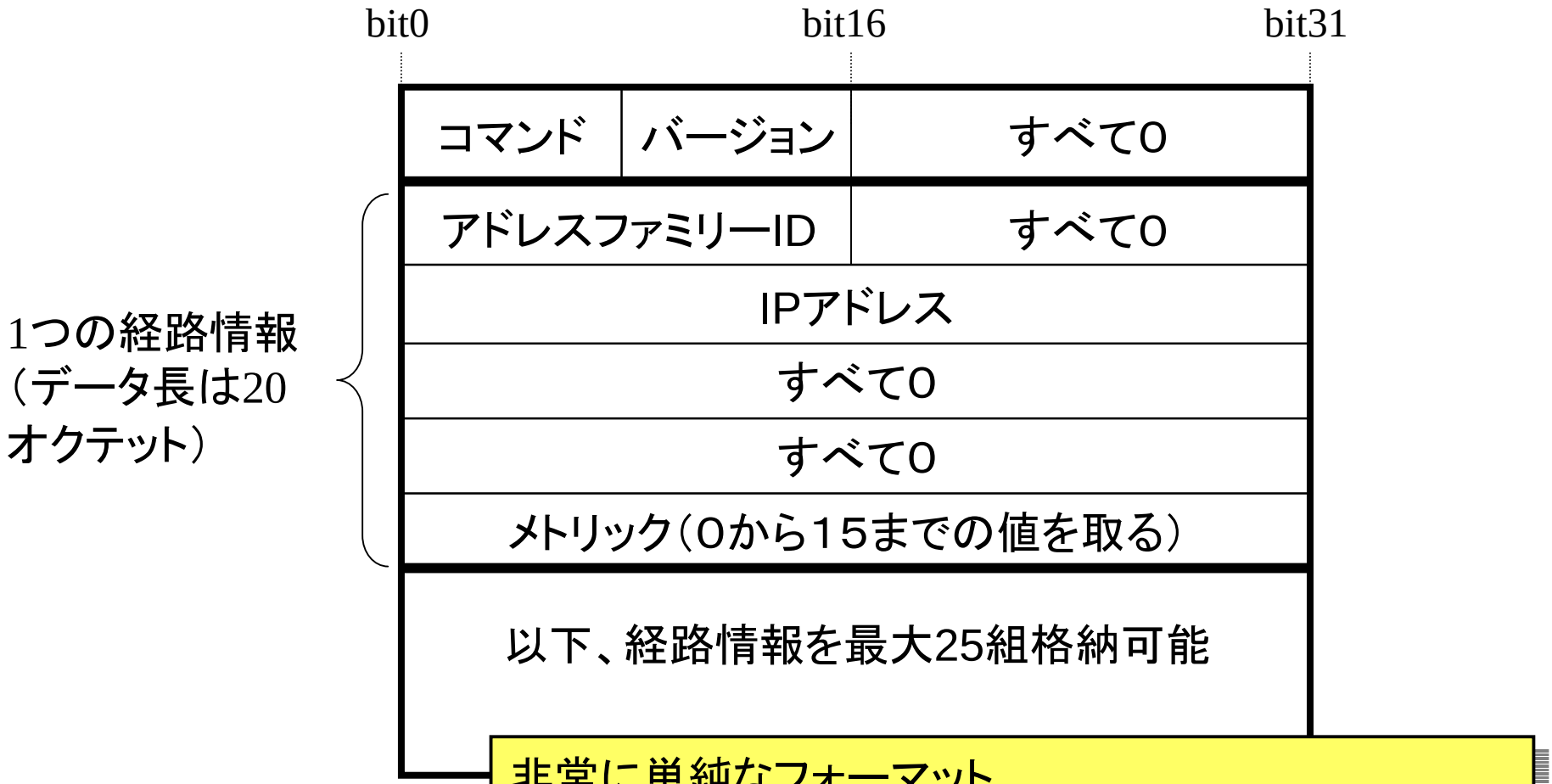
RIP – Routing Information Protocol , リップ

- 原始的で簡素なルーティングプロトコル
  - 古くからUNIXシステム上で実装されている(routed)。
  - UDP 520番ポートを利用(セッションを確立しない)
  - 自分のもっているルーティングテーブルを接続しているネットワークに30秒ごとにブロードキャストする
  - ネットマスクの情報を運ばない
  - 隣接したルータから受け取った情報(ネットワークアドレス)に自分の知っている情報を付加し送信する
  - これが全ルータの間で繰り返し行われることでルータは接続されたすべてのネットワークとそこへの道筋を知る。

自分が知っている情報を加えて隣に広告する



パケットフォーマット



非常に単純なフォーマット
IPアドレスに対するサブネット情報のフィールドがない

RIPのメリットとデメリット

- メリット
  - 処理の負荷が小さい（軽い）・実装が簡単
  - どんなルータでも利用可能，ホストも対応可能
- デメリット
  - 30秒に一度全経路情報を伝える
    - 経路数が多くなると無駄が多い
  - ネットマスクの情報を運ばない
    - クラスフルなルーティングプロトコル
  - 収束に時間がかかる
  - 最大のホップ数は15までしか対応できない
  - ホップ数で比較なので、回線の帯域に応じて適切な経路を選ぶことが難しい

RIP2 – RIPの改良版

- ネットマスク長情報を運ぶように改良
  - クラスレスプロトコルとなった

- しかし、RIP1と同様、30秒に一度全経路情報を伝えようとする
- 実装が少ない – RIP最大のメリットが消失

結局あまり普及せず今にいたる

france telecom

クラスフルとクラスレス

ISPバックボーンネットワークにおける経路制御設計～ 理論編 ～
ルーティング基礎事項の整理

france telecom

Class A, B, C とは
– クラスフルアドレス –

- IPv4の初期設計において、上位3ビットによってIPアドレスのどこまでがネットワークを示し、どこまでがネットワーク中のホストを示すか、自動的に分かるようにした。

4bits	Octet#1	Class	ネットワークアドレス	ホスト数
0???	0—127	A	第一オクテットだけ	16.8M
10??	128—191	B	第二オクテットまで	65534
110?	192—223	C	第三オクテットまで	254

クラスフルなルーティングとルーティングプロトコル

- プロトコルにネットマスクを扱う機能がない
- Class A, B, Cに従ってネットワークアドレスの認識を行う
- その中を更に分割したものをサブネットと言う
  - プロトコルを通じてネットマスクを伝達できない
  - サブネッティングは、自身のインターフェースに定義したものを参照して解釈する
    - 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
- 自身が属するネットワークアドレス以外は、サブネットとして認識できない
  - サブネットが2方向以上に散らばっていると経路制御不能

クラスレスルーティング

- プロトコルがネットマスクの情報も扱う
  - ネットワークを示すもの==プリフィクス(Prefix)
  - 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)

IGP, EGP, AS
アルゴリズム

ISPバックボーンネットワークにおける経路制御設計～ 理論編 ～
ルーティング基礎事項の整理

france telecom

IGP, EGP, AS

- AS – Autonomous System = 自律システム
  - (基本的には)ISPを1つの単位としてASを定める
  - ASの中外で別々の独立したルーティングを行う
- IGP – Interior Gateway Protocol
  - 内部(ゲートウェイ)プロトコル
  - 個別のネットワークセグメントまでに至る詳細なルーティングを実施する
  - RIP, OSPF, IS-IS, IGRP, EIGRP,,,
- EGP – Exterior Gateway Protocol
  - 外部(ゲートウェイ)プロトコル
  - ASを単位としてAS間の経路制御を行う
  - 実質的にBGPだけ。原型にEGP, 他にIDRP

アルゴリズムの種類

- それぞれのプロトコルのアルゴリズムの特徴を表現したもの
- ディスタンスベクター – distance vector
  - RIPなど。「距離」と「方向」を扱う、という意味
- パスベクター – path vector
  - BGP。「パス属性」と「方向」を扱う、という意味
- リンクステート – link state
  - 管理領域内の全てのネットワークセグメント（リンク）の状態情報を収集して管理する方式

OSPF
– Open Shortest Path First

ISPバックボーンネットワークにおける経路制御設計～ 理論編 ～

france telecom

セクション目次

- OSPF – Open Shortest Path First
  - OSPFを簡潔に説明する
  - OSPFのプロセスを確立する
  - リンクステートデータベース
  - リンクステート広告
  - OSPFの方路選択と耐障害性
  - 基本的な適用技術
  - IS-ISとの比較

OSPFを簡潔に説明する

ISPバックボーンネットワークにおける経路制御設計～ 理論編 ～

OSPF – Open Shortest Path First

france telecom

OSPFのメリット：
リンクステート型の精密な動作

- リンクステート型プロトコル
  - 全ルータが全リンク状態をDBで管理して、そこからルーティングテーブルを生成する
- リンクステート型ならではの精密なネットワーク管理
  - バックアップパス，イコールコストマルチパスの形成，障害発生時の素早い収束

リンクステート – Link State

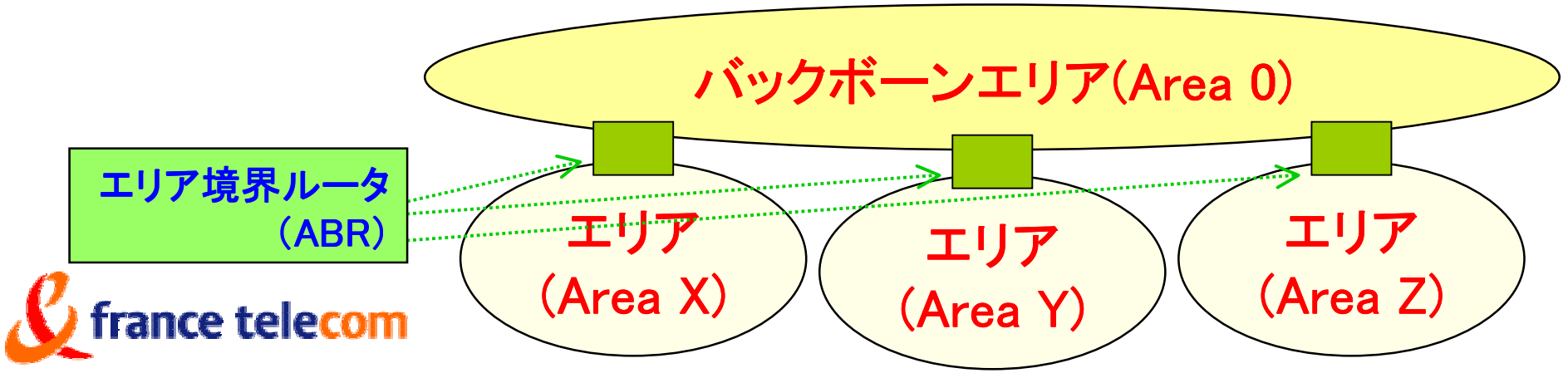
- リンク（ネットワークセグメント）の状態情報
  - 死活，IPアドレス，マスク，ネットワークタイプ，接続ルータ，メトリック
  - 「リンクステートアドバタイズメント（LSA）」で伝達
- 管理領域内の全てのルータは、管理領域全域のリンクステートを全て収集し、データベース化して管理する
  - リンクステートデータベース（LSDB）
  - RIPのような原始的なダイナミックルーティングとは全く異なる
- リンクの状態が変化したときにその変更を他のルータに伝える
  - リンクの死活や接続関係（トポロジ）の変化
  - 実は、変化のないときでも定期的に30分に1回リフレッシュを実施

OSPFのネットワーク設計

- 管理するネットワークをエリアに分割して、エリア外の管理を簡素化
  - クラスレス対応，エリアごとの経路集約が可能
- DR（指名ルータ），BDR（バックアップ指名ルータ）を利用した、マルチプロアクセスネットワークにおけるプロトコル負荷の軽減
  - ネットワーク設計上の考慮が必要

エリア

- エリア
  - OSPF管理領域をいくつかのエリアに分割する
  - エリアIDは32ビット.IPv4アドレスを割り当てることが可能
  - バックボーンエリア(area0)に他のエリアがぶら下がる形
  - LSDBはエリア内の全ルータで共通,エリア外のトポロジは管理せず、サマリのみを管理
- エリア境界ルータ(ABR: Area Border Router)
  - バックボーンエリア(エリア0)と他のエリアをつなぐルータ



OSPFその他の特徴

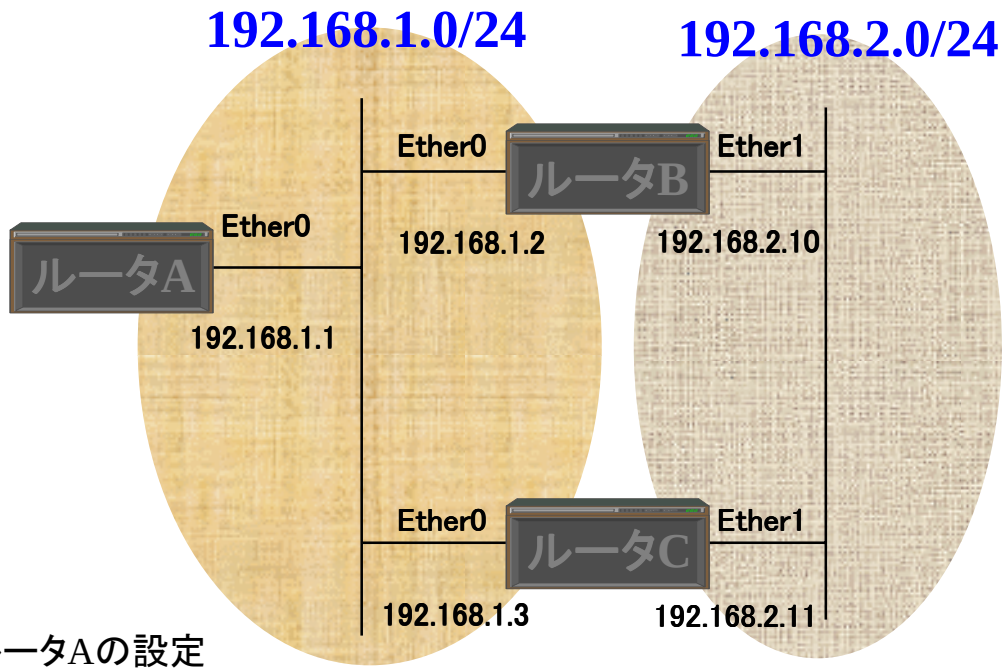
- 精密なルーティングを目指して、
  - OSIのIS-ISに見習って開発されたルーティングプロトコル
  - 原型としてSPFプロトコル
- IPマルチキャストを利用
- IPの上に直接実装されたプロトコル。
  - プロトコル番号89番
- DR（指名ルータ），BDR（バックアップ指名ルータ）を利用したマルチプルアクセスネットワークにおけるやり取りの簡素化

OSPF設定の一例（ciscoの場合）

- router ospf <process ID>
  - 自分のASと同じ番号にすることが多い
    - 一つのAS内で一つしかOSPF processを走らせない場合
    - process IDは1～65535の何番にしてもいい
- network 192.168.0.0 0.0.0.15 area 0
  - 0.0.0.15はワイルドカードマスク
    - アドレスのうち無視する部分をマスクする
  - 192.168.0.0～192.168.0.15の範囲にあるアドレスのインタフェースで
    - OSPFを area 0 で話す
    - そのインタフェースのネットワークをOSPFに広告する

上記2つが基本で、最低限のOSPFのconfig

OSPF設定の一例（ciscoの場合）



ルータAの設定

```
interface Ethernet0
 ip address 192.168.1.1  255.255.255.0
!
router ospf 1
 network 192.168.1.0  0.0.0.255  area  0
```

ルータBの設定

```
interface Ethernet0
 ip address 192.168.1.2  255.255.255.0
!
interface Ethernet1
 ip address 192.168.2.10  255.255.255.0
!
router ospf 1
 network 192.168.1.0  0.0.0.255  area  0
 network 192.168.2.0  0.0.0.255  area  0
```

ルータCの設定

```
interface Ethernet0
 ip address 192.168.1.3  255.255.255.0
!
interface Ethernet1
 ip address 192.168.2.11  255.255.255.0
!
router ospf 1
 network 192.168.1.0  0.0.0.255  area  0
 network 192.168.2.0  0.0.0.255  area  0
```

OSPFプロセスを確立する

ISPバックボーンネットワークにおける経路制御設計～ 理論編 ～

OSPF – Open Shortest Path First

france telecom

パケット種別

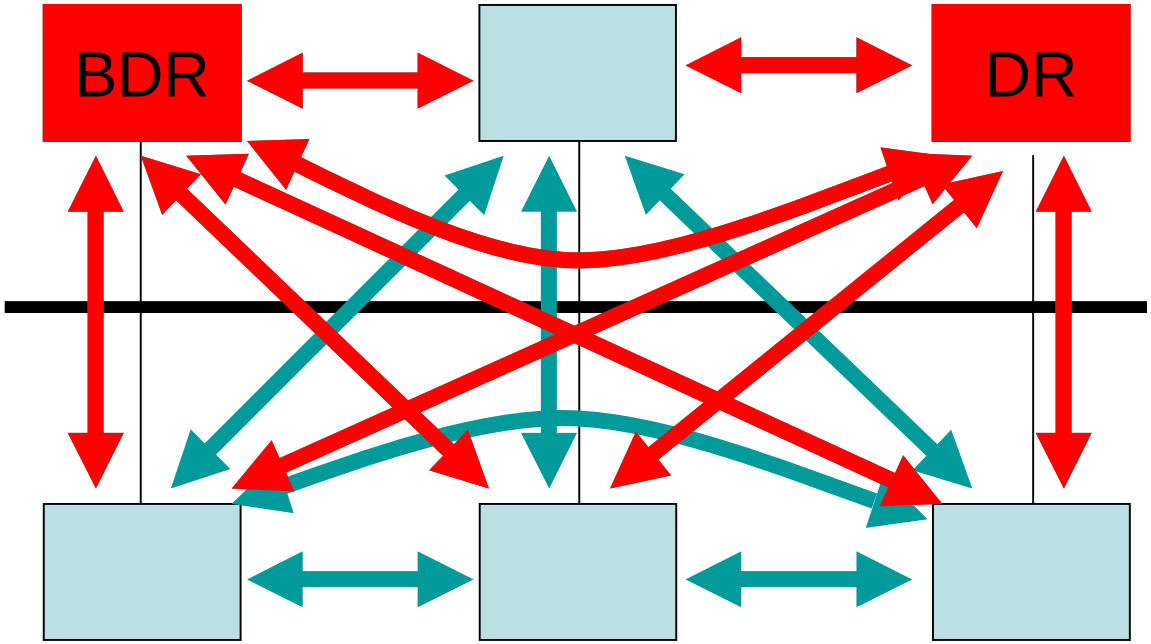
	パケットタイプ	内容
1	Hello	ネイバー確立と維持
2	Database Description	アジャセンシー確立時のDB内容の伝達
3	Link-State Request	リンクステート情報の要求
4	Link-State Update	リンクステート情報の伝達・更新
5	Link-State Acknowledgement	Link-State Updateの受領確認

アジャセンシー，DR，BDR

- アジャセンシー(adjacency:隣接) ==リンクステートのやり取りを行う関係
- マルチプルアクセスネットワークでは指名ルータ(designated router: DR)，バックアップDR(BDR）が代表して、他のルータとのリンクステートのやり取りを行う。

ネイバー
Helloのやりとり

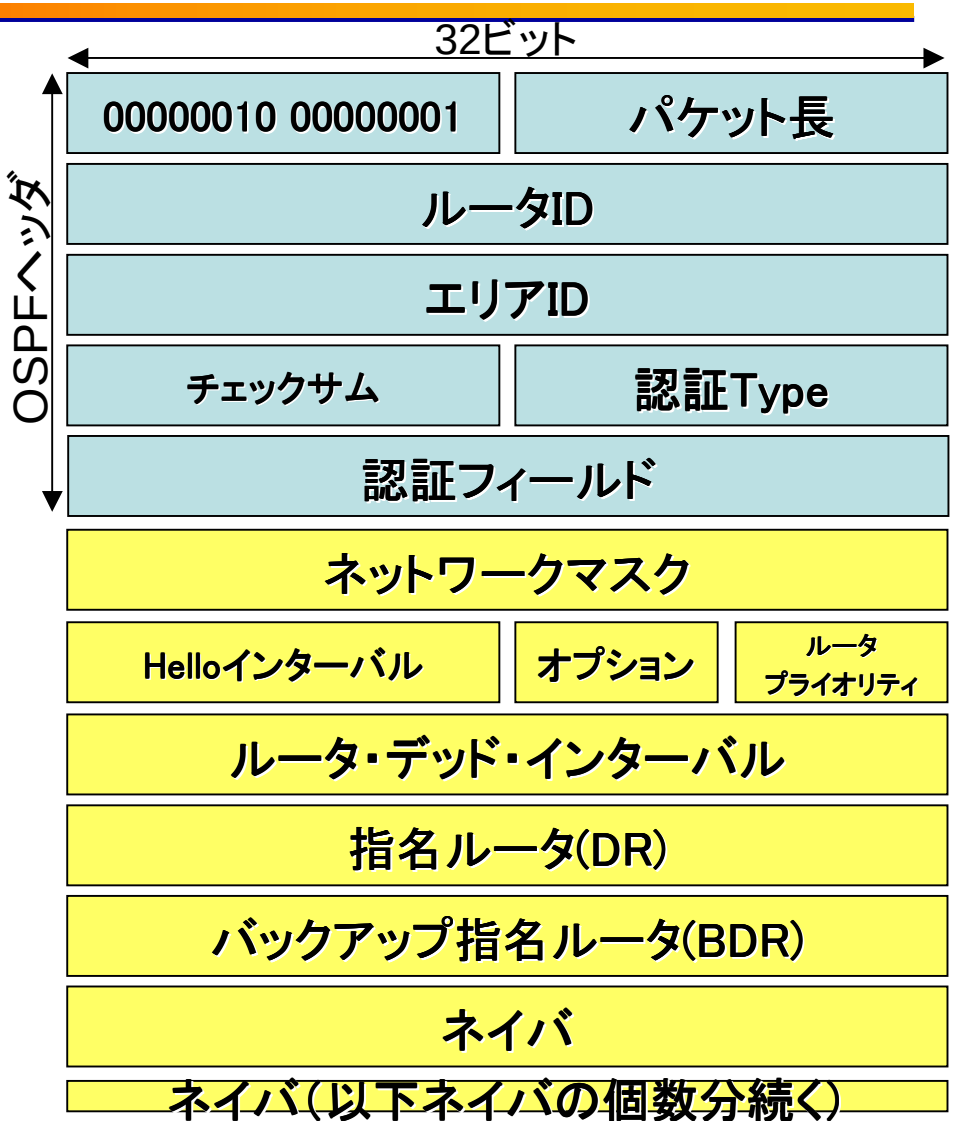
ネイバーであり
アジャセンシー
リンクステートのやりとり



france telecom

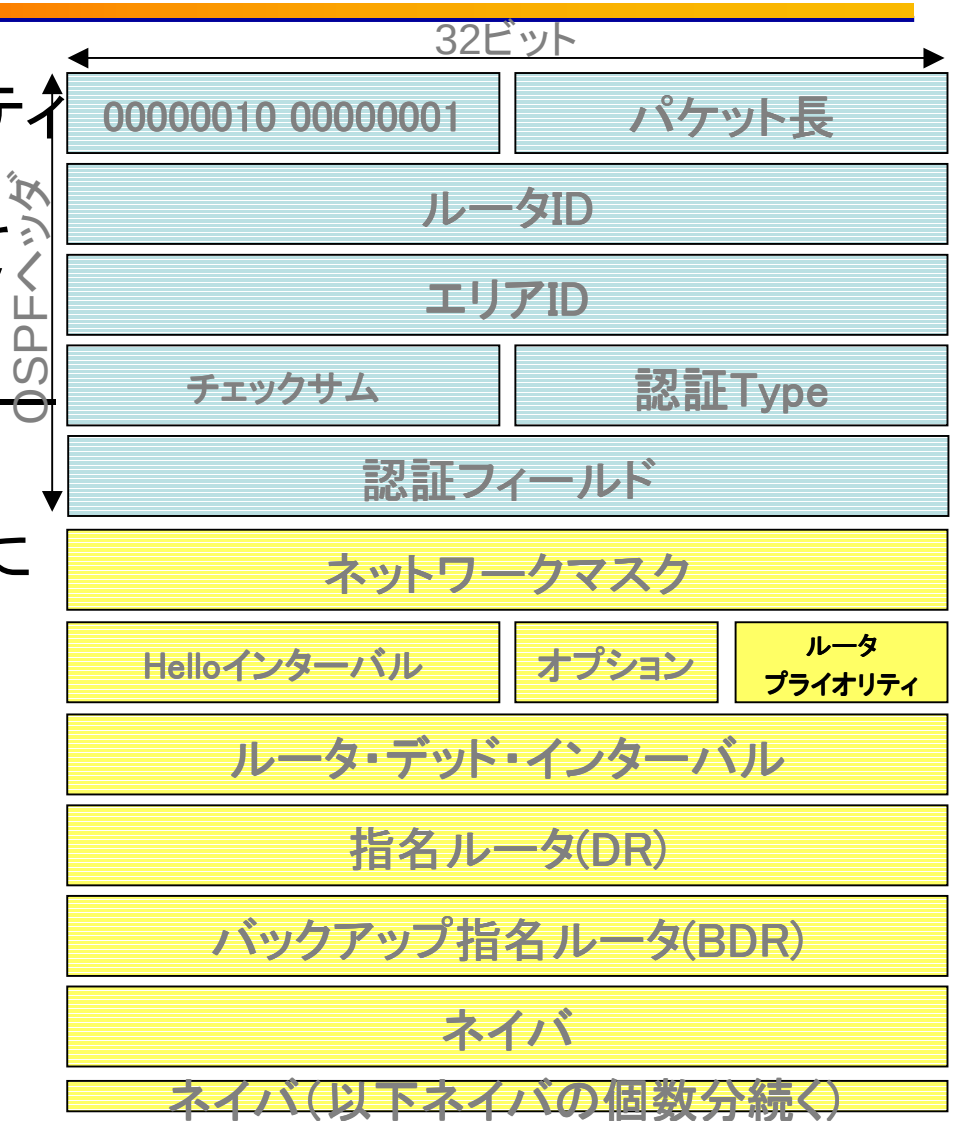
HELLOパケット

- マルチキャストでネットワークセグメント上に発信される
  - Ackなどは返らない
- ルータIDで自分を名乗り、それまでに知っているDR,BDR,ネイバを示す
  - ネイバ同士は、相手からのHELLOパケットのネイバフィールドに自分のルータIDがあることで、ステータスを認識する
- DR, BDRはルータプライオリティの比較，なければルータIDの大きいほう。但し後発によって置き換わらない



ルータプライオリティ

- DR, BDRはルータプライオリティの比較，なければルータIDの大きいほう。但し後発によって置き換わらない
- Ip ospf priority をインターフェースで定義（ciscoの場合）
- Priority 0 で、決してDR,BDRにならない
  - 高負荷なルータに設定すると有効



アジャセンシー確立までの道

- HELLOパケットでアジャセンシーであることが決定
- Database Descriptionパケットでリンクステートデータベースの内容比較を行う
- 相手が持っていて自分が持っていない（もしくは古い）リンクステートに関して、リンクステート広告を要求（リンクステートリクエスト）
- 全リンクステートの交換完了
- 完了

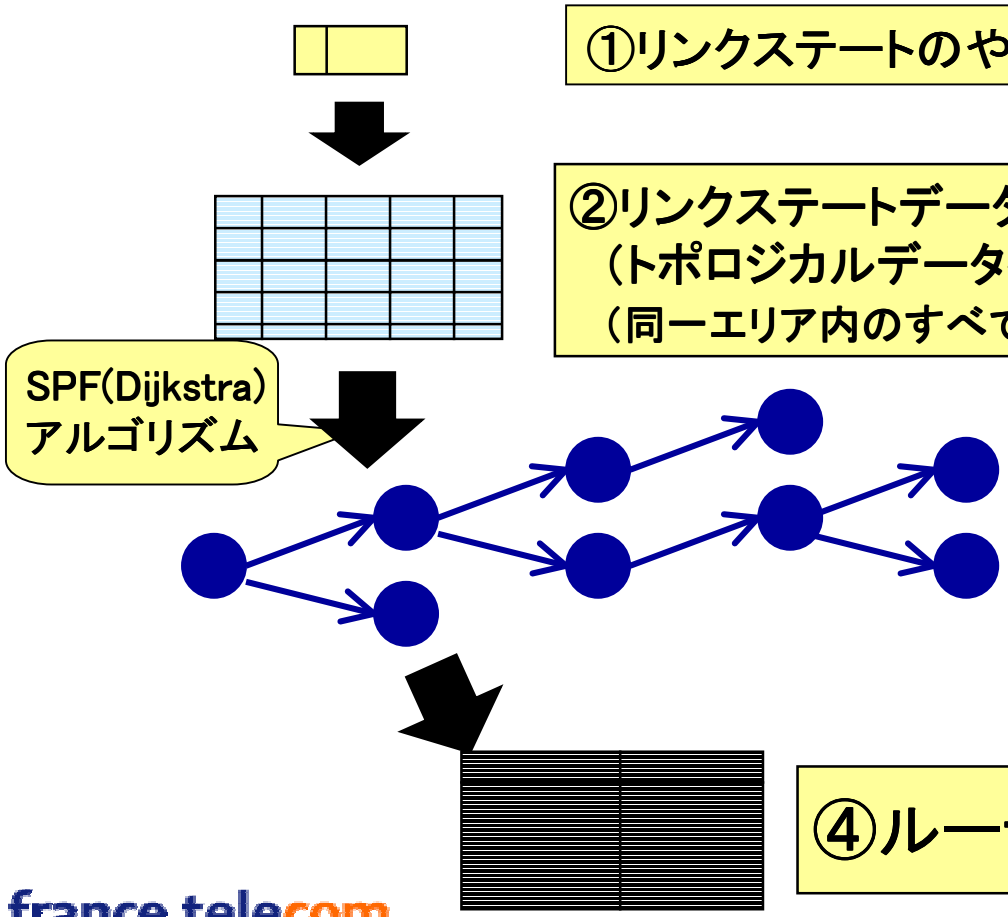
リンクステートデータベース

ISPバックボーンネットワークにおける経路制御設計～ 理論編 ～

OSPF – Open Shortest Path First

france telecom

OSPFにおける
ルーティングテーブルの生成

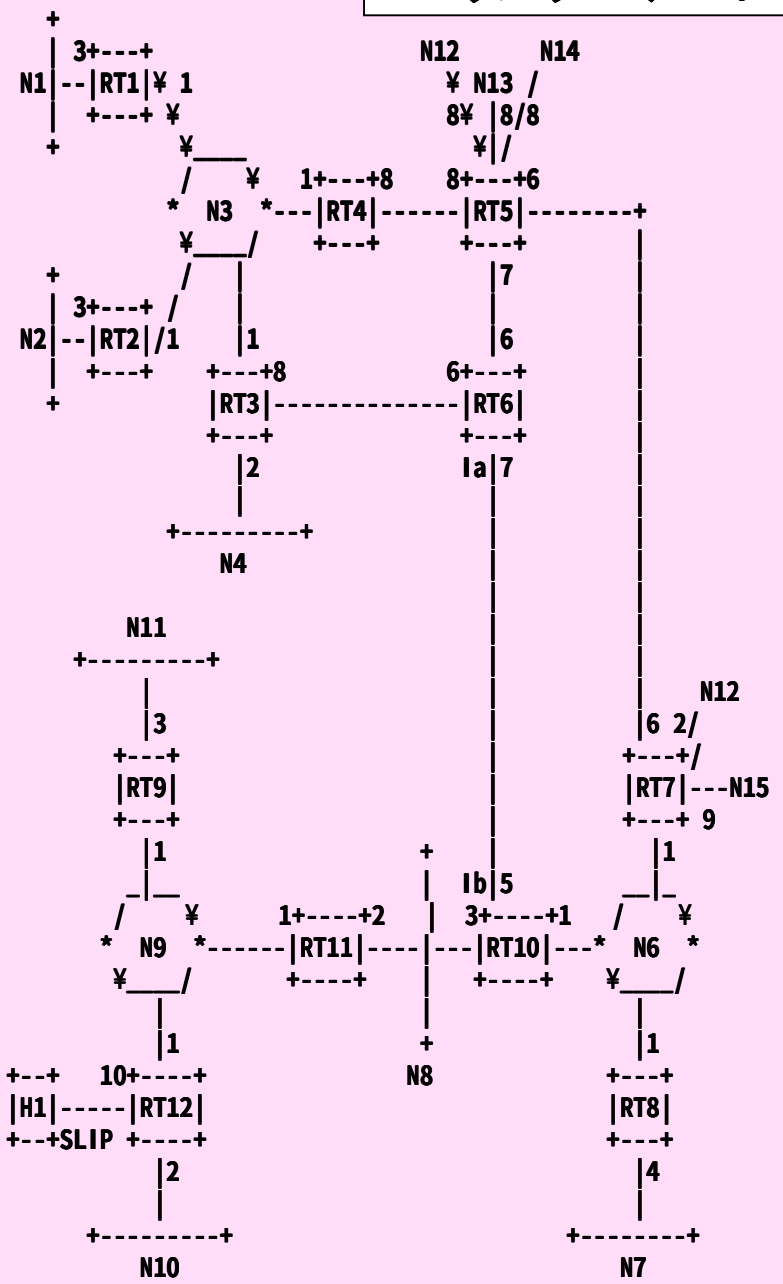


france telecom

リンクステートデータベース

- ルータとネットワークで構成される有向グラフ
  - ルータがネットワークにインタフェースを持っているときは、ルータとネットワークをつなぐ
  - ２つのルータが物理的にpoint-to-pointで結ばれているときは、ルータ同士をつなぐ
    - Numberedな/30を持つ場合、双方のルータにぶら下がるネットワークとして解釈される.
  - データベースの中はコストを値とする
    - ルータから見てネットワークに対してコスト値がつく
    - ネットワークからルータに向かうところは常にコスト0

リンクステートデータベースの例



A sample Autonomous System

		FROM															
		RT1	RT2	RT3	RT4	RT5	RT6	RT7	RT8	RT9	RT10	RT11	RT12	N3	N6	N8	N9
	RT1													0			
	RT2													0			
	RT3						6							0			
	RT4					8								0			
	RT5				8		6	6									
	RT6			8		7					5						
	RT7					6									0		
*	RT8														0		
*	RT9																0
T	RT10						7								0	0	
O	RT11															0	0
*	RT12																0
*	N1	3															
	N2		3														
	N3	1	1	1	1												
	N4			2													
	N6							1	1		1						
	N7								4								
	N8										3	2					
	N9									1		1	1				
	N10												2				
	N11									3							
	N12					8		2									
	N13					8											
	N14					8											
	N15							9									
	H1												10				
	Ia										5						
	Ib						7										

The resulting directed graph

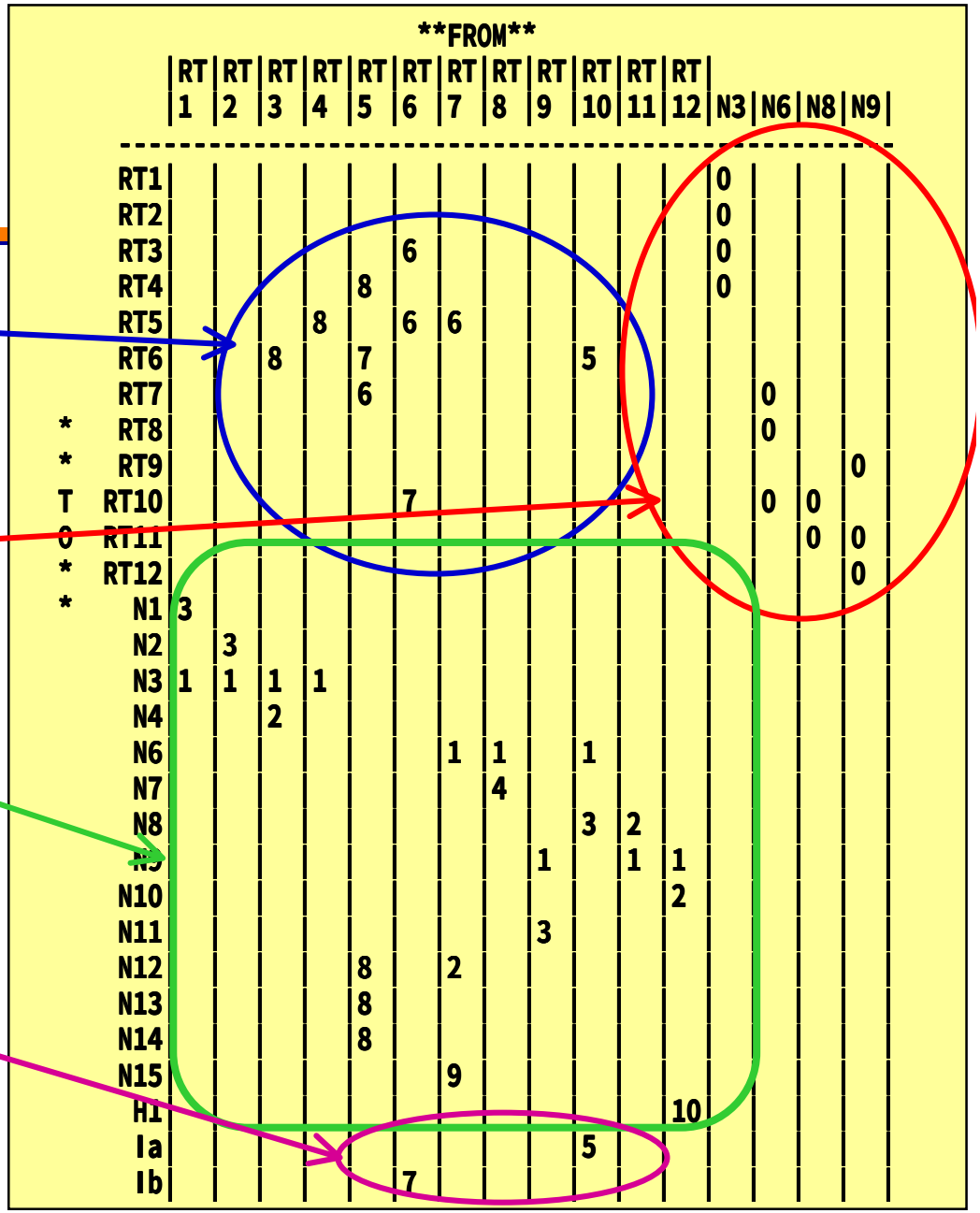
LSDBの内容

p-to-pはRT同士の辺となる

・FROMでNWがあるところは複数のルータがあるマルチアクセスネットワークとなる
・NWからRTに向かうのは常に0

・RTからNやHに向かうのはそのルータがそのネットワークにインタフェースを持つことを意味する
・値はそのインタフェースでのコストを示す

・RTからIに向かうのは、P2Pにおいて対抗のルータのインタフェースにアドレスが割り当てられていることを示す

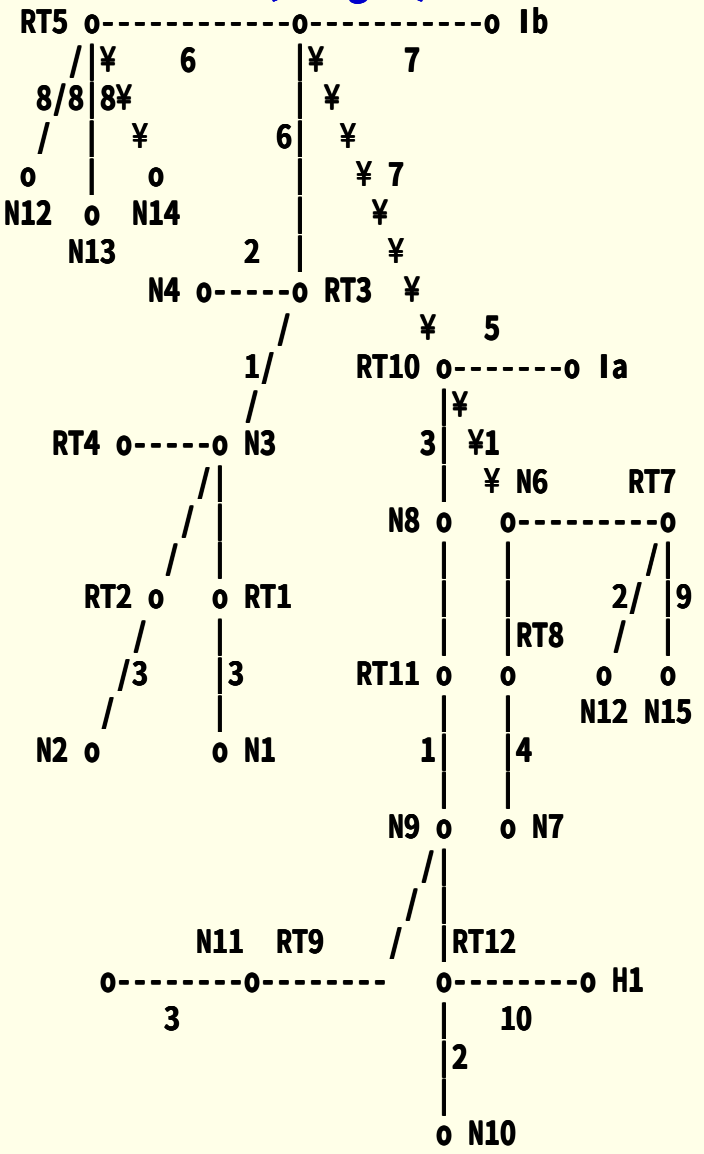


france telecom

最短パスツリー

		FROM															
		RT 1	RT 2	RT 3	RT 4	RT 5	RT 6	RT 7	RT 8	RT 9	RT 10	RT 11	RT 12	N3	N6	N8	N9
	RT1													0			
	RT2													0			
	RT3						6							0			
	RT4					8								0			
	RT5				8		6	6									
	RT6			8		7					5						
	RT7					6									0		
*	RT8														0		
*	RT9																0
T	RT10						7								0	0	
O	RT11															0	0
*	RT12																0
*	N1	3															
	N2		3														
	N3	1	1	1	1												
	N4			2													
	N6							1	1		1						
	N7								4								
	N8										3	2					
	N9									1		1	1				
	N10												2				
	N11									3							
	N12					8		2									
	N13					8											
	N14					8											
	N15							9									
	H1												1				
	Ia										5						
	Ib						7										

SPF(Dijkstra)
アルゴリズム



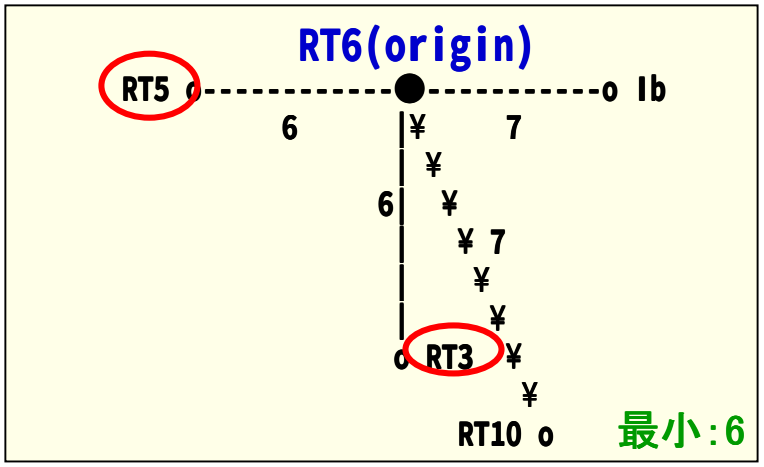
SPF(Dijkstra)アルゴリズム（１）

すべての中で最小のものを確定していき、次はそこから次のノードまでを加えていく

		FROM																
		RT1	RT2	RT3	RT4	RT5	RT6	RT7	RT8	RT9	RT10	RT11	RT12	N3	N6	N8	N9
	RT1													0			
	RT2													0			
	RT3						6							0			
	RT4					8								0			
	RT5				8		6	6									
	RT6			8		7					5						
	RT7					6									0		
*	RT8														0		
*	RT9																0
T	RT10						7								0	0	
O	RT11															0	0
*	RT12																0
*	N1	3															
	N2		3														
	N3	1	1	1	1												
	N4			2				1	1		1						
	N6								4								
	N7										3	2					
	N8									1		1	1				
	N9												2				
	N10									3							
	N11					8		2									
	N12					8											
	N13					8											
	N14							9									
	N15												10				
	H1										5						
	Ia						7										
	Ib																

データベースを見て、RT6から次のノードまでのツリーを作る

１回目



：確定

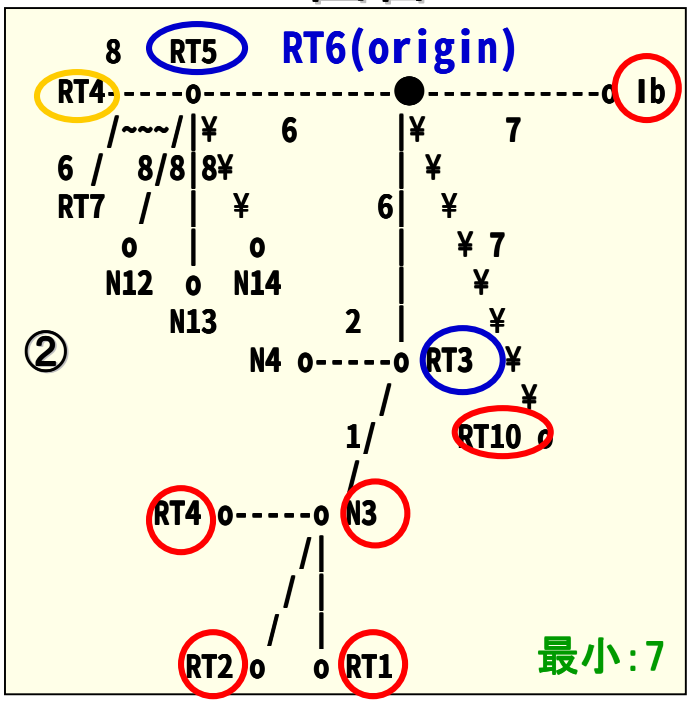
現在リーフにあるノードの中でRT6からのコストが最小である6のノードを確定する

SPF(Dijkstra)アルゴリズム(2)

確定したところからDBを見て次のノードまで伸ばす(RT6などの既に確定しているノードは除く)

		FROM															
		RT1	RT2	RT3	RT4	RT5	RT6	RT7	RT8	RT9	RT10	RT11	RT12	N3	N6	N8	N9
	RT1													0			
	RT2													0			
	RT3						6							0			
	RT4					8								0			
	RT5				8		6	6									
	RT6			8		7					5						
	RT7					6									0		
*	RT8														0		
*	RT9																0
T	RT10						7								0	0	
O	RT11															0	0
*	RT12																0
*	N1	3															
	N2		3														
	N3	1	1	1	1												
	N4			2													
	N6							1	1		1						
	N7								4								
	N8										3	2					
	N9									1		1	1				
	N10												2				
	N11									3							
	N12					8		2									
	N13					8											
	N14					8											
	N15							9									
	H1												10				
	Ia										5						
	Ib						7										

2回目



:旧確定 :新確定 :消去

現在リーフにあるノードの中でRT6からのコストが最小である7のノードを確定する

RT4はRT6→RT3→N3→RT4で確定したのでRT5→RT4のところは消去する

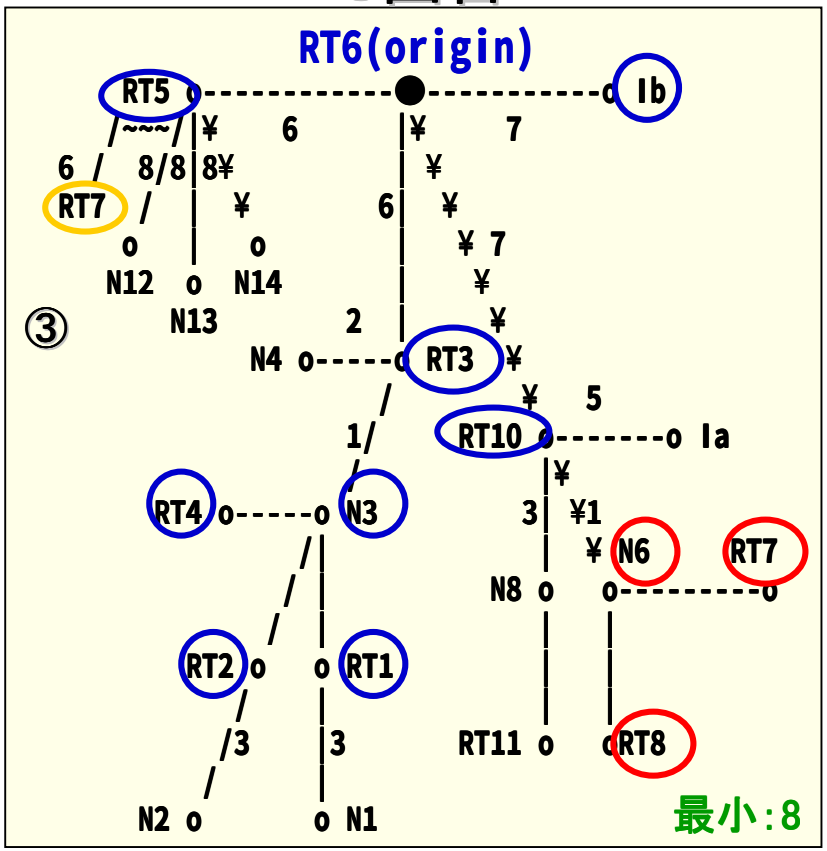
SPF(Dijkstra)アルゴリズム（３）

確定したところからDBを見て次のノードまで伸ばす.これを繰り返す.

FROM

		RT1	RT2	RT3	RT4	RT5	RT6	RT7	RT8	RT9	RT10	RT11	RT12	N3	N6	N8	N9
	RT1													0			
	RT2													0			
	RT3						6							0			
	RT4				8									0			
	RT5			8			6	6									
	RT6			8		7					5						
	RT7					6									0		
*	RT8														0		
*	RT9																0
T	RT10						7								0	0	
O	RT11															0	0
*	RT12																0
*	N1	3															
	N2		3														
	N3	1	1	1	1												
	N4			2													
	N6							1	1		1						
	N7								4								
	N8										3	2					
	N9									1		1	1				
	N10												2				
	N11									3							
	N12				8		2										
	N13				8												
	N14				8												
	N15							9									
	H1												10				
	Ia										5						
	Ib						7										

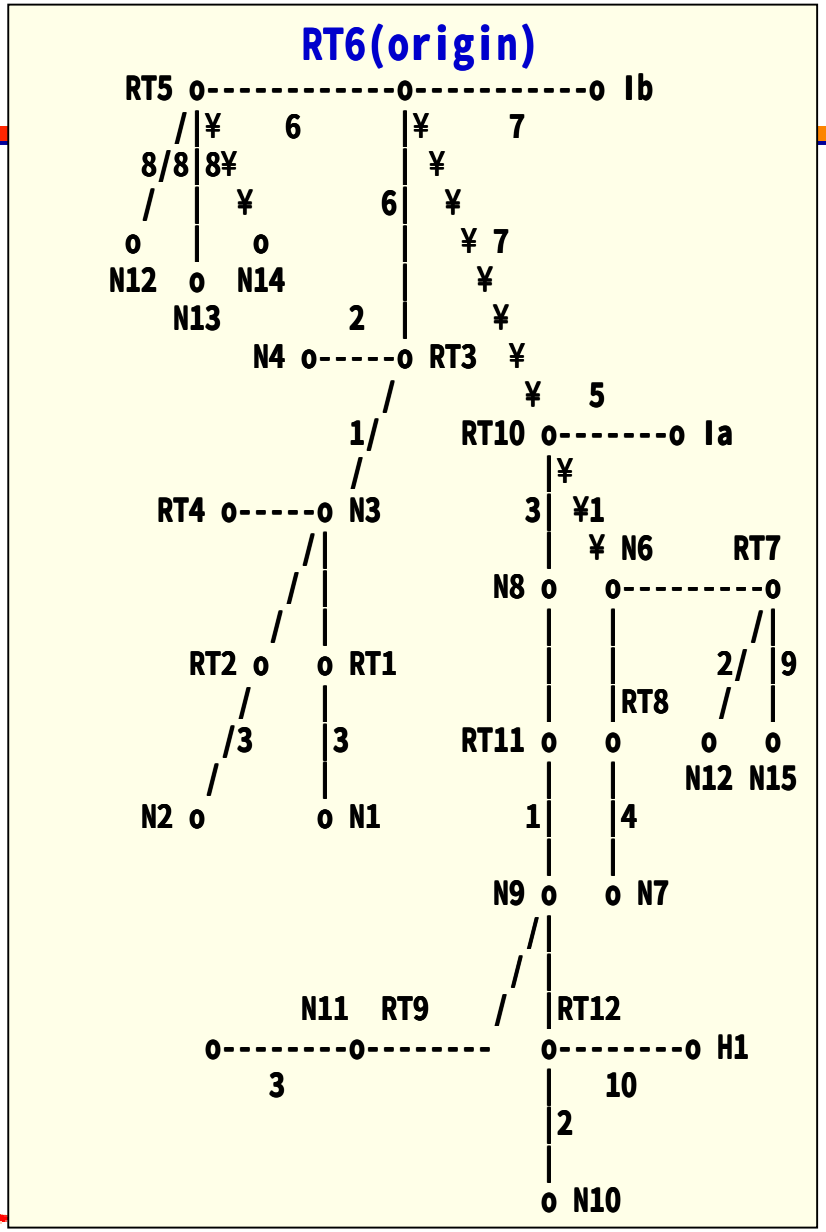
３回目



：旧確定　：新確定　：消去

RT7はRT6→RT10→N6→RT7で確定したので
RT5→RT7のところは消去する

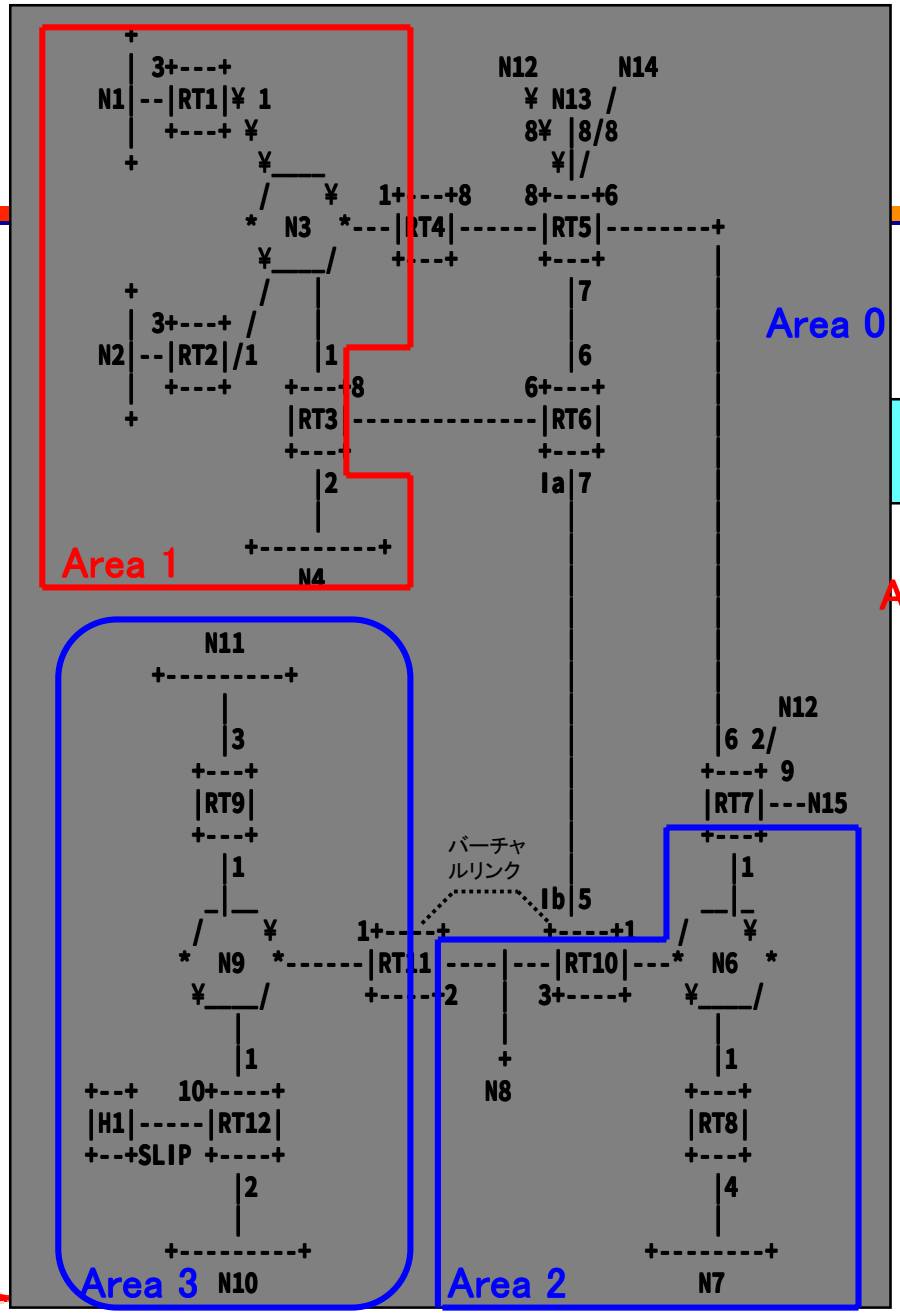
ルーティングテーブルの作成



Destination	Next Hop	Distance
N1	RT3	10
N2	RT3	10
N3	RT3	7
N4	RT3	8
Ib	*	7
Ia	RT10	12
N6	RT10	8
N7	RT10	12
N8	RT10	10
N9	RT10	11
N10	RT10	13
N11	RT10	14
H1	RT10	21
RT5	RT5	6
RT7	RT10	8

The portion of Router RT6's routing table listing local destinations.

エリアで分けられている場合



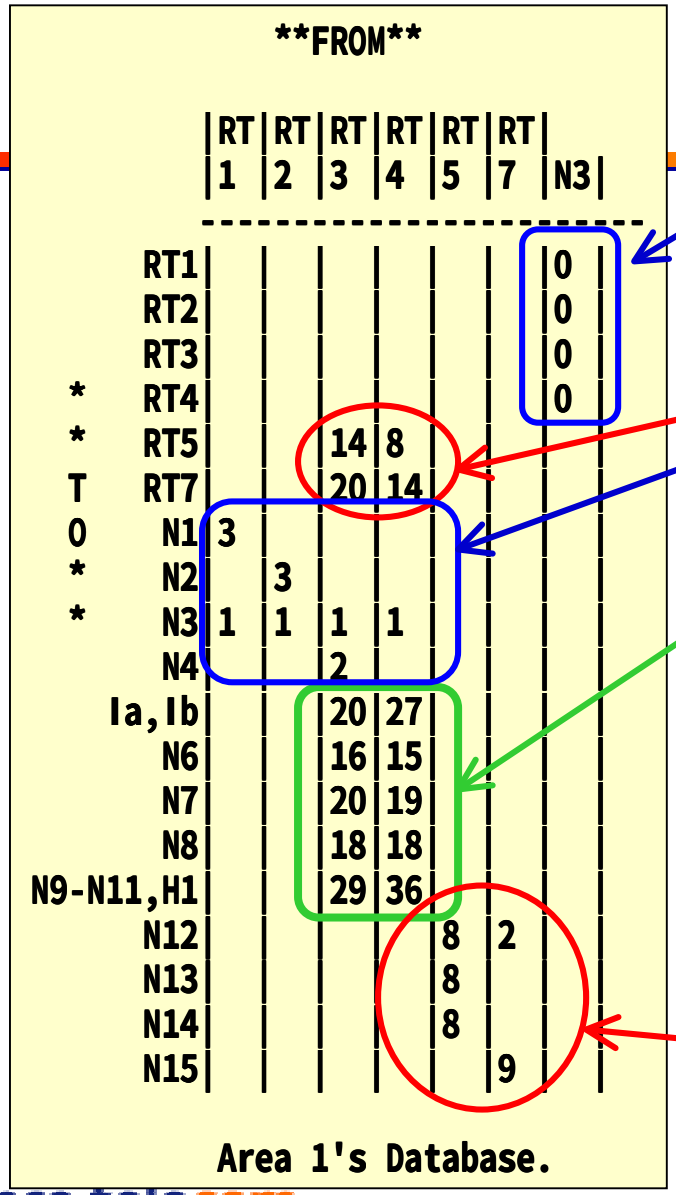
Area 1

FROM

		RT 1	RT 2	RT 3	RT 4	RT 5	RT 7	N3
	RT1							0
	RT2							0
	RT3							0
*	RT4							0
*	RT5			14	8			
T	RT7			20	14			
0	N1	3						
*	N2		3					
*	N3	1	1	1	1			
	N4			2				
	Ia,Ib			20	27			
	N6			16	15			
	N7			20	19			
	N8			18	18			
	N9-N11,H1			29	36			
	N12					8	2	
	N13					8		
	N14					8		
	N15						9	

Area 1's Database.

リンクステートデータベースの内容



FROM

		RT1	RT2	RT3	RT4	RT5	RT7	N3
	RT1							0
	RT2							0
	RT3							0
*	RT4							0
*	RT5			14	8			
T	RT7			20	14			
0	N1	3						
*	N2		3					
*	N3	1	1	1	1			
	N4			2				
	Ia,Ib			20	27			
	N6			16	15			
	N7			20	19			
	N8			18	18			
	N9-N11,H1			29	36			
	N12					8	2	
	N13					8		
	N14					8		
	N15						9	

Area 1's Database.

LS Type2: ネットワークLSA

エリア境界ルータから
AS境界ルータまで
LS Type4: サマリ LSA

LS Type1:
ルータLSA

エリア境界ルータからエリア外の
ネットワーク(inter-area)まで
LS Type3: サマリ LSA

これは以下の情報からわかる。
・エリア境界ルータ(RT3、RT4)から全部のエリア境界ルータ(RT7、RT10など)までのコストがバックボーンエリアのSPF treeから計算される。
・各エリアのエリア境界ルータからバックボーンにサマリ情報を流している。
これがつまりエリア境界ルータがバックボーンに属していなければならない理由でもある。

AS境界ルータから
AS externalなネットワークまで
Type5: AS external LSA

リンクステート広告（LSA)

ISPバックボーンネットワークにおける経路制御設計～ 理論編 ～

OSPF – Open Shortest Path First

france telecom

リンクステート広告(LSA)

- リンクステート広告(LSA: Link State Advertisement)
  - OSPFにおいて経路情報生成のために交換される本質的情報
  - Link-State Update(OSPFパケットタイプ#4)パケットの中に複数収容される

LS type	LSAの名前
1	ルータLSA
2	ネットワークLSA
3,4	サマリLSA
5	AS-external LSA

ルータLSA（LSAタイプ1）

- 全てのルータで生成する
- ルータの接続情報
  - そのルータにどういうリンクがついているか、それぞれのリンクの種類*とリンクの情報(Link ID, Link DATA)とコストを情報としてもつ
- エリア内しか伝わらない
- これにより、エリア内の各ルータが各ネットワークにどのように接続されているかが分かる

*参考：ルータLSAの中で表すリンクのType

Link Type	Description
1	他のルータとp-to-p接続**
2	透過ネットワーク***への接続
3	stubネットワークへの接続
4	virtual link

ネットワークLSA（LSAタイプ2）

- DRがそのネットワークを代表して生成、広告する
- このネットワークに接続されるルータのリスト

LSAタイプ3～5

- サマリLSA(LSAタイプ3,4)
  - エリア境界ルータによって生成される
  - エリア外の情報
  - Type 3 はエリア外のネットワークの情報
  - Type 4 はAS境界ルータの情報（AS外部のネットワークについてはType5）
- AS external LSA(LSAタイプ5)
  - AS境界ルータによって生成される
  - 他のプロトコルからredistribute(再分配)された経路

AS境界とスタブエリア

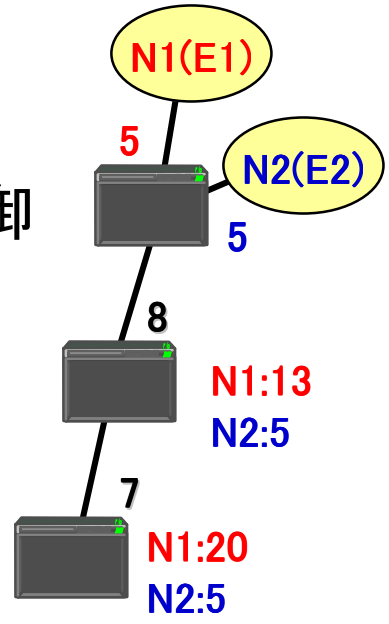
- AS境界
  - OSPFの議論でAS境界と言う場合、「他のプロトコル（connected, static含む）が接するところ」という意味で使う.
- スタブ(stub)エリア
  - AS境界を一切含まないエリア
  - そのスタブエリアであることを明示的に定義することで、ABRでデフォルトルートを生成してエリア内のルーティングを実現し、エリア内のOSPF負荷を軽減することが出来る

NSSA – Not So Stubby Area

- RFC1587 ”The OSPF NSSA Option”
- Not So Stubby –　そんなにスタブっぽくない – 準スタブエリア
- Type 7 LSAを使う
- NSSAではType7 LSAを流すことができる
- NSSAのAS境界ルータでAS外部経路をType7 LSAとしてredistributeする
  - Type 7 LSAsはNSSAのASBRでしか生成されず、NSSA内にしか流れない
  - 他のエリアに対しては、ABRでType 7 LSAsをType 5 LSAsに変更する。そのときサマライズやフィルターすることもできる。

外部経路(External Routes)の取り扱い

- 扱い方（同一プリフィクスの選択方法）に2つのタイプ
  - Type1: redistributeされたメトリック＋OSPFコスト
    - Redistribute時のメトリックを一定にすれば、最も近いASBRに経路制御される
  - Type2: redistributeされたコストのまま
    - Redistribute時のメトリックで優先制御を決定付ける
  - 同一プリフィクスに対して以下のように優先制御
    - IntraArea>InterArea>External 1>External 2



OSPFの方路選択と耐障害性

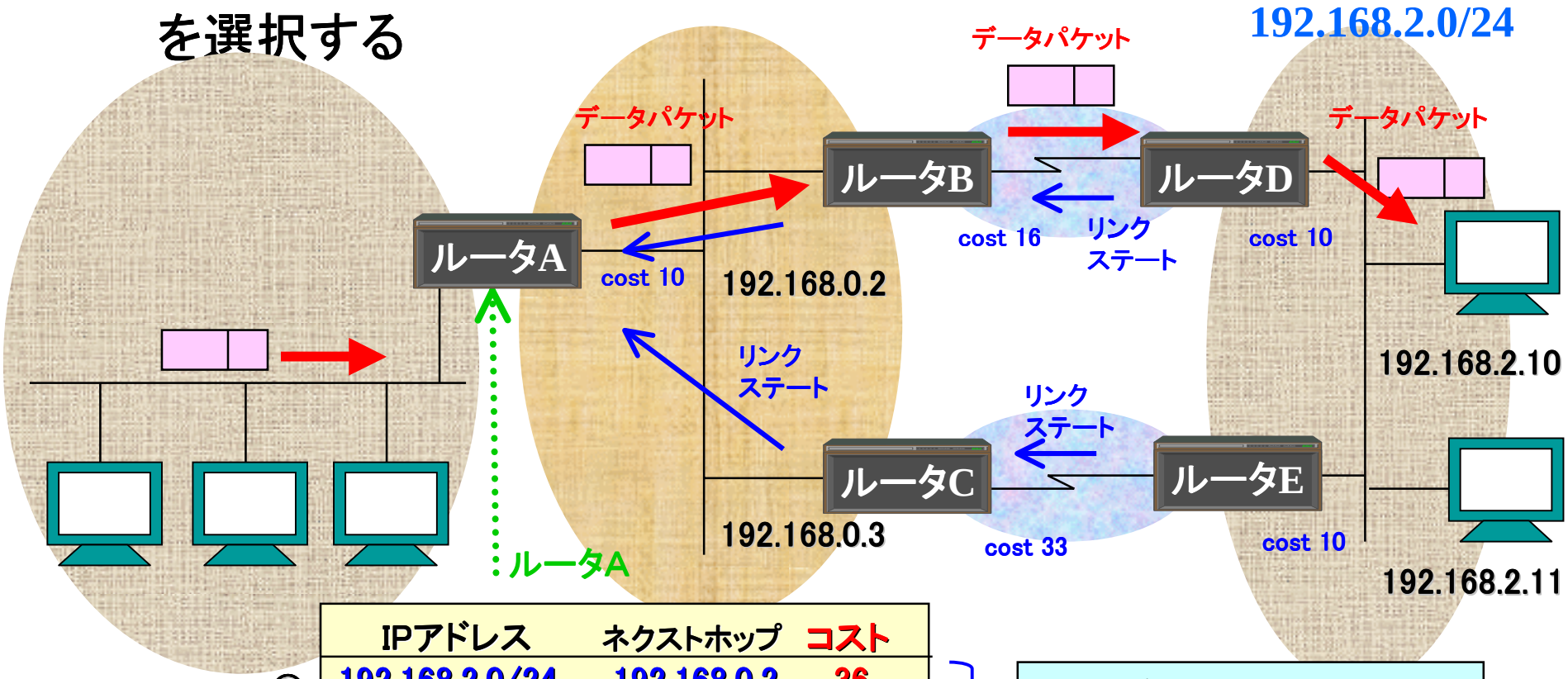
ISPバックボーンネットワークにおける経路制御設計～ 理論編 ～

OSPF – Open Shortest Path First

france telecom

コスト

- 同じネットワークが複数見える場合、コストが一番低い経路を選択する



IPアドレス	ネクストホップ	コスト
192.168.2.0/24	192.168.0.2	36
192.168.2.0/24	192.168.0.3	53

◎

コストが一番低い経路を選択

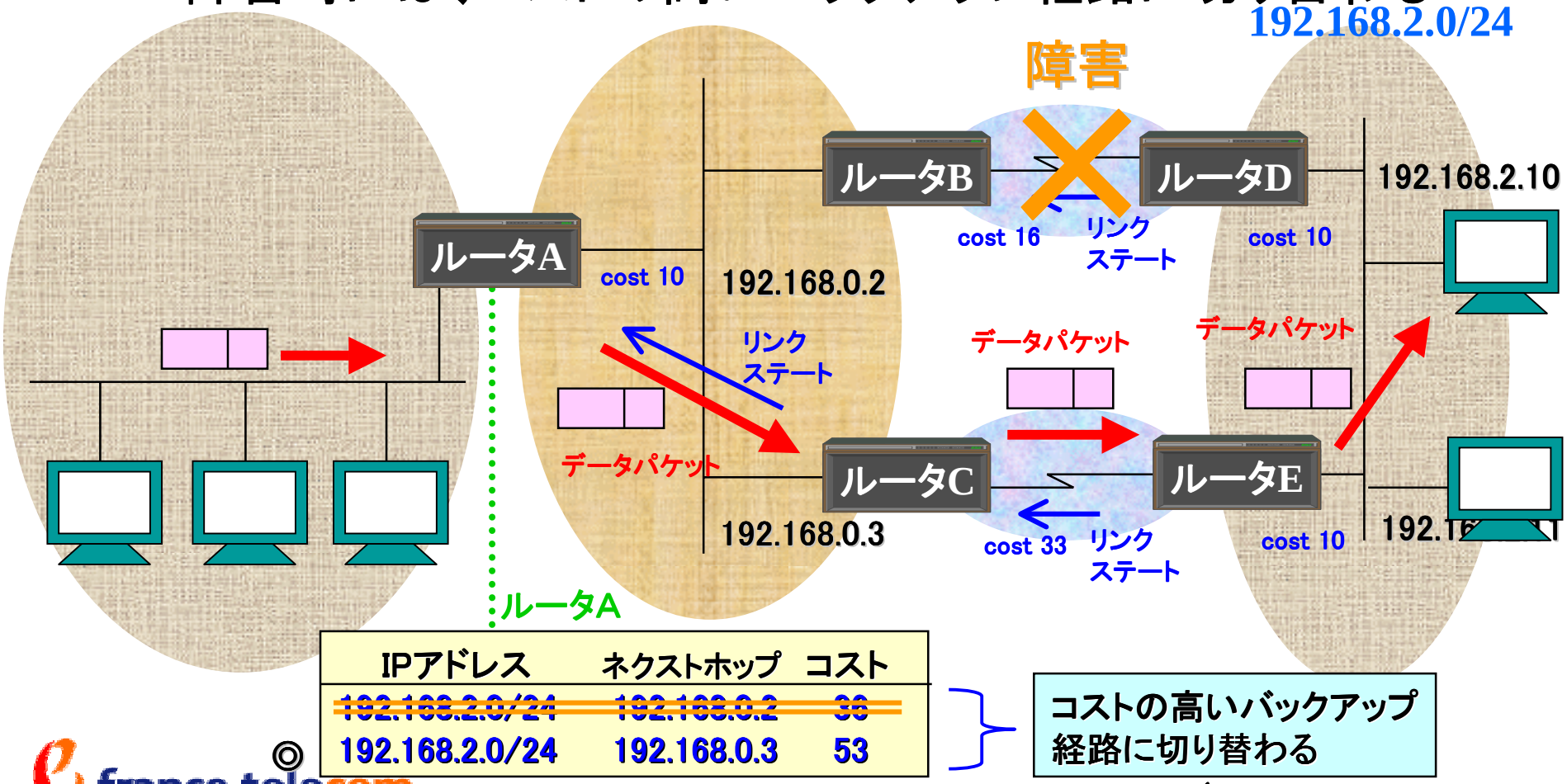
※正確にはルーティングテーブル（フォワーディングテーブル）にのるのがそれだけになる

コストの割り当てと計算

- コスト値の範囲:1 ～ 65535
- インターフェースに割り当てる
  - ip ospf cost <cost値>
  - 非対称でもよい
    - そのインタフェースからデータパケットが出るときのためのコスト
- 自動割り当ても可能
  - デフォルト 100M/回線速度(bps)

障害時にはバックアップ経路に切り替わる

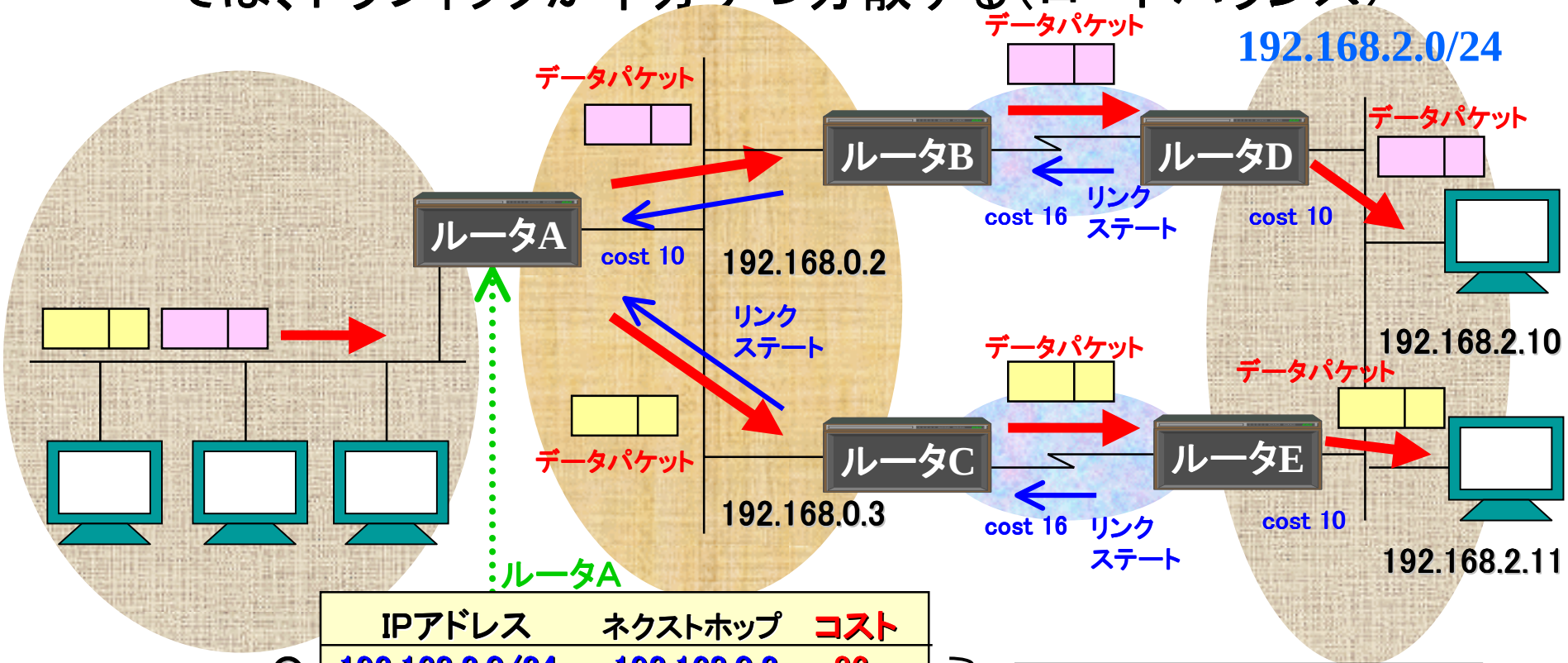
- 障害時には、コストの高いバックアップ経路に切り替わる



※正確にはコストの低い経路がルーティングテーブルから消え、コストの高い経路が現れる

ロードバランス –
イコールコストマルチパス

- 同じネットワークが同じコストで見えるネットワークに対しては、トラフィックが半分ずつ分散する（ロードバランス）



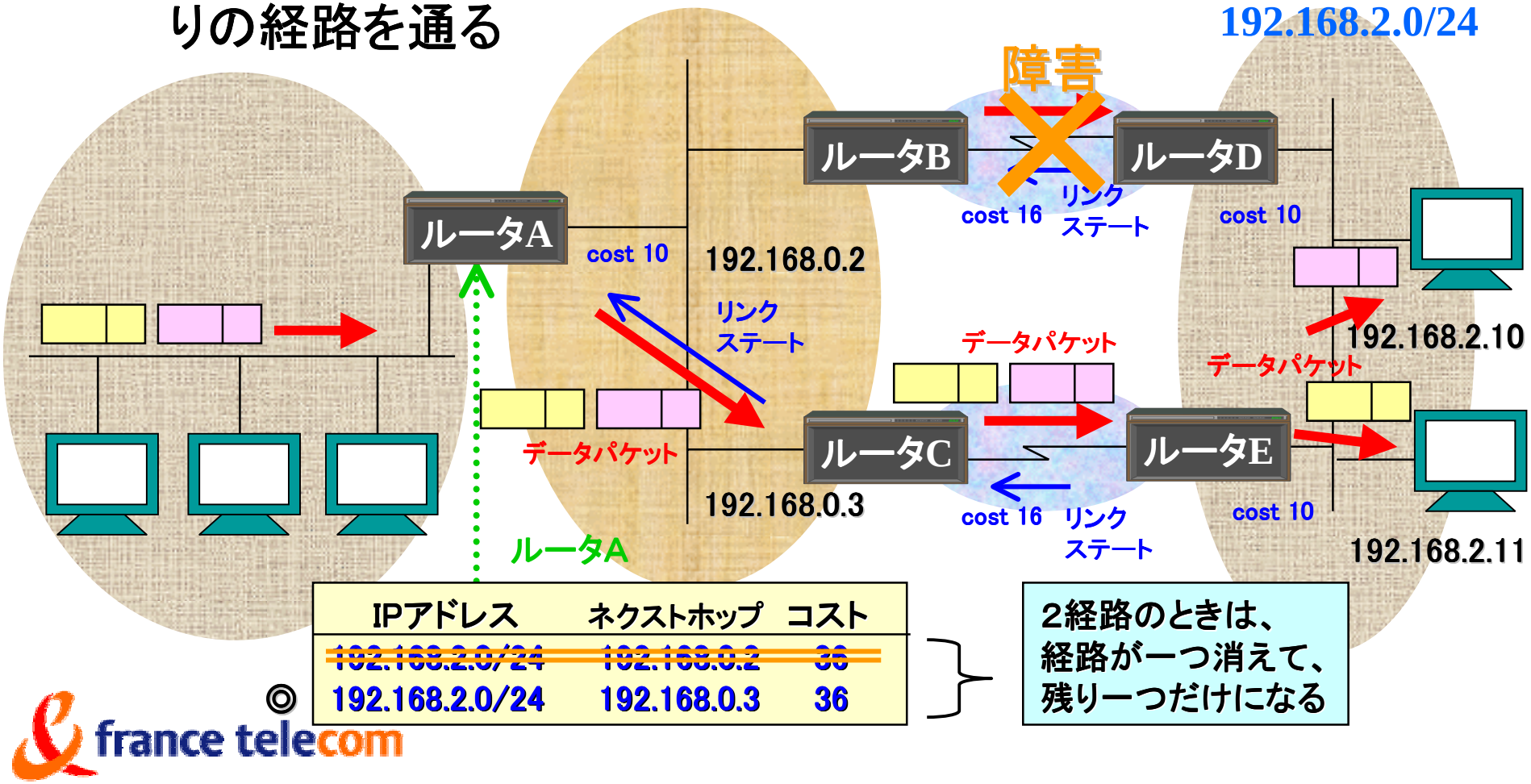
	IPアドレス	ネクストホップ	コスト
○	192.168.2.0/24	192.168.0.2	36
○	192.168.2.0/24	192.168.0.3	36

等コストでロードバランスする

france telecom

イコールコストマルチパスの障害時の挙動

- 障害時には、障害した経路が消えて全てのトラヒックが残りの経路を通る



イコールコストマルチパスの取り扱い

- OSPFでは同コストの複数方路を同時にルーティングテーブルに乗せられる機能を提供する
- パケットフォワーディングを実現するのはルータの実装
  - Ciscoでは最大6
    - maximum-paths 6　（rotuer ospf セクション）

基本的な適用技術

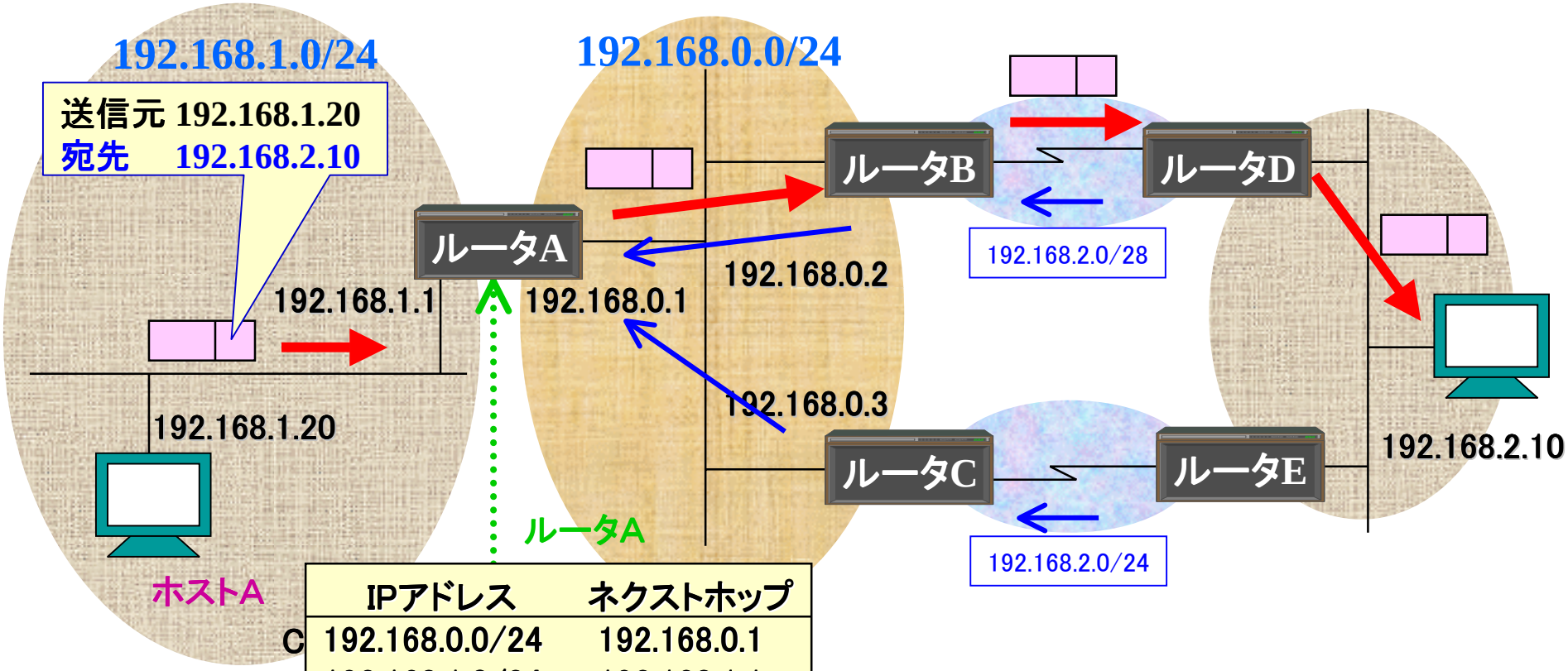
ISPバックボーンネットワークにおける経路制御設計～ 理論編 ～

OSPF – Open Shortest Path First

france telecom

最長一致法アルゴリズム (longest-match)

- IPパケットの宛先アドレスを調べて、一致するネットワークアドレスが複数ある場合には、ビット列が長い方のネットワークアドレスを選択する

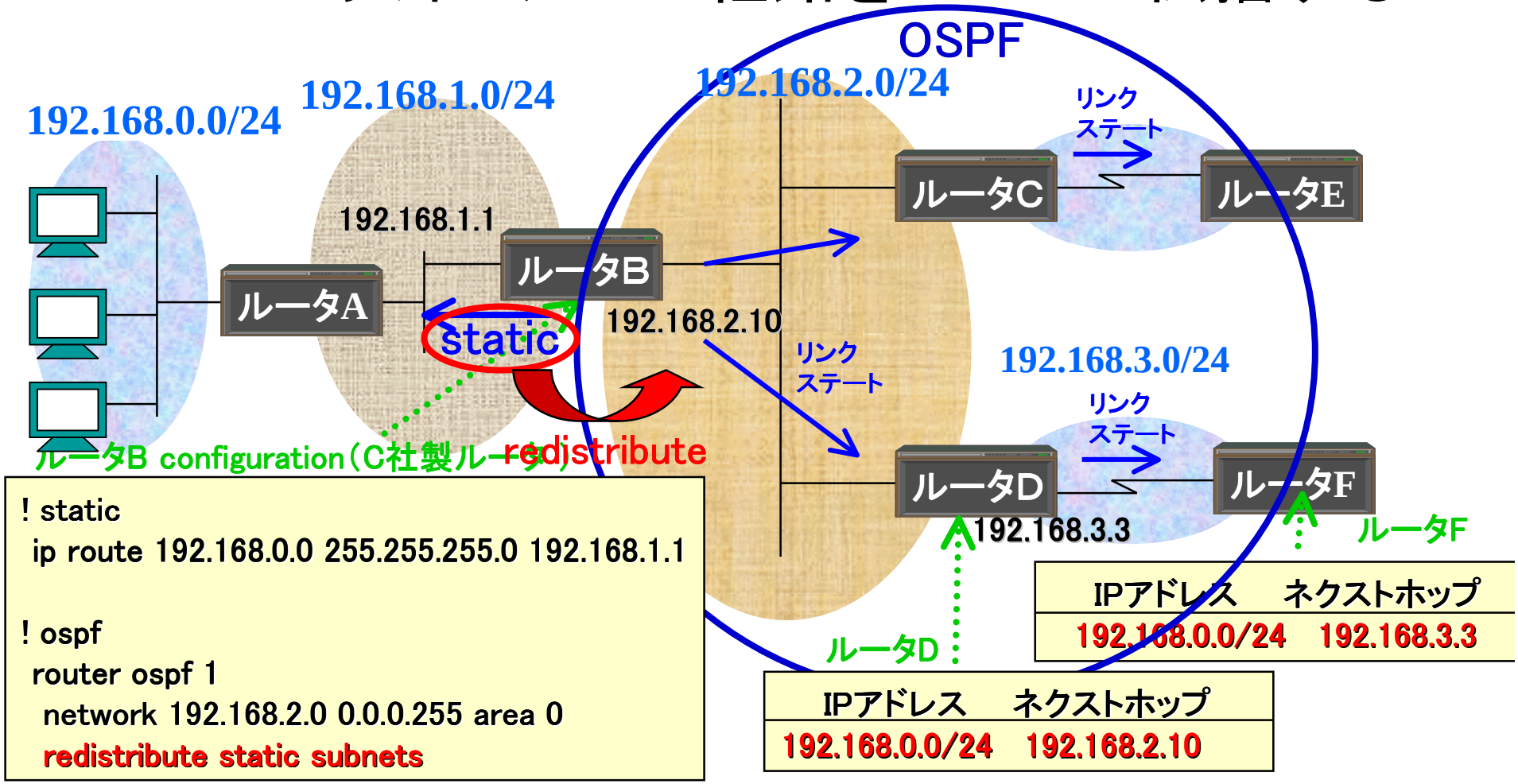


IPアドレス	ネクストホップ
192.168.0.0/24	192.168.0.1
192.168.1.0/24	192.168.1.1
192.168.2.0/28	**192.168.0.2**
192.168.2.0/24	192.168.0.3

Longest matchで192.168.0.2を選択する

Redistribution – 再分配

- OSPF以外で知った経路をOSPFで伝播する



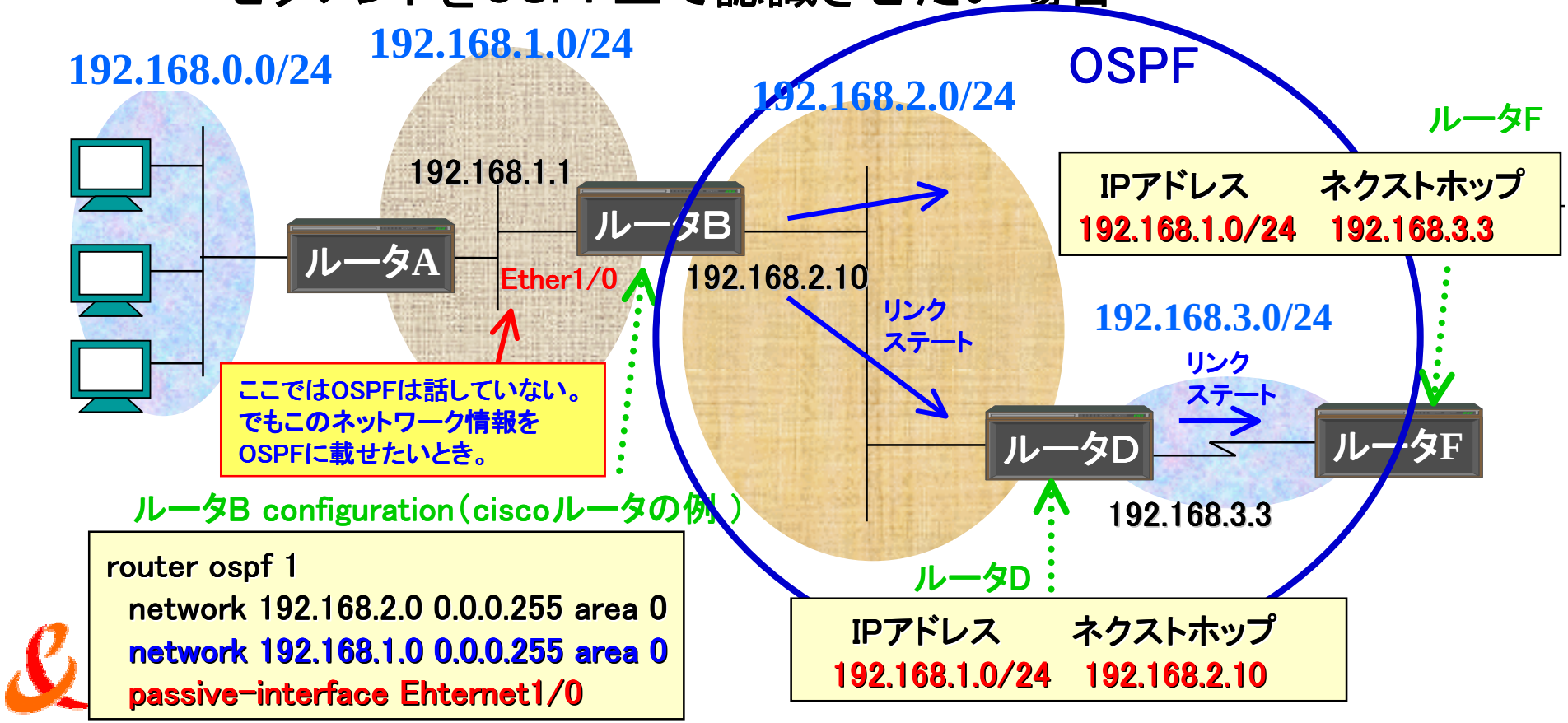
Redistribution – 再分配

- 最も一般的な例 – スタティック経路をOSPFで認識させる
  - 特にデフォルトルート
    - Default-information originate (ciscoのコマンド)
- コネクテッド経路をredistributeする例もある
  - Passive-interfaceでも実現可能
- 他のルーティングプロトコルからのredistribution

- OSPF的には「外部経路(external routes)」となり、redistributeを設定するルータはAS境界ルータとなる

Passive-Interface

- passive-interface Ethernet1/0
  - そのインタフェース上にOSPFのネイバはいないが、そのセグメントをOSPF上で認識させたい場合



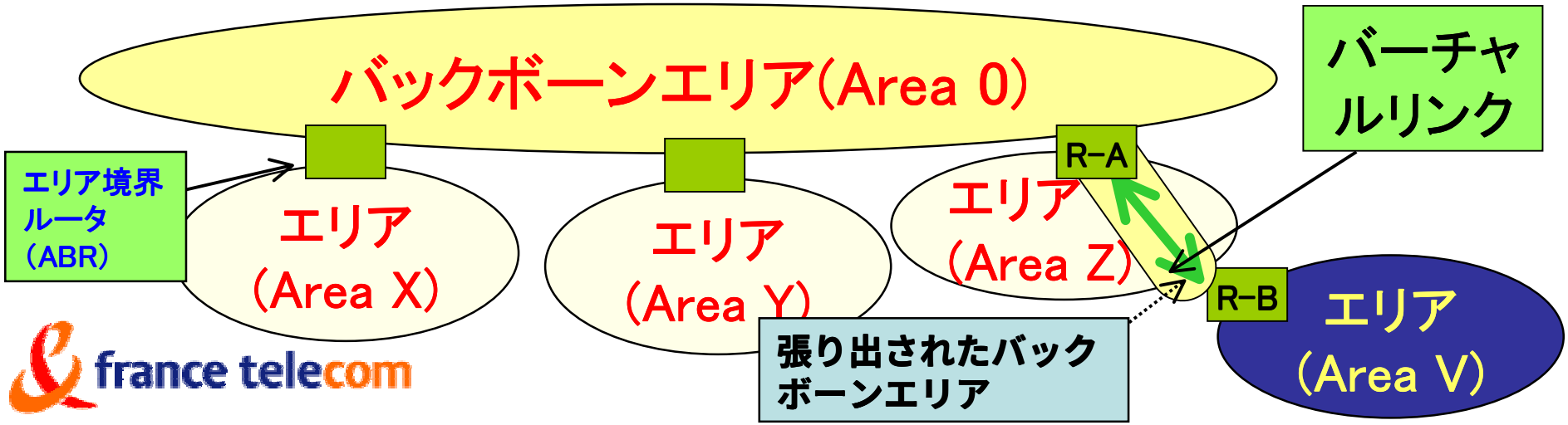
エリア設計

- 基本 – 可能な限りシンプルに : 最初からエリア分けを考えない. まずはエリア0のシングルエリアで考える.
- 信頼性を必要であれば、一つのエリアに複数のエリア境界ルータを設置する
- 一つのエリア境界ルータが所属するエリアはなるべく2つまでにすべき
  - つまりエリア0ともう一つのエリア、というようになる
- 経路の集約
  - エリア境界ルータにて経路の集約をする
  - エリアごとに経路を集約できるように、アドレス設計をする
    - area ** range <address> <mask>　　(エリア境界ルータ)
  - OSPFにredistributeされる経路も集約できるように、アドレス設計する
    - summary-address <address> <mask>　(AS境界ルータ)

france telecom

バーチャルリンク

- バックボーンに対して物理的コネクションを持たないエリアを取り込む
- Area Z virtual-link [R-B loopback addr]
- 設計が過度に複雑になるのを避けるため、定常状態での利用は避ける．障害時の緊急利用や移行作業の経過状態などで利用



ルータID

- ループバックインターフェースを定義して、IPアドレスをOSPFのルータID，iBGPのネイバアドレスとして利用する
  - 物理インターフェースをルータIDやiBGPネイバアドレスとすると、障害でそのIPアドレスが利用不可能となったときにルータIDを変更する必要があり、不安定要素となる.
  - ループバックインターフェースは仮想インターフェースで、決してダウンしない：ルータIDとして常に利用可能
  - /32で十分
  - その名の通りルータの識別子として利用可能.

IS-ISとの比較

Intermediate-System to Intermediate-System

ISPバックボーンネットワークにおける経路制御設計～ 理論編 ～

OSPF – Open Shortest Path First

france telecom

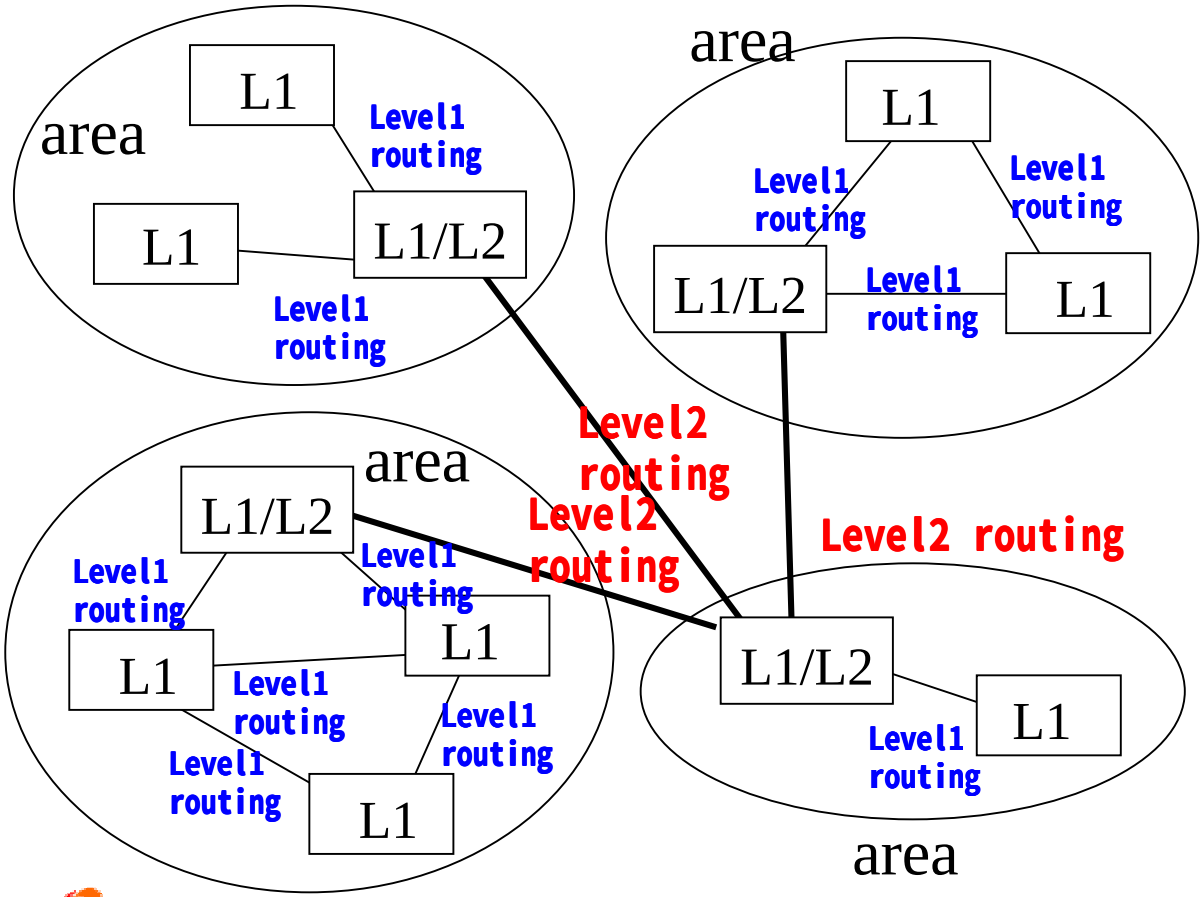
IS-IS
OSIのリンクステート型プロトコル

- OSPFと非常に類似したリンクステート型プロトコル
- もともとはOSI CLNS(connectionless network service)用のCLNP(CLN Protocol)の経路情報交換用で、それをIPも扱えるように拡張(Enhanced IS-IS, DUAL IS-IS)
- OSPFよりも前に実装されたため、IGPの規模対応を迫られた大手ISPを中心に利用。未だに多くの大手ISPが利用
  - 多くの大手ISPが利用するため安定性が高く、スケーラビリティも高そうだが、OSPFとアーキテクチャ的に大差ないと言われている。

IS-ISとOSPF・プロトコルの比較

	IS-IS	OSPF
プロトコルスタック	OSI	TCP/IP
情報伝達に利用するプロトコル	CLNP (Connectionless Network Protocol)	IP(インターネットプロトコル)
ノードID	NET (Network Entry Title, NSAPアドレス)	ルータID(IPアドレス)
階層化, 区画化	レベル1(エリア内) レベル2(エリア間)	エリアによる区画化, 上位エリアはエリア0
リンクステートの交換	LSP(リンクステートPDU)	LSA(リンクステート広告)
コストの値	1リンク63(6ビット)まで, 総和1,023(10ビット)まで	65,535(16ビット)まで

IS-ISネットワーク構成例



- areaの概念がある。ボーダーはリンク上となる（OSPFはルータ上）
- area内のroutingをlevel1，area間をlevel2 routingという
- level1 routingを話すルータをlevel1(L1)ルータという。Level1 ISともいう。
- leve2 routingを話すルータをlevel2(L2)ルータという。Level2 ISともいう。
- level1 routing、level2 routingどちらも話すルータをlevel1/level2(L1/L2)ルータという
- 多くのL2ルータは、自areaのためlevel1 routingを話すため、L1/L2ルータとなる

france telecom

Level1 / Level2 ルーティング

- SPFアルゴリズム
  - Level1とLevel2両方それぞれに関して独立に走る
- Level1 IS ルータにおいて
  - 他エリアへの通信に関しては、metric的に最も近いL1/L2ルータに向けてdefault routeを向けることによって通信が可能となる。
- 実際的には、Level2のみの起動の場合が多い

BGP
– Border Gateway Protocol

ISPバックボーンネットワークにおける経路制御設計～ 理論編 ～

france telecom

セクション目次

- BGP – Border Gateway Protocol
  - ASと階層的経路制御
  - BGPの特徴
  - パケットタイプとセッション確立
  - パス属性値 – Path Attributes
  - 単純なBGPシステムの導入
  - iBGPシステムの構成
  - iBGPシステムのスケーラビリティ
  - 現在の経路制御におけるBGPとOSPFの関係
  - ポリシルーティング

ASと階層的経路制御

ISPバックボーンネットワークにおける経路制御設計～ 理論編 ～

BGP – Border Gateway Protocol

france telecom

ASとAS番号

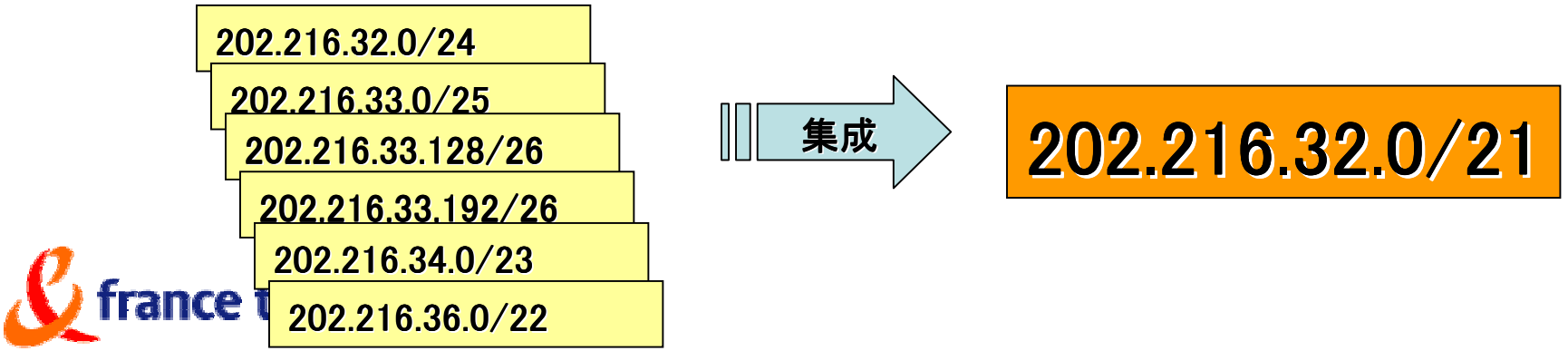
- AS – Autonomous System
  - 自律システム
  - 別名：ルーティングドメイン
  - 単一管理・単一ルーティングポリシで運用されるネットワーク領域 – 一般的には一つのISP
- AS番号 – ASN : Autonomous System Number
  - 16ビット(1～65535)の番号空間を持つ
    - AS5511==FT/*Opentransit Internet* , AS2914== *NTT/Verio*,
    - 64512～65534はプライベートAS, 65535はIANAリザーブ
    - 現在最大値は30000程度, 16,000個程度が観測される
  - The Internet とは、ASが相互接続された全体

The Internetにおける階層的経路制御(1)

- CIDR – Classless Inter-Domain Routing
  - クラスレスにAS間ルーティングを実施する
  - 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

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。

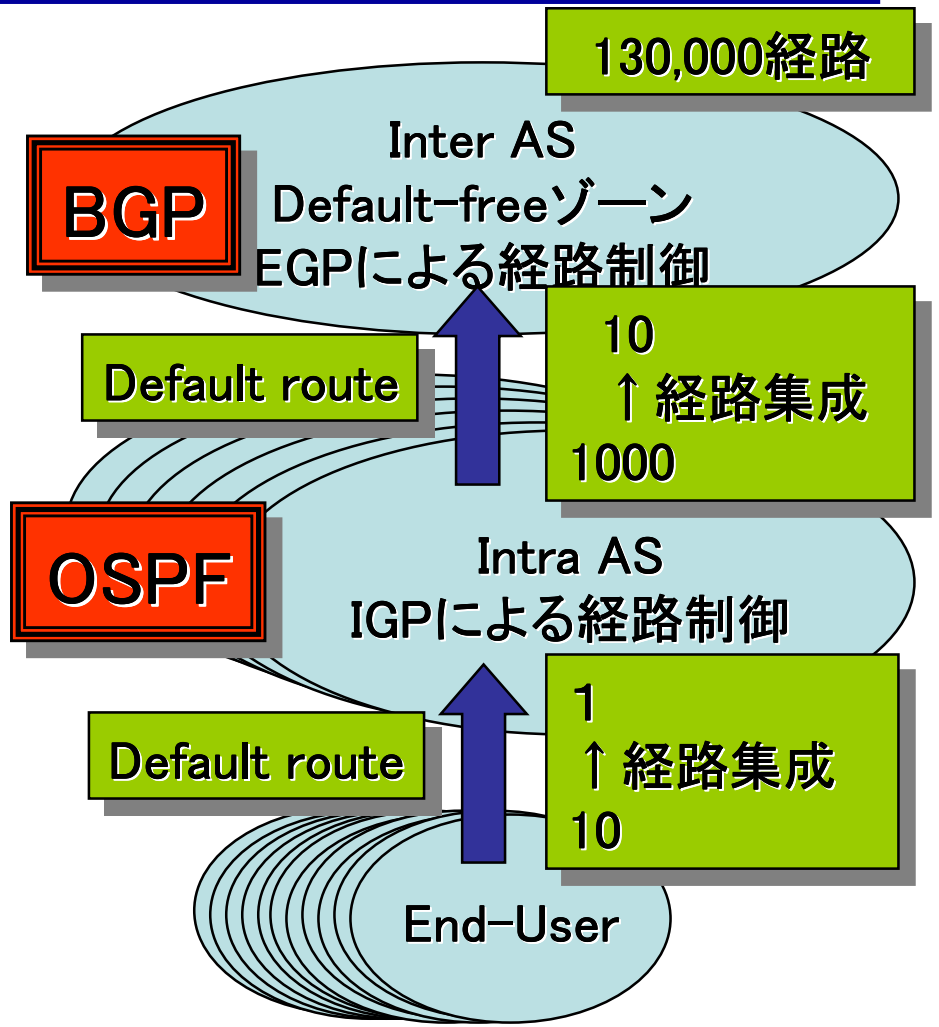


The Internetにおける階層的経路制御(3)

- 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

The Internetにおける階層的経路制御(4)

- 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



The Internetにおける階層的経路制御(5)

- 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
  - マルチホーム用の小さいブロックの割り当てが容認されつつある
    - 階層的経路制御の崩壊の兆し

BGPの特徴

ISPバックボーンネットワークにおける経路制御設計～ 理論編 ～

BGP – Border Gateway Protocol

france telecom

基本事項の確認

- 現在のバージョンは4 – BGP4, RFC1771
- AS間経路制御に用いる
- ピアリング(peering) – 明示的に定義した隣接ルータとの間にTCP上でセッションを確立し、経路情報を交換する
- パスベクター型プロトコルと呼ばれ、プリフィクス単位の経路情報レコードにはパス属性と呼ばれる属性情報が添えられている。
- パス属性により経路の優先制御を行う。
- パス属性を調整することで経路制御ポリシを実装することができる。

BGPとOSPFの比較(1)

OSPF	BGP
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

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	明示的に定義した隣接ルータのみと経路交換
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	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	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
30分でリフレッシュ--flooding	リフレッシュなし

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

BGPとOSPFの比較(4)

OSPF	BGP
IGP AS内の経路制御・トポロジ管理	EGP AS間の経路制御
各ルータは AS内の経路情報とネットワーク状態を交換し合っている	各ルータは 他のASの経路情報を交換し合って同期を取っている

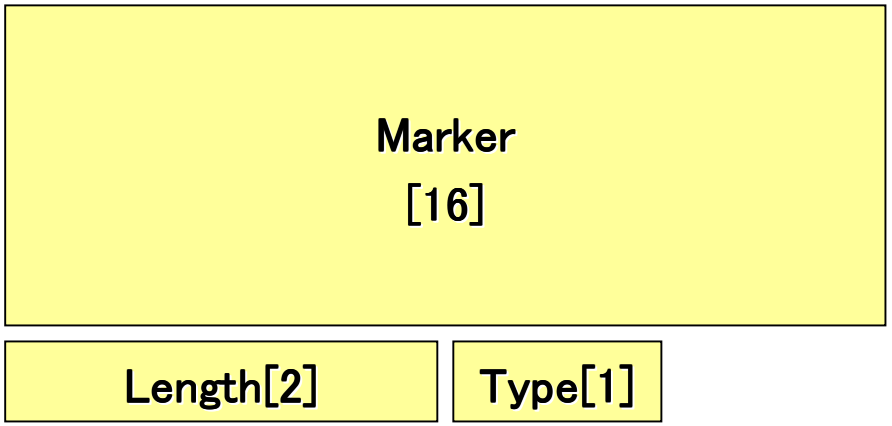
パケットタイプと
セッション確立

ISPバックボーンネットワークにおける経路制御設計～ 理論編 ～
BGP – Border Gateway Protocol

france telecom

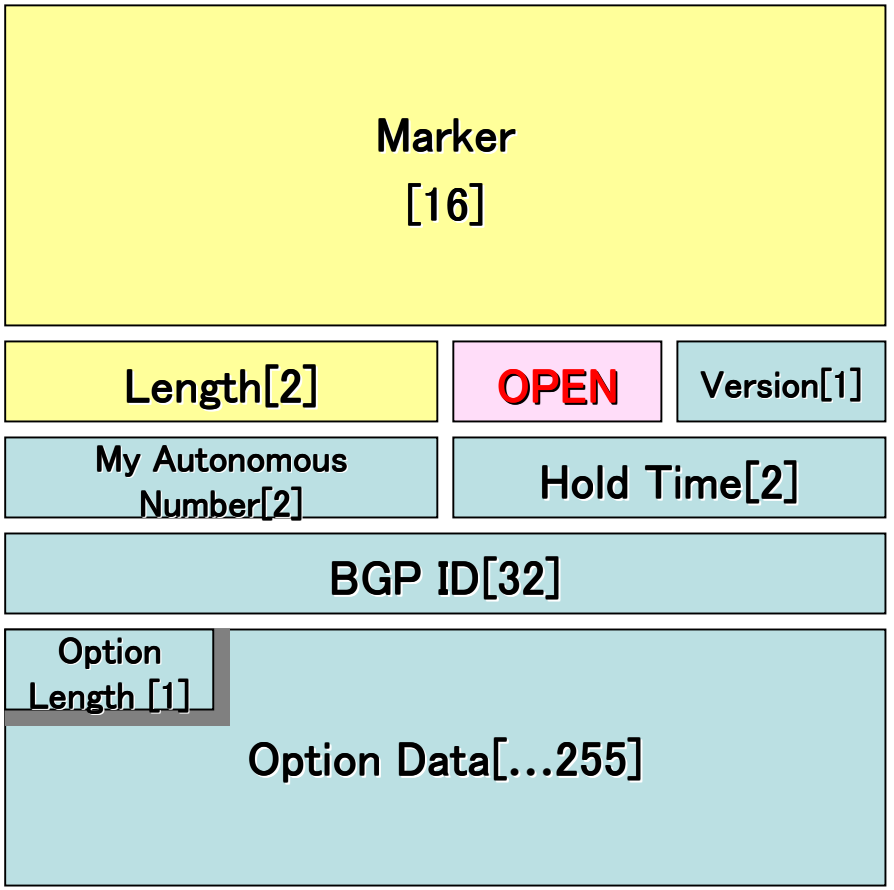
メッセージヘッダ

- マーカー(Marker)
  - セキュリティ目的に利用
- 長さ(length)
- タイプ(Type)
  - メッセージタイプ
    - オープン(OPEN)
    - 更新(UPDATE)
    - 通知(NOTIFICATION)
    - キープアライブ(KEEPALIVE)



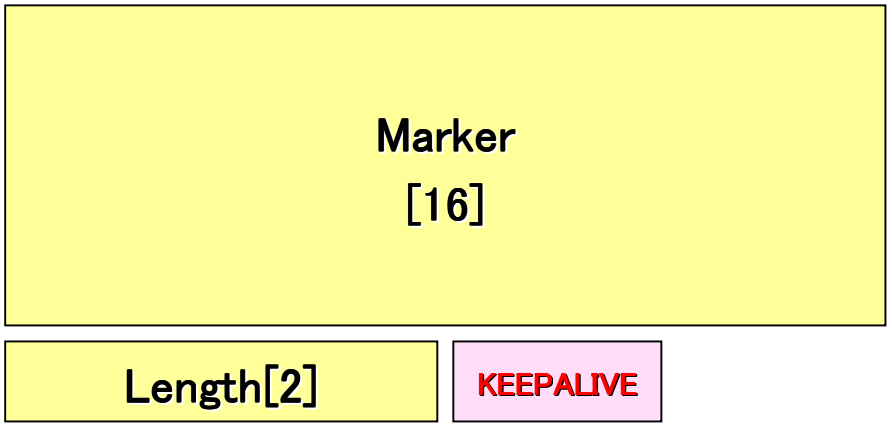
OPENメッセージ

- BGPセッション開設に利用
- 以下のパラメータ提示してネゴシエーションを実施
  - バージョン
    - 現在はバージョン4
  - 自AS番号
  - ホールドタイム
    - キープアライブの間隔を指定
  - BGP ID
    - 自分が持つIPアドレスからひとつをIDとして利用する
  - オプションは現在のところ認証情報のみが定義されている
- 受諾にはKEEPALIVE，拒否にはNOTIFICATIONを返す



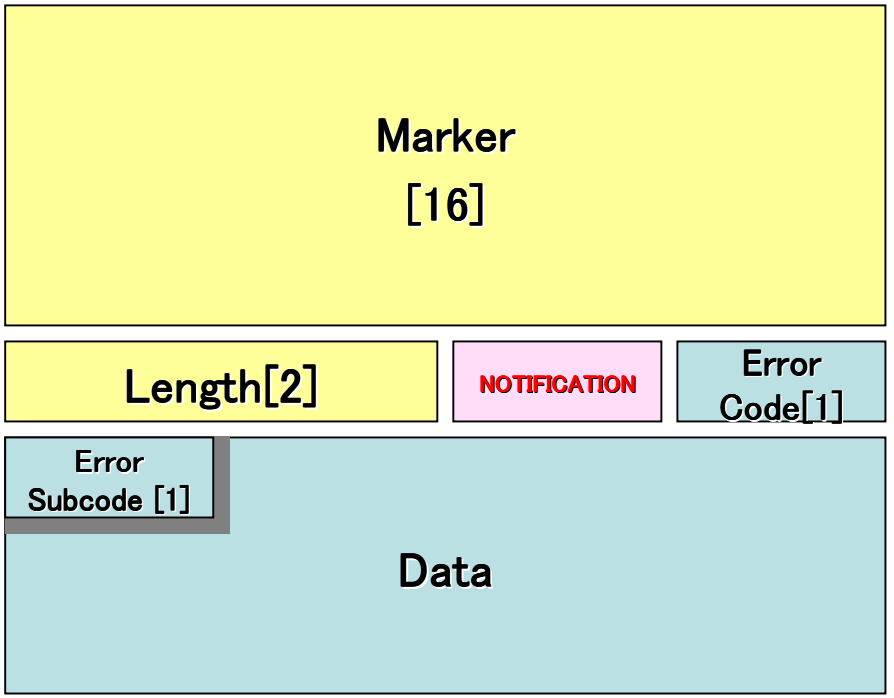
KEEPALIVEメッセージ

- セッションの正常性確認に利用
- UPDATEが一定期間以上発生しない場合に送信
  - Hold Time より頻繁に



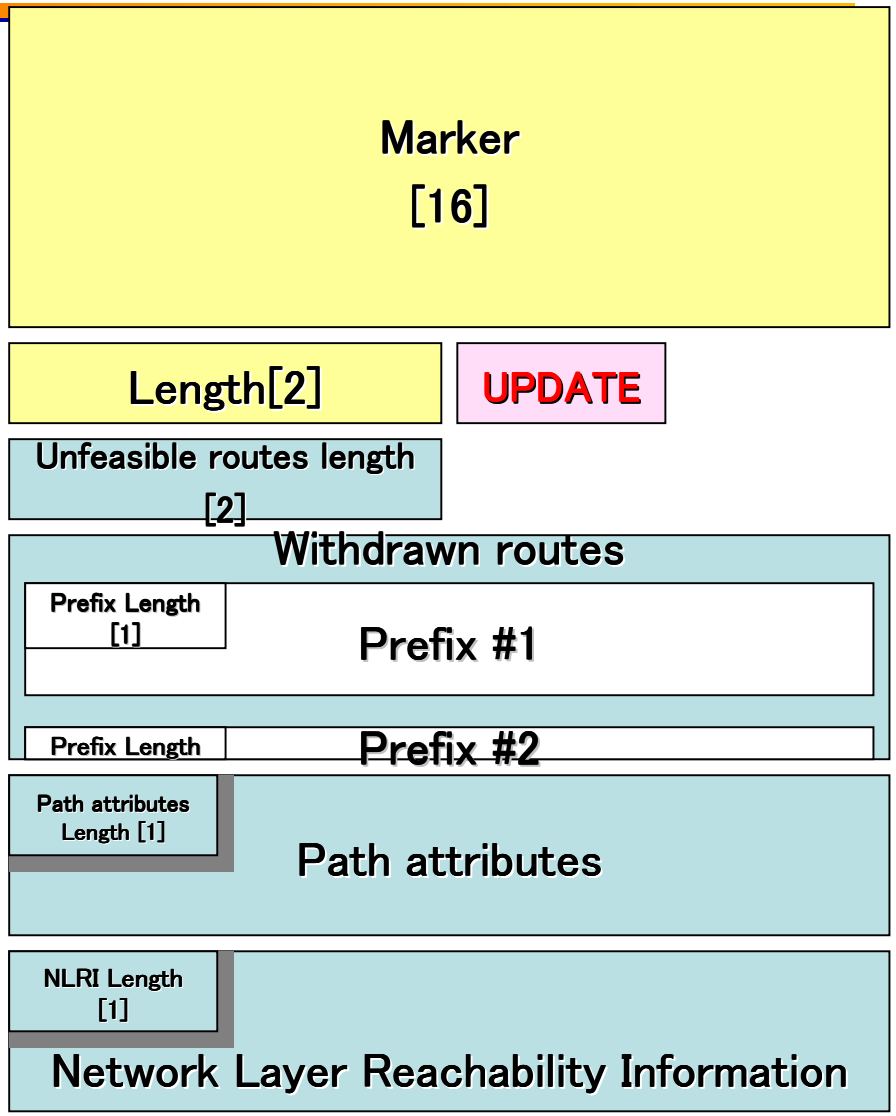
NOTIFICATIONメッセージ

- エラー通知し、セッションを終了する
  - フォーマット不良
  - 無効値
  - 状態遷移エラー
  - Hold Time 満了
  - セッション中止



UPDATEメッセージ

- 経路情報の交換に利用
- 取り消される(withdrawn)プリフィクスを複数、及びパス属性と到達可能プリフィクス(NLRI)の組一対を伝達可能
  - 経路情報を消すのは一気にできる
  - パス属性の変更や新しい経路情報は1プリフィクスずつUPDATEを送信



BGPセッションのライフサイクル

- 両ルータでのコンフィグレーション
- OPEN
  - バージョン不一致などなど不具合があれば NOTIFICATIONを発して切断
  - 不具合なければKEEPALIVEを返してセッション確立
- UPDATEパケットで経路情報を交換
- 新たな経路情報がなければ、KEEPALIVEでセッション維持
- セッション切断のためにはNOTIFICATION

パス属性値
– Path Attributes

ISPバックボーンネットワークにおける経路制御設計～ 理論編 ～
BGP – Border Gateway Protocol

france telecom

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

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
- 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

Path Attributes（パス属性）
ポリシ制御のプレイヤーたち

- LOCAL_PREF – 任意
  - 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 – 非透過性
  - 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 – 透過性
  - 任意の32ビットの情報を伝達する

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)

単純なBGPシステムの導入

ISPバックボーンネットワークにおける経路制御設計～ 理論編 ～

BGP – Border Gateway Protocol

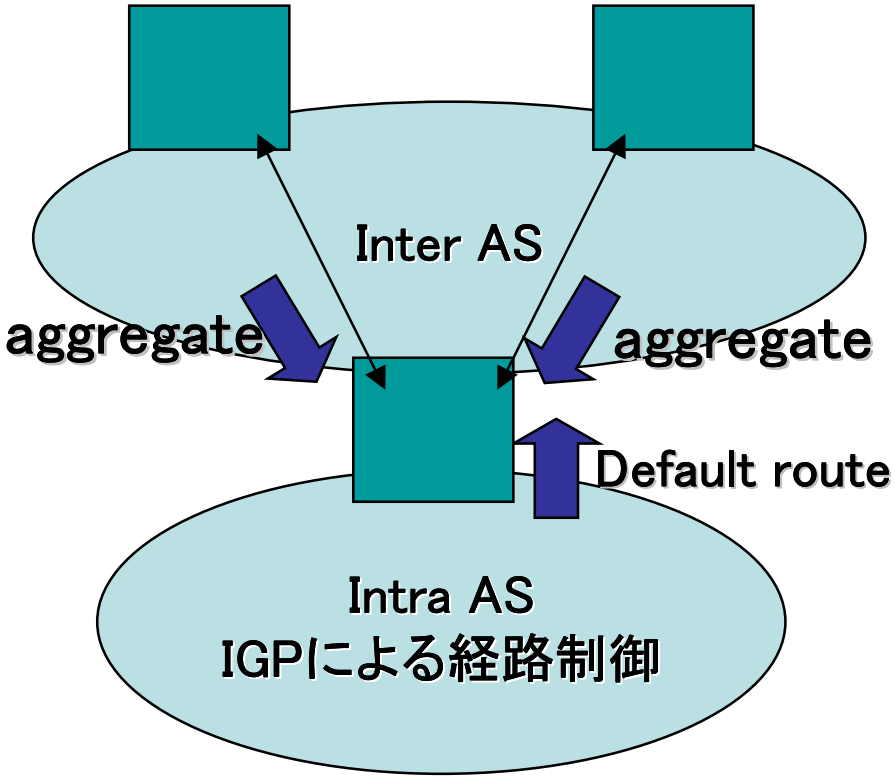
france telecom

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 http://www.nic.ad.jp/ja/translation/rfc/1930.html
  - マルチホーム接続（IX接続を含む）となっていることが条件
    - シングルホームで済むのであれば、上流ASに含まれても問題なく、別のASを構成する必要がない

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
- 2つ以上のASに接続



france telecom

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
- DIX-IE, NSPIXP3, 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

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
 network 192.0.1.0
 aggregate-address 223.224.0.0 255.255.0.0 summary-only
 neighbor 210.171.224.110 remote-as 5511
 neighbor 210.171.224.110 route-map AS551
 neighbor 210.171.224.110 route-map jpix-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

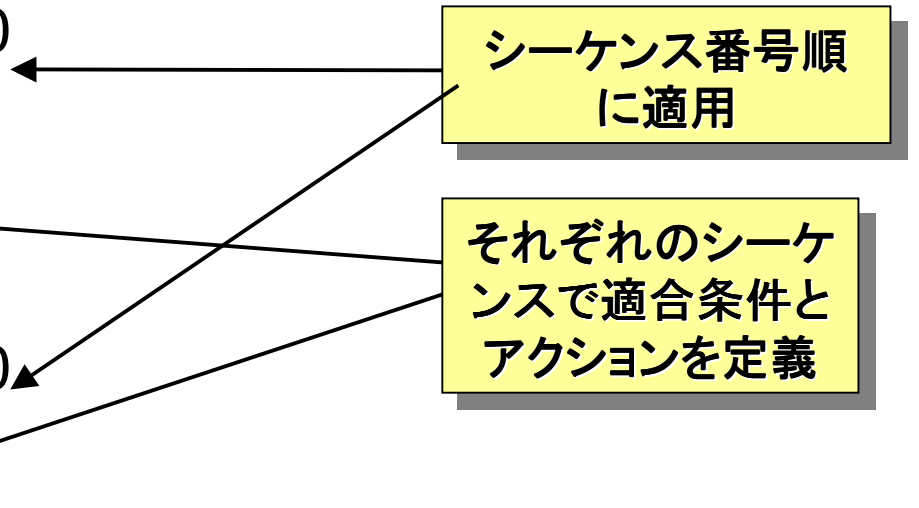
Peer確立

Route-mapで
ポリシを記述

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 AS5511-in permit 10
 match as-path 10
 set local-preference 110
!
route-map AS5511-in permit 20
 match as-path 20
 set local-preference 100
!
```



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 jpix-out permit 10
 match as-path 30
 set metric 1000
!
```

iBGPシステムの構成

ISPバックボーンネットワークにおける経路制御設計～ 理論編 ～

BGP – Border Gateway Protocol

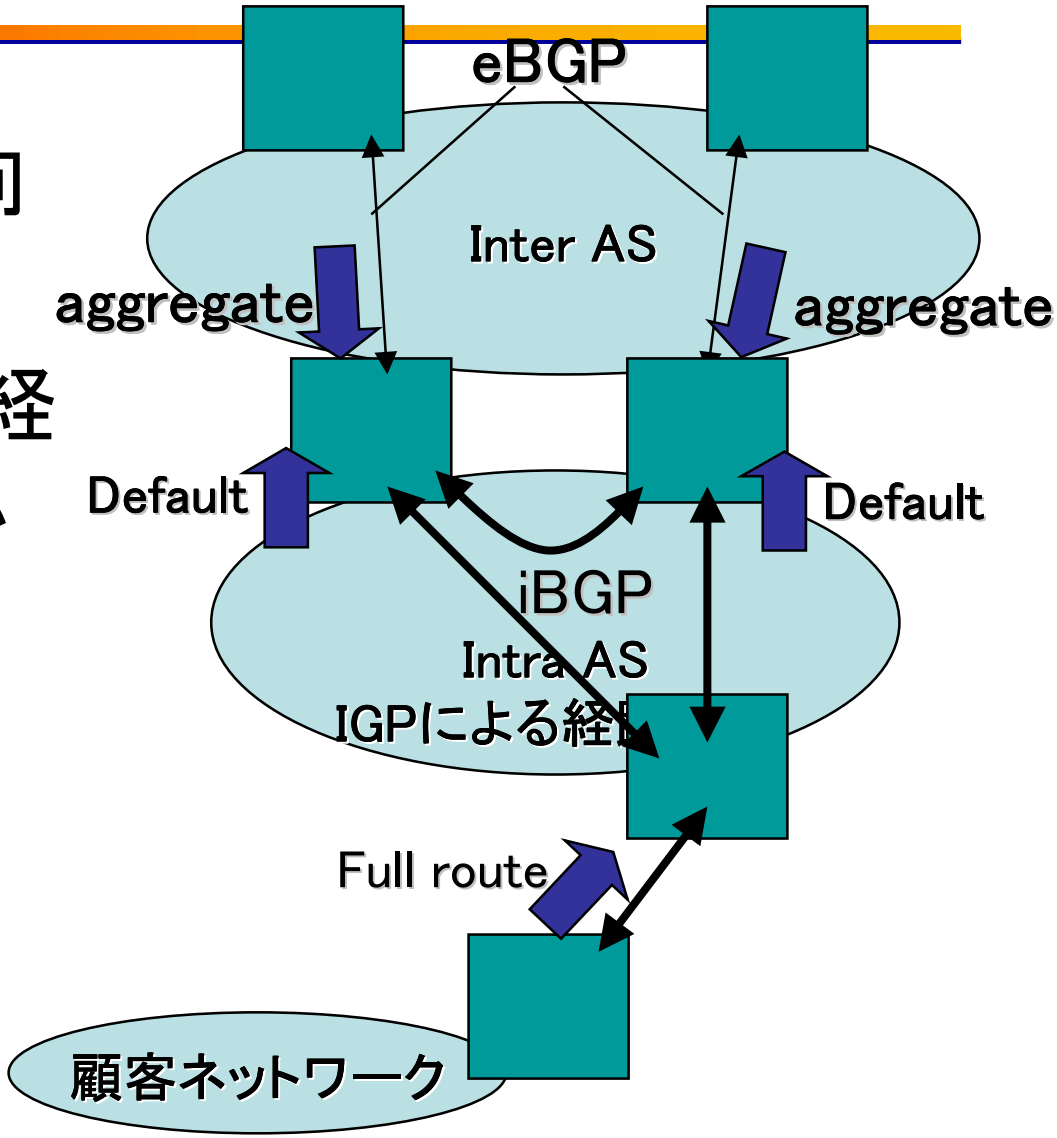
france telecom

複数のボーダルータを置く

- 上流に2つ，顧客向けに1つ
- ボーダルータ間の経路情報の同期が必要

↓

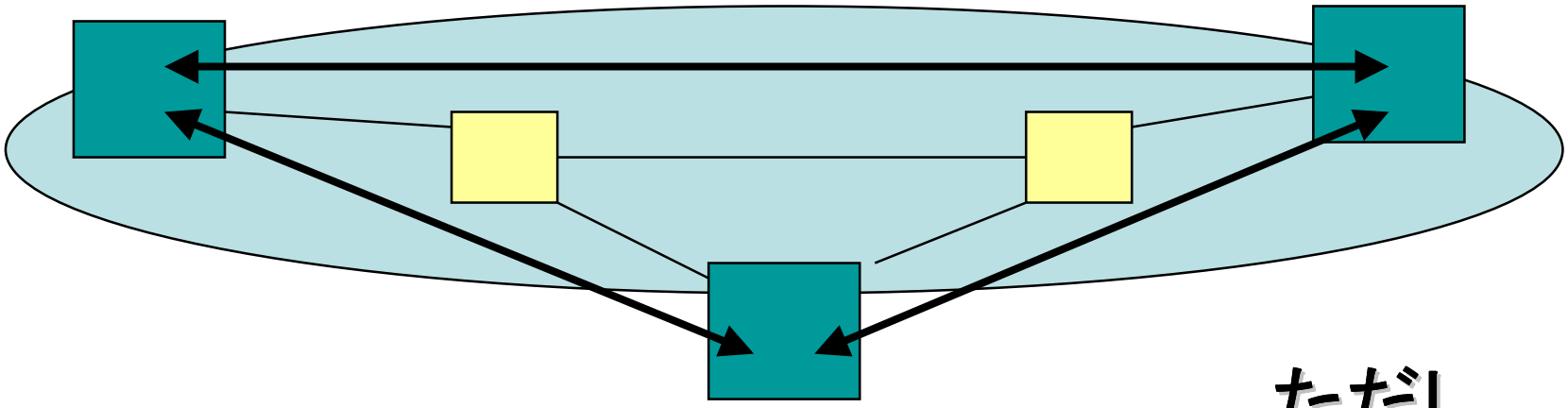
<u>iBGPの確立</u>



france telecom

iBGPの注意点

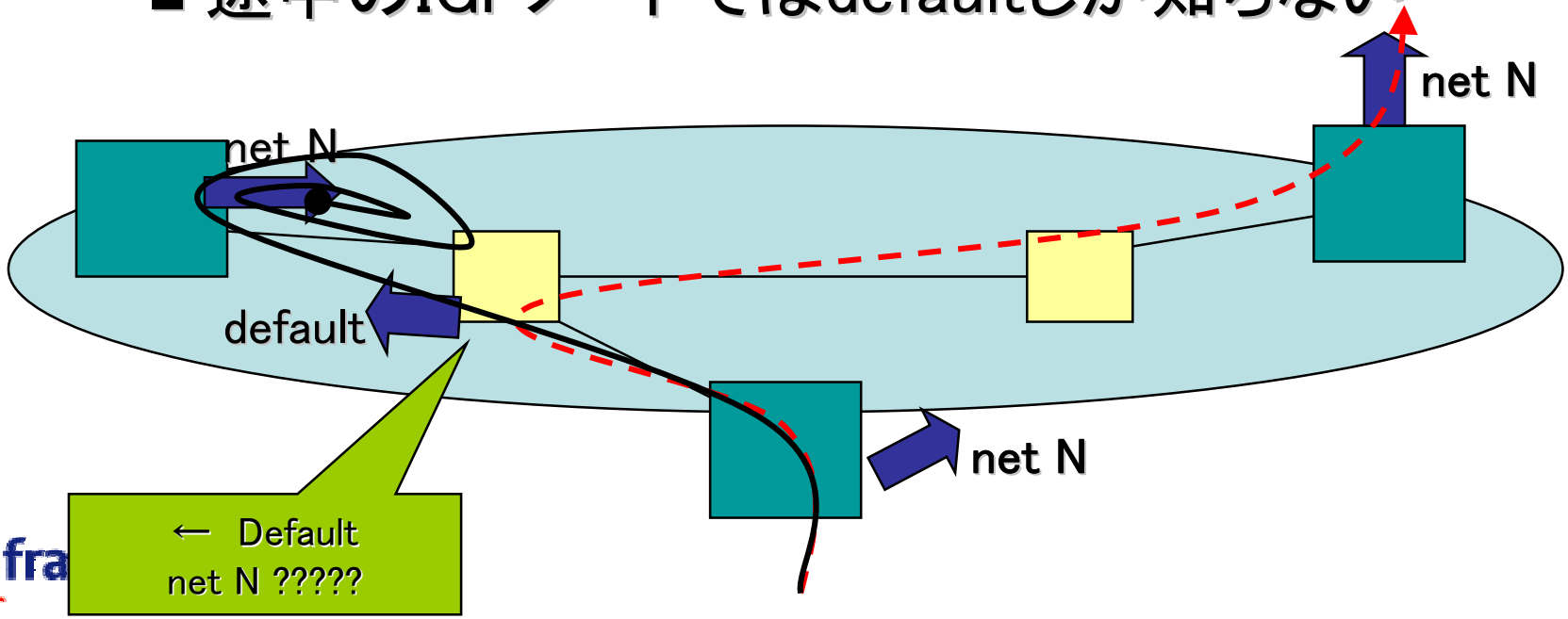
- eBGPは直接隣接を必要とするが、iBGPはAS内での同期が目的なので離れていても確立可能
- iBGPは全てのボーダルータとセッションを張る必要がある



ただし、、

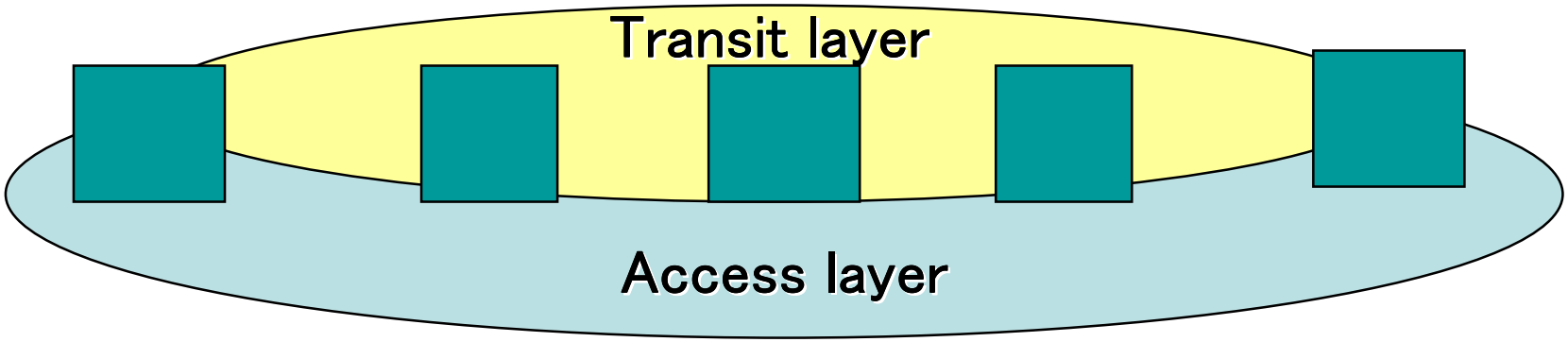
iBGP・仕様上の問題点

- 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



iBGPシステムの解

- No synchronization
  - IGP synchronizationの縛りを解くコマンド(cisco)
  - 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



france telecom

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

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

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」
  - 1oopback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

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 4000
 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 4000
 neighbor IBGP peer-group
 neighbor IBGP remote-as 4000
 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

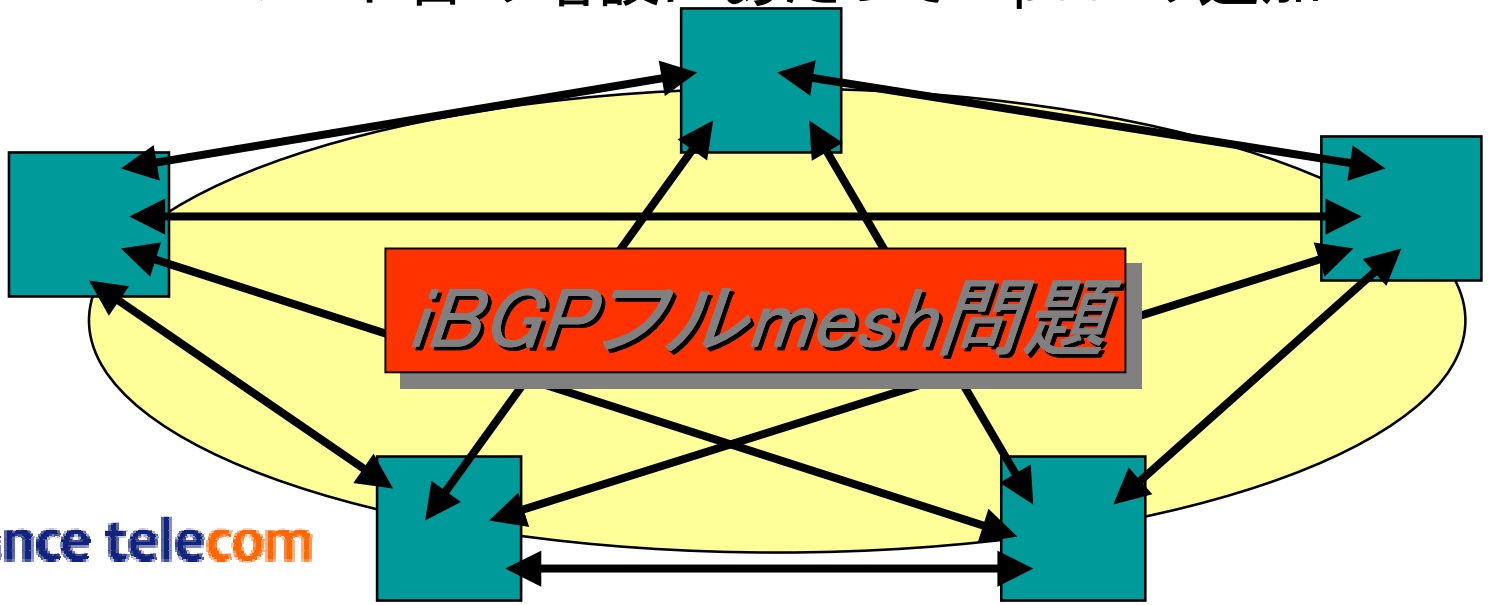
iBGPシステムの スケーラビリティ

ISPバックボーンネットワークにおける経路制御設計～ 理論編 ～
BGP – Border Gateway Protocol

france telecom

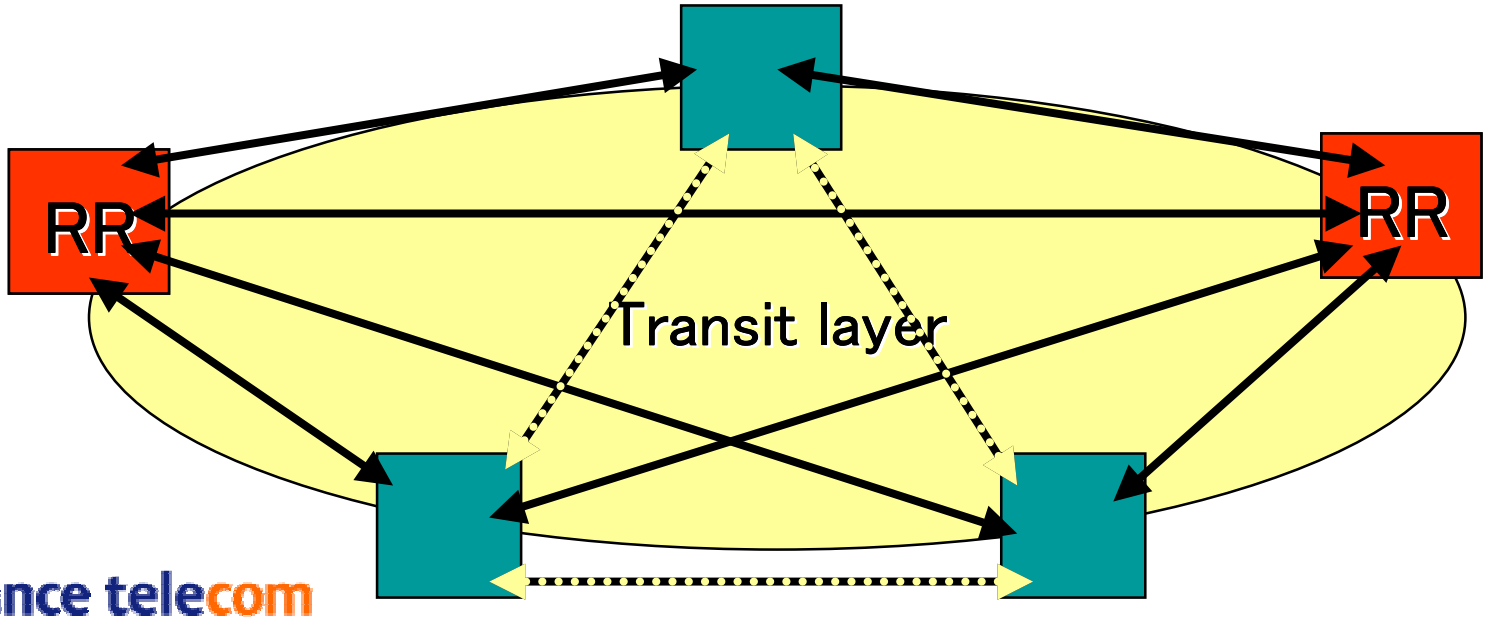
iBGPシステムのスケーラビリティ

- 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



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



france telecom

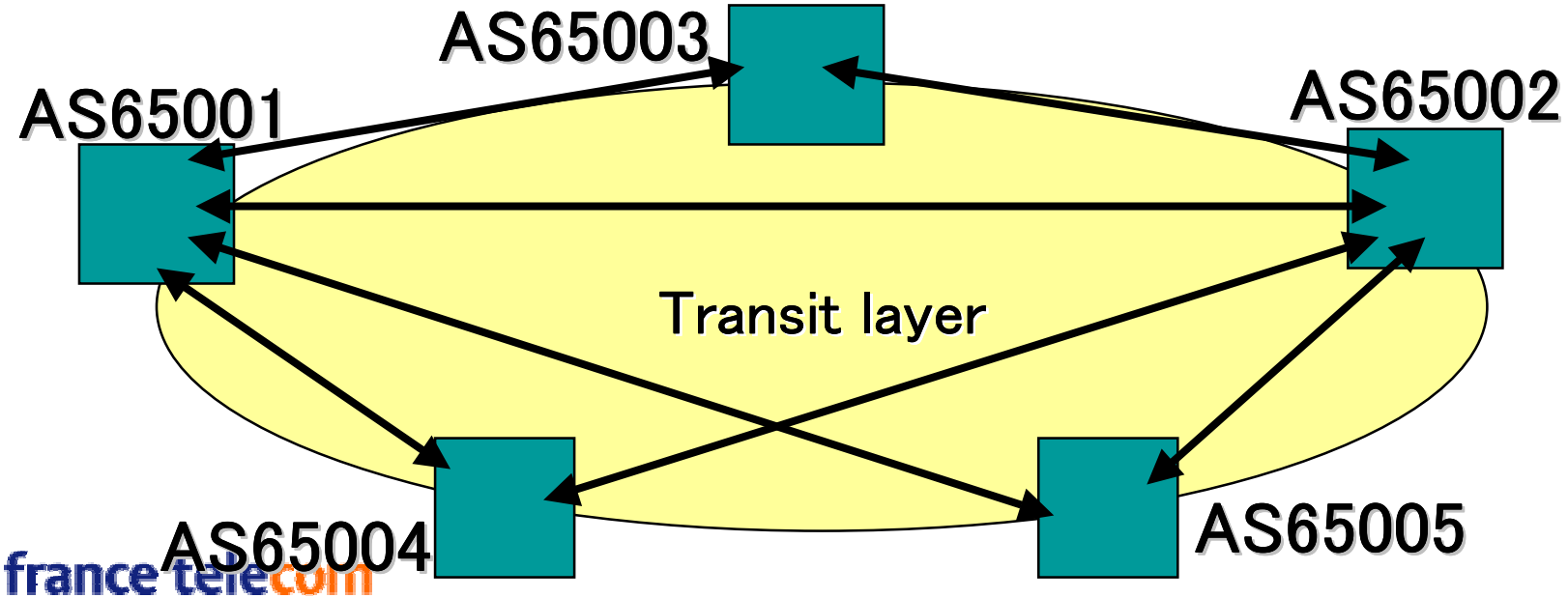
iBGPフルmesh問題解決策
iBGPルートリフレクタ(2)

- 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

```
router bgp 5511
 bgp cluster-id FOUR-BYTE-CLUSTER-ID
 neighbor CLI.ENT.IPA.DDR remote-as 5511
 neighbor CLI.ENT.IPA.DDR route-reflector-client
```

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



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

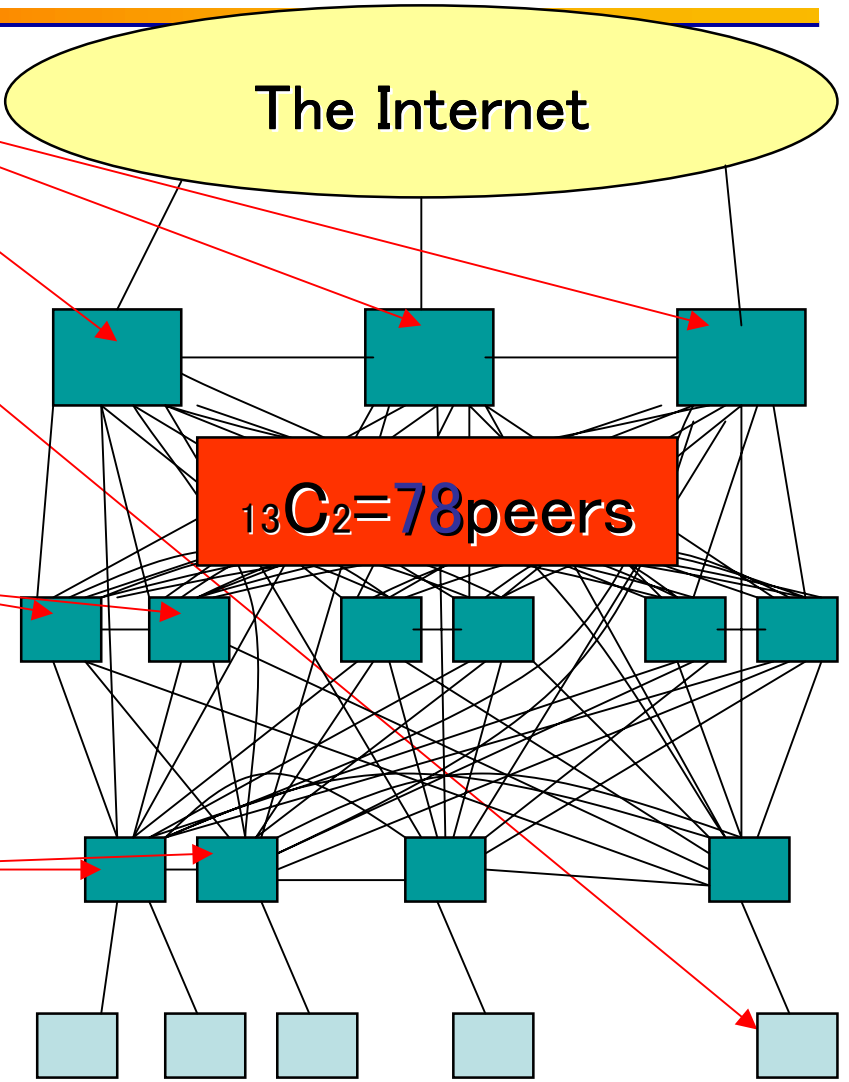
```
router bgp 65000
  bgp confederation identifier 5511
  bgp confederation peers 65001 65002 65003 65004
  network …….
```

- 但しBGPコンフェデレーションはフルmesh回避よりも、管理領域分割や複数ネットワーク統合，IGP分割などに真価を発揮する

france telecom

AS内BGPスケーラビリティ問題の実際

- 複数の対外接続
- 地域/POP毎にBGP接続加入者がいる
  - それぞれBGPノードが必要
- 冗長性確保が必要
  - POPにコアルータを2台
- BGP加入者増加
  - BGP加入者収容ルータの増加



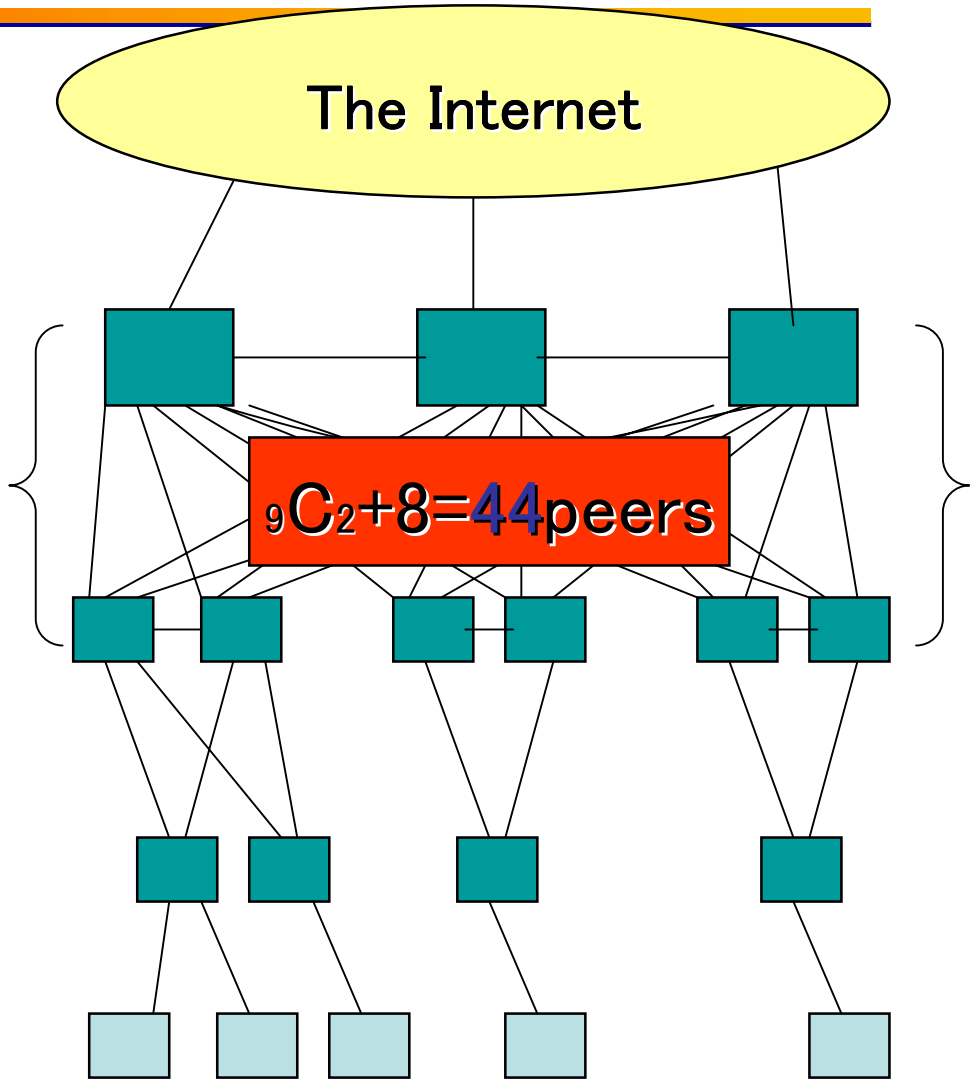
france telecom

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

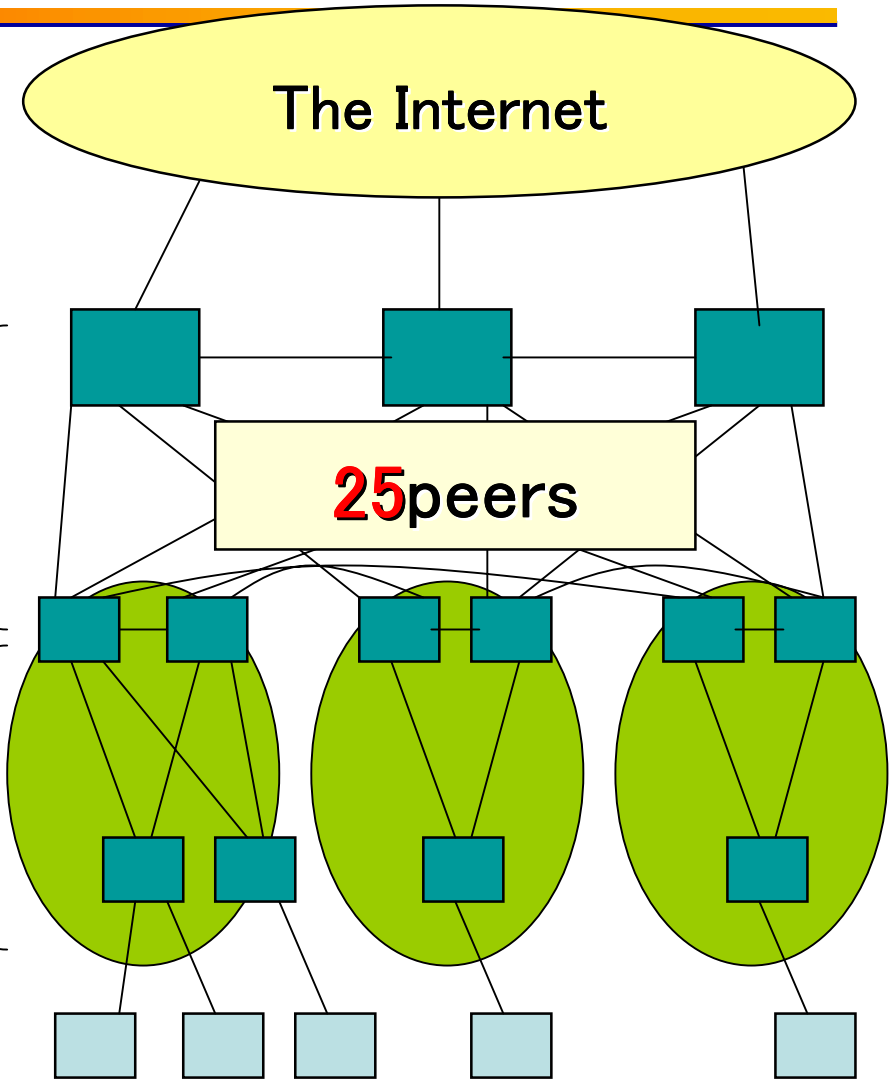
加入者ルータがリフレクタクライアント



france telecom

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



france telecom

現在の経路制御における BGPとOSPFの関係

ISPバックボーンネットワークにおける経路制御設計～ 理論編 ～

BGP – Border Gateway Protocol

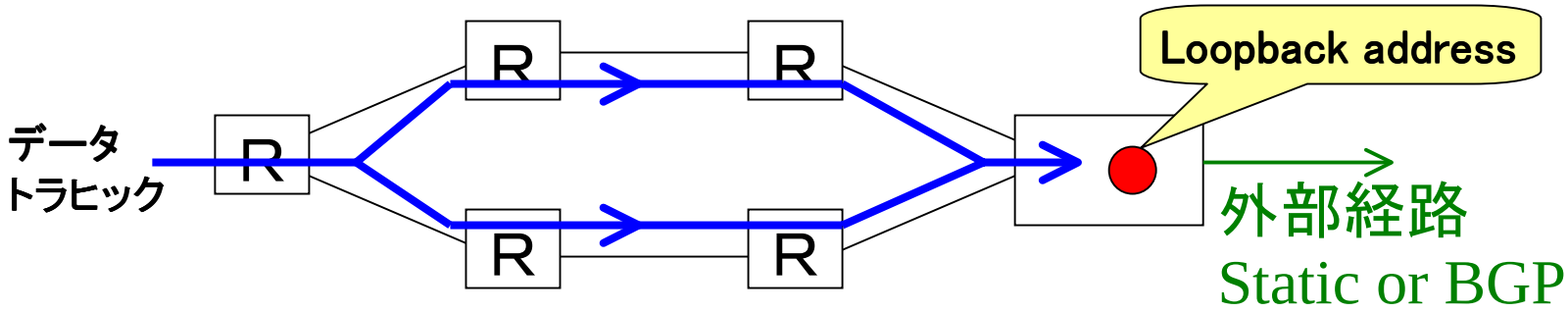
france telecom

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し、フルmeshを回避
- 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経路はBGPにredistributeされる

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



BGPとOSPFの分担(1)

- 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

BGPとOSPFの分担(2)

- いまやゲートウェイは複数・多数ある
  - IGPのデフォルトルートは外部対地の経路制御を行うことは非現実的
- iBGPがAS内のBGP経路の同期だけでなく、AS内での外部対地に対する経路制御を荷っている。

BGP：	staticも含めた外部対地に対する経路制御
OSPF：	内部トポロジの管理

ポリシルーティング

ISPバックボーンネットワークにおける経路制御設計～ 理論編 ～

BGP – Border Gateway Protocol

france telecom

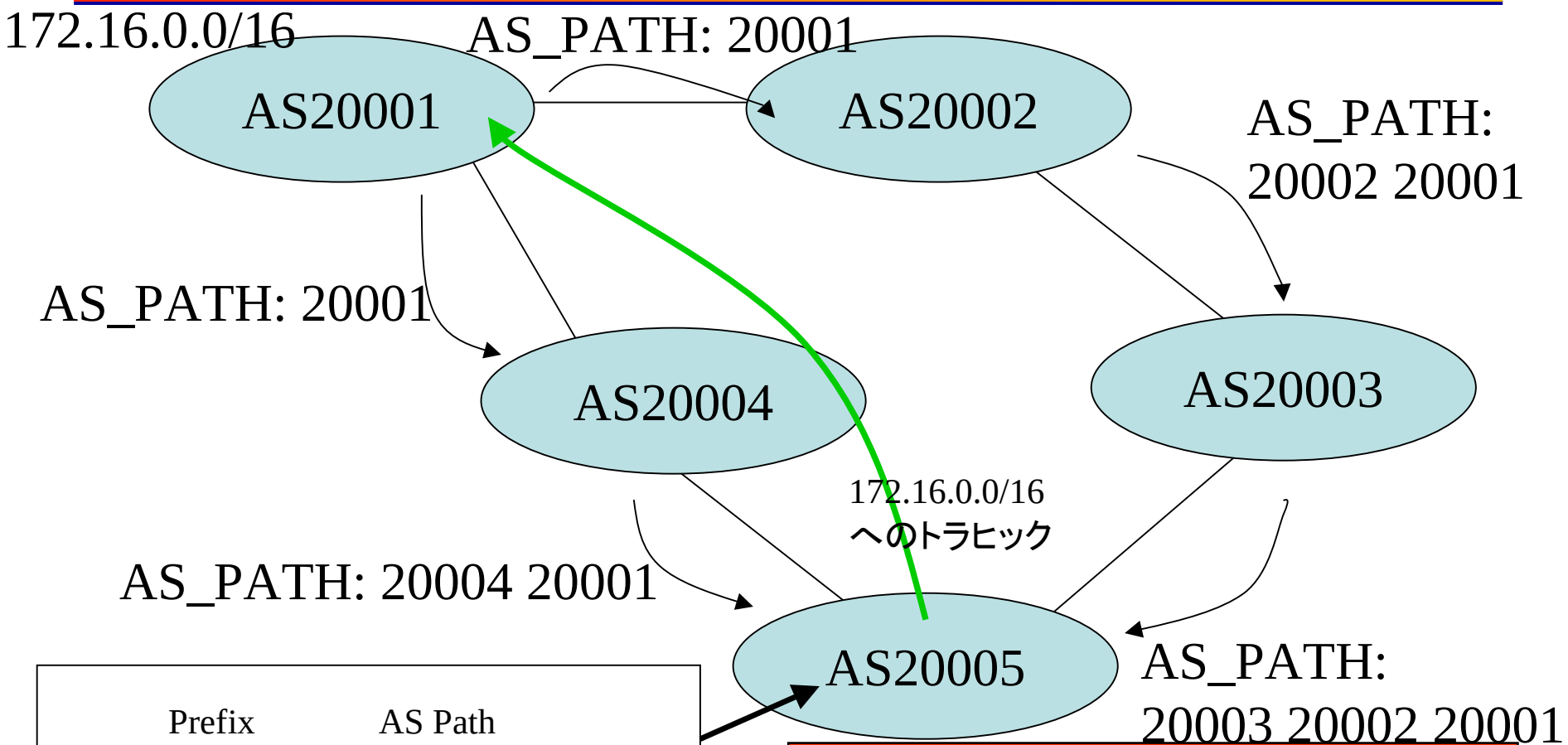
ポリシルーティング

- 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

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。
  - 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
  - 隣接する同一ASの複数peer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

AS_PATH

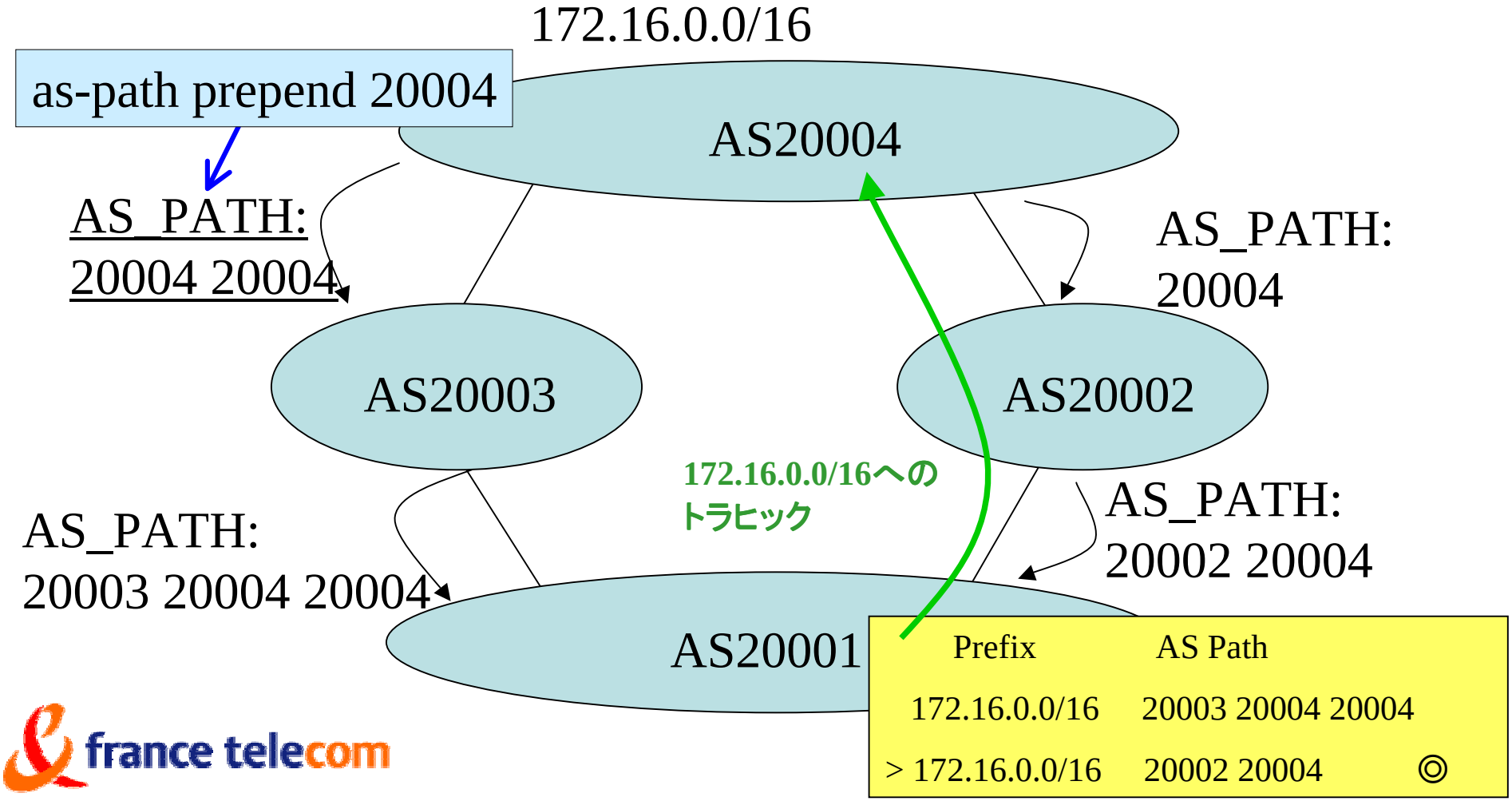


通常、AS_PATHが短い（AS数が少ない）ものを選択する

france telecom

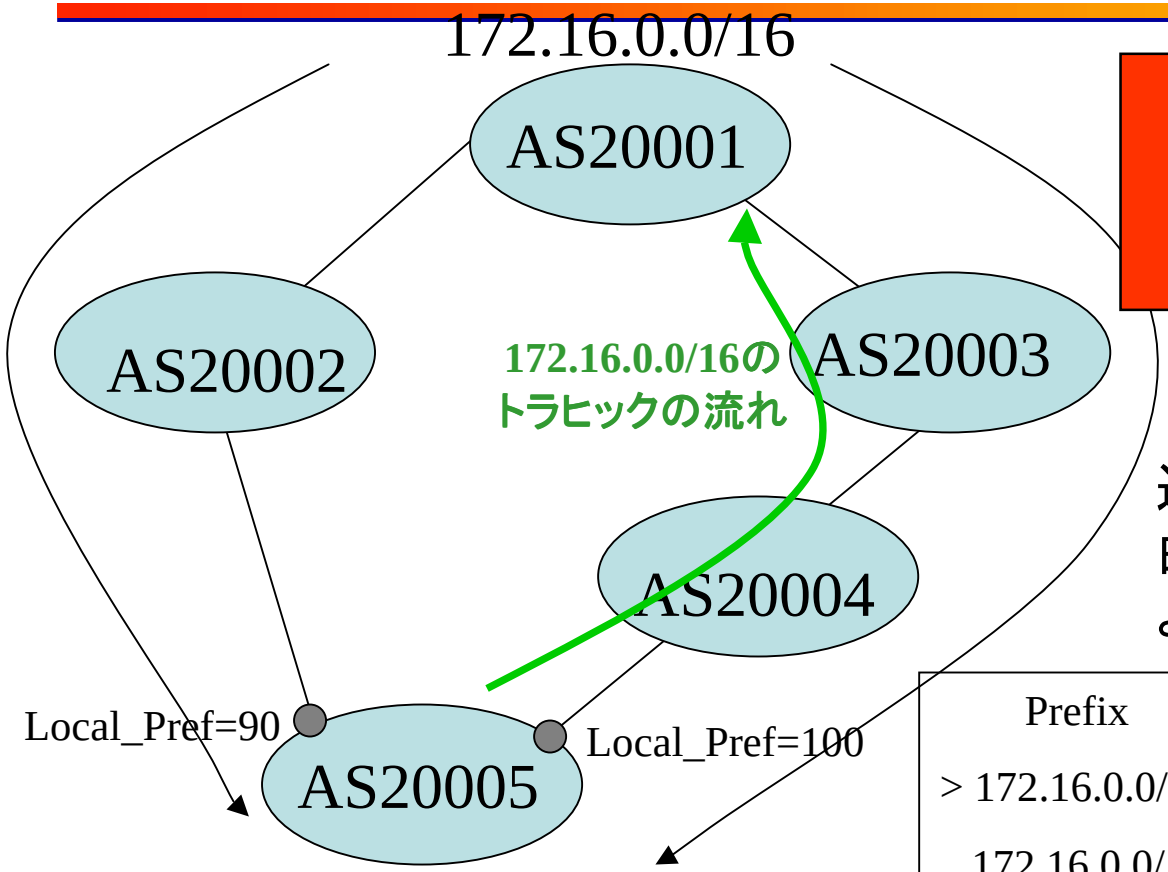
AS Path Prepend

ASを余計につけて、AS_PATH_lengthを長く見せるテクニック



france telecom

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。

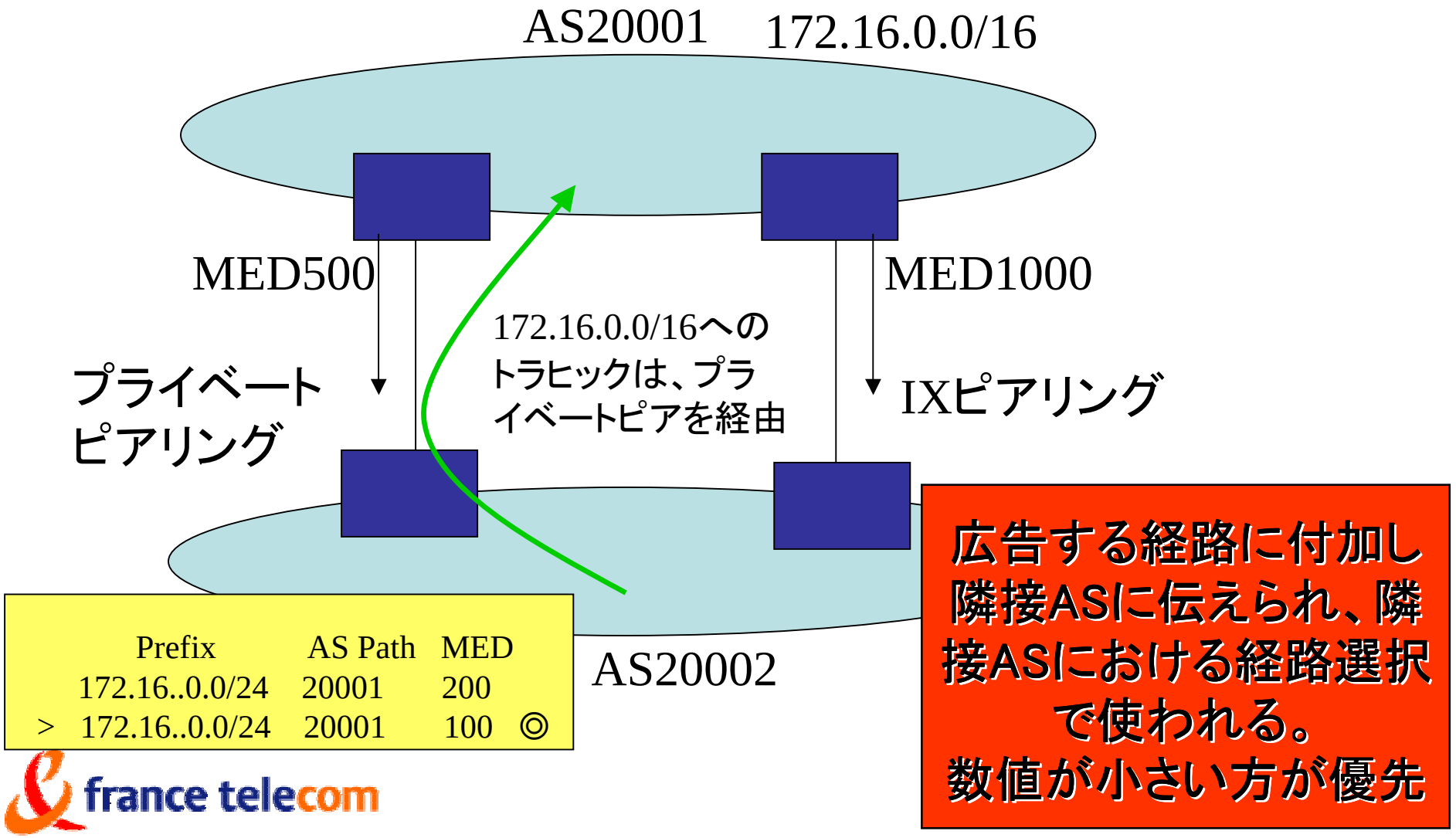
Prefix	AS Path	LocPref	
> 172.16.0.0/16	20004 20003 20001	100	◎
172.16.0.0/16	20002 20001	90	

AS_PATH:
20002 20001

AS_PATH:
20004 20003 20001

france telecom

MULTI_EXIT_DISC



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を発生させる

COMMUNITY(2)

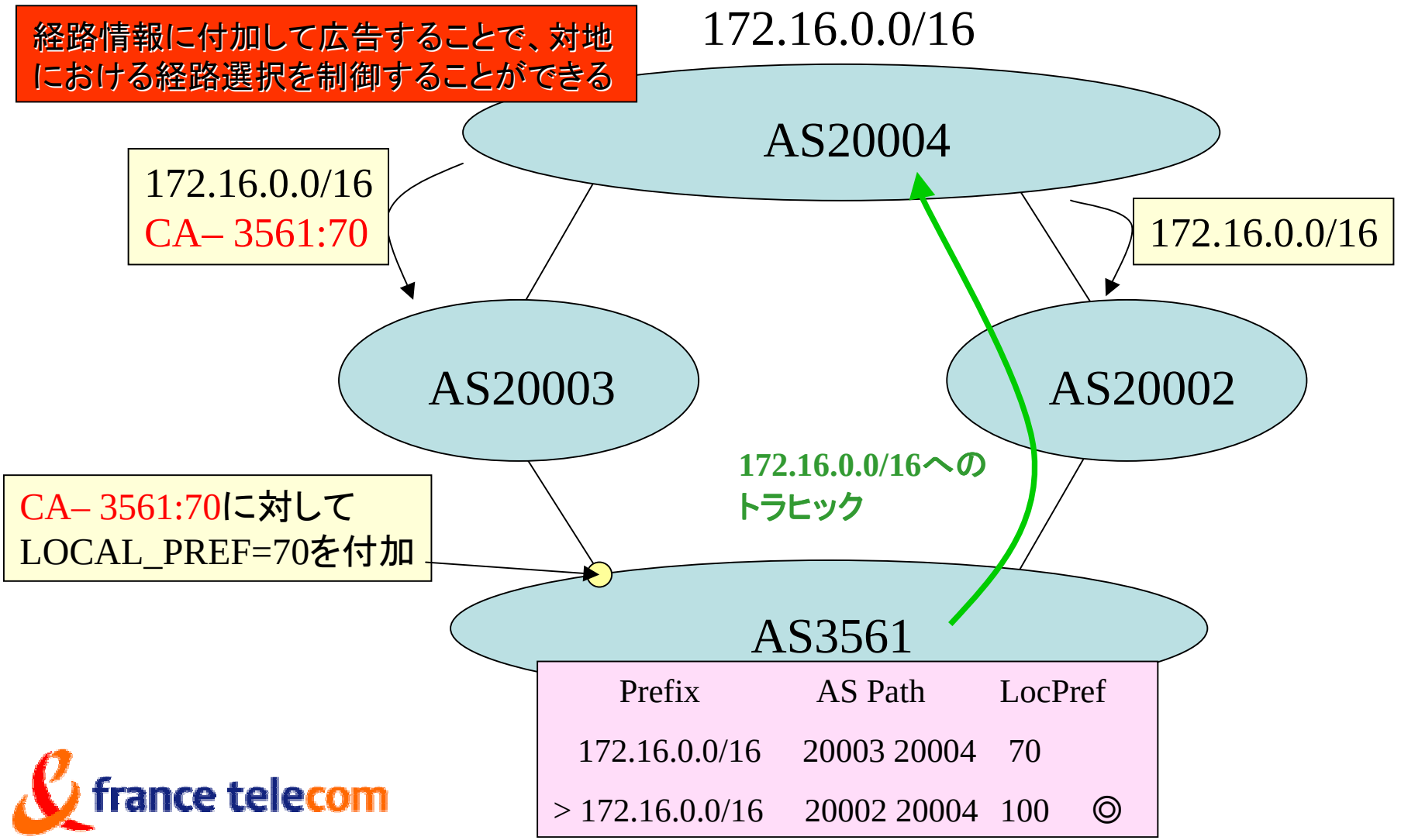
- 一般的な利用法
  - New-format – 32ビットを16ビットずつに二分
    - 5511:1000
  - 上位 – ターゲットAS
  - 下位 – ターゲットASでの動作
- 例1：RFC1998 MCI（現C&Wnet）における実装例
  - 3561:70　そのプリフィクスにLocPref=70付与
  - 3561:80　そのプリフィクスにLocPref=80付与
  - . . . .
  - そのASからの戻りトラヒックの制御に便利！

COMMUNITY(3)

- AS5511 Opentransit Internet の例

5511:1000	米国ピアに非広告	5511:2000	欧州ピアに非公告
5511:1001	米国ピアにプリペンド1	5511:2001	欧州ピアにプリペンド1
5511:1002	米国ピアにプリペンド2	5511:2002	欧州ピアにプリペンド2
5511:1101	Sprintlinkに非公告	5511:2101	Equantに非公告
5511:1102	ICMに非公告	5511:2102	Eboneに非公告
5511:1201	Sprintlinkにプリペンド1	5511:2201	Equantにプリペンド1
5511:1202	ICMにプリペンド1	5511:2202	Eboneにプリペンド1
5511:1301	Sprintlinkにプリペンド2	5511:2301	Equantにプリペンド2
5511:1302	ICMにプリペンド2	5511:2302	Eboneにプリペンド2
5511:1401	SprintlinkにRelayPOPで非公告	5511:3000	アジア太平洋に非公告
		5511:3001	アジア太平洋にプリペンド1
		5511:3002	アジア太平洋にプリペンド2
		5511:4000	欧米亜太以外に広告しない

COMMUNITYの利用方法



france telecom

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

ご静聴ありがとうございました。

ISPバックボーンネットワークにおける経路制御設計
～ 理論編 ～

フランステレコム　ネットワーク・アンド・キャリア・ディビジョン
アジア地域IPプロダクト担当

前村 昌紀
akinori.maemura@francetelecom.com , maem@opentransit.net

france telecom