大規模ネットワークにおける経路制御設計

友近 剛史（NTT コミュニケーションズ（株））

前村 昌紀（日本電気（株））

1999 年 12 月 16 日

Internet Week 99　パシフィコ横浜

（社）日本ネットワークインフォメーションセンター編

目次

1 概要 1

2 IGP でのシステム設計論 1

3 BGP によるシステム設計論 22

4 文献紹介：Jessica Yu による Internet Draft『Scalable Routing Design Principles：規模対応性の高い経路制御設計の指針』..... 30

5 事例紹介：スタティック経路の BGP への再分配 33

6 事例紹介：コンフェデレーションの応用 37

1　概要

ネットワークが大規模なものとなっていくと、トラフィックだけでなくルーティングに関する問題にも対処しなければならなくなります。この講演では、次の 2 つを中心に、大規模ネットワークでの経路制御について説明します。

- OSPF（Open Shortest Path Fast）
- BGP（Boarder Gateway Protocol）

2　IGP でのシステム設計論

米国では 3 年くらい前から大規模ネットワークにおける IGP（Interior Gateway Protocol）のスケーラビリティが問題視され始め、日本でも 1 年ほど前から実際にいくつかの問題が出てきています。ここでは、最初に、なぜ IGP のスケーラビリティが問題視されるのかを示します。

2.1　概要

まず、ルーティングの基本について復習しておきます。図 1 に示すように、ルータでは、IP パケットのヘッダ部分に書かれている宛先 IP アドレスとルータ内のルーティングテーブルによって、IP パケットの転送先が決定されています。

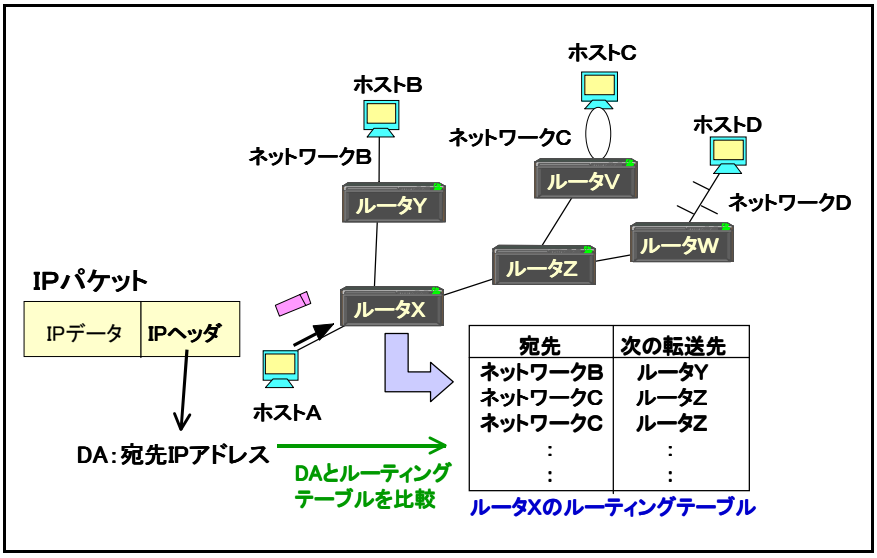

図 1：ルータの動作

このような動作で利用される、ルーティングテーブルをどのようにして作成するかが「ルーティング」です。ルーティングには、各ルータ上でルーティングテーブルを手作業で作成する「スタティックルーティング」と、各ルータのルーティングテーブルを自動的に作成する「ダイナミックルーティング」があります。

ダイナミックルーティングでは、次のように実際のデータの流れとは逆向きの順序で、ルーティングテーブルが作成されます。

1. ルータ B からルータ A に対して経路情報が渡される。
2. ルート A 内でルーティングテーブルが作成される。
3. ルーティングテーブルの内容に従って、ルータ A からルータ B にデータが渡される。

そして、このようなルーティングのためのプロトコルには、AS（Autonomous System：自律システム）内で利用される IGP（Interior Gateway Protocol）と、AS 間で利用される EGP（Exterior Gateway Protocol）の 2 種類があります。

このうち IGP として利用されるプロトコルには、次のものがあります。

- RIP（Routing Information Protocol）（2.2 を参照）
- OSPF（Open Shortest Path Fast）（2.3 を参照）
- IS-IS（Intermediate System-to-Intermediate System）（2.4 を参照）

EGP として利用されるプロトコルには、次のものがあります。

- BGP（Boarder Gateway Protocol）（3 を参照）

また、ルーティングプロトコルでは、次の 3 種類のアルゴリズムが利用されています。

- ディスタンスベクターアルゴリズム
- リンクステートアルゴリズム
- パスベクターアルゴリズム

このうち、ディスタンスベクターアルゴリズムでは、隣接するルータ同士が経路情報を交換し合うことで、各ルータがネットワーク情報を入手していきます。このときには、他のルータから受け取ったルーティングテーブルに、自らが直接接続しているネットワークを追加した後に、そのルーティングテーブルを別のインタフェースに渡します。

これに対してリンクステートアルゴリズムでは、図２ に示すように、各ルータが自ら接続しているネットワークの情報等を、ネットワーク全体に通知します。そして、各ルータは共通のトポロジカルデータベースを作成します。これは地図にあたるもので、ネットワーク構成がどのようになっているかがわかるデータベースです。最近ではこれを「リンクステートデータベース」と言いますが、今回は意味を理解しやすいように、昔言われていた「トポロジカルデータベース」という呼び方にします。

各ルータでは、このトポロジカルデータベースから、自らをルートとした最短パスツリーを作成し、ルーティングテーブルを作成します。

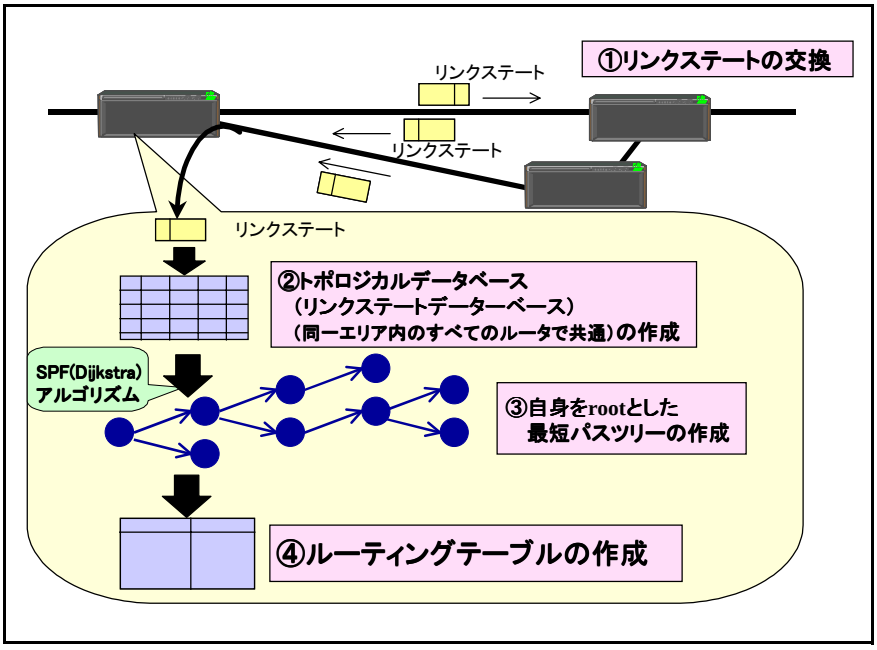

図２：リンクステートアルゴリズム

3 番目のパスベクターアルゴリズムは、BGP で利用されているアルゴリズムです。このアルゴリズムでは、経路情報にパス属性というものが付加され、それを基に経路選択されます。図 3 に AS パス属性という、パス属性の 1 つの例を示します。図 3 のように、パス属性が経路情報に付加されて伝わり、それを基に経路選択されます。

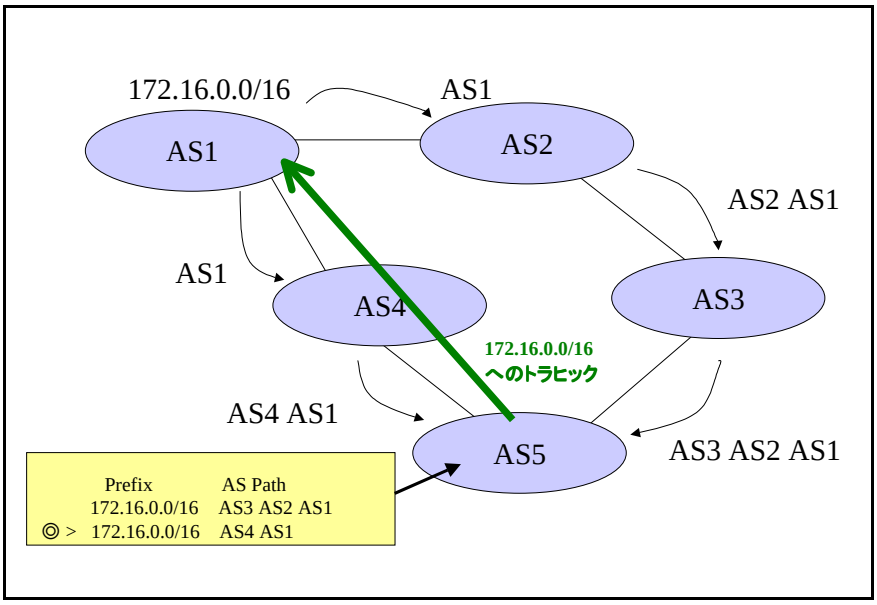

図 3：パスベクターアルゴリズム

2.2 RIP

ディスタンスベクターアルゴリズムを利用する RIP（Routing Information Protocol）は、実装が容易で処理負荷が小さいため、現在では多数のネットワーク機器が対応しています。ただし、RIP には、次のような欠点があります。

- サブネットマスクの情報が渡されないため VLSM（Variable Length Subnet Mask）に対応していない。
- 送信元と受信先間での最適な経路がホップ数によって決定されるため、回線の帯域幅に応じて適切な経路を選択することが難しい。
- 最大ホップ数が 15 までに制限されている。
- デフォルトでは、経路情報が 30 秒ごとに各ルータによってブロードキャストされるため、トラフィックが増加する。

また、ディスタンスベクターアルゴリズムを利用する RIP は、網変更等を実施すると収束に時間がかかってしまうため、大規模ネットワークでは利用できません。

このような RIP の欠点のいくつかを、互換性を保ったまま改善したものが RIP2 です。RIP2 では、認証機構が提供され、経路情報をブロードキャストだけでなくマルチキャストでも受け渡すことができます。また、RIP2 では VLSM に対応するため、クラス C 等で 1 つのネットワークを /26 や /27 等、複数の長さに分割したサブネットマスクの情報も渡されるようになっています。ただし、RIP2 でもディスタンスベクターアルゴリズムを利用しているため、網変更後の収束に時間がかかってしまい、大規模ネットワークには適していません。

2.3 OSPF

リンクステートアルゴリズムを利用する OSPF（Open Shortest Path Fast）は、RFC 2328 によって規定されています。

OSPF は、VLSM にも対応し、IP アドレス、サブネットマスク、接続されるネットワークタイプ等を表すリンクステートというリンクの状態情報をマルチキャストによって配布します。このリンクステートは、30 分ごとのリフレッシュ時以外では、トポロジに変更があったときにのみ更新内容が送信されます。ルーティングテーブル自体は交換されません。OSPF では、トポロジカルデータベースと呼ばれるデータベースを持っています。これはネットワーク構成がどのようになっているかわかるデータベースで、地図のようなものです。同じエリアのすべてのルータで共通のトポロジカルデータベースを持っており、このことが OSPF の最大の特徴と言えます。

ここからしばらくは、Cisco Systems 社のルータでの OSPF の設定例を説明します。

まず、OSPF を設定するときにはプロセスを起動し、インタフェースを指定します。これらの設定は次のようなコマンドを使って実行します。

```
router ospf 〈プロセス ID〉
 network 192.168.0.0 0.0.0.15 area 0
```

このうち〈プロセス ID〉には、1 〜 65535 までの任意の番号が指定できます。通常、1 つの AS 内で 1 つの OSPF プロセスしか動作させないときには、プロセス ID を AS 番号と同じものとすることが多いようです。また、network コマンドは、そのルータの connected なインタフェースのうち、その network の範囲に当てはまる Interface について、次の 2 つの意味を持ちます。

- OSPF を話す。
- その connected なネットワークの情報を OSPF に広告する。

図 4 のように、あるインタフェースのネットワークを OSPF に広告したいが、そのインタフェースでは OSPF パケットのやりとりをしないようなときには、passive-interface というコマンドを設定します。network コマンド＋ passive-interface コマンドと同様の設定は、redistribute connected subnets でも実行できます。

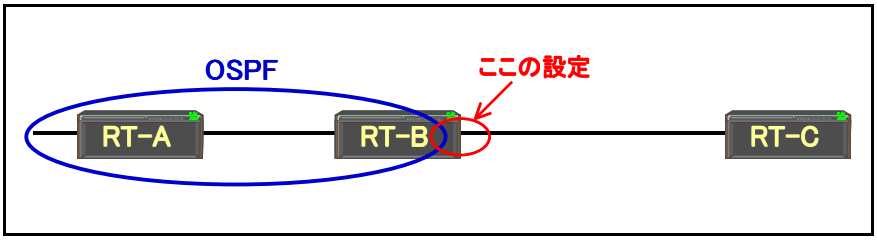

図 4：スタブなネットワークの参加

さらに、インタフェース用のコマンドを使って、次のような設定ができます。

- コストによるネットワークごとの重み付け
- HELLO パケットの送出間隔
- HELLO パケットを受け取れなかったときに、障害だと判断するまでの間隔
- OSPF パケットの認証

また、同一コストの複数パスを同時に使用して、ロードシェアリングもできます。

OSPF の設定には、いくつか注意すべき点があります。

まず、OSPF では、デフォルトルートの情報が再分配されません。このため、default-information originate や default-information originate always 等のコマンドによって、デフォルトルートを広告する必要があります。デフォルトルートは、BGP スピーカーでないルータ（エッジに近いルータ）が、BGP スピーカー（インターネット側に近いルータ）までデータパケットを転送できるようにするために設定します。デフォルトルートを広告しているルータでは、プライベートアドレス等の未知なアドレス向けのパケットが送られてきたときに、そのパケットを破棄します。ただし、このような処理は高負荷なものとなるため、インタフェースでパケットを破棄できるようなルータ（例：GSR）でデフォルトルートを広告すべきです。

また、OSPF では、コスト（メトリック）のタイプについても注意する必要があります。スタティックや他のルーティングプロトコルから OSPF に再分配（redistribute）される外部ルート（External routes）のコストには、タイプ 1 とタイプ 2 の 2 種類があります。このうちタイプ 1 は、外部ルートでのコストに、そこまでの内部ルートのコストを加えたものとなります。これに対してタイプ 2 は、外部ルートでのコストのままとなります。デフォルトでは、タイプ 2 のコストが利用されます。また、同一のネットワークに関する複数の経路情報については、コストで比較される前に、エリア内の内部ルート、エリア外の内部ルート、タイプ 1 の外部ルート、タイプ 2 の外部ルート、の順で優先されることにも注意してください。

また、OSPF のルータ ID を安定させる意味でも、loopback アドレスは設定しておくべきです。

さらに、OSPF では、ループバックアドレスが設定されているときは、ルータ ID はそのループバックアドレスとなります。これに対して、ループバックアドレスが設定されていないときは、最大の IP アドレスがルータ ID となります。この場合、ルータ ID が変化するとリンクステートを再度受け渡す必要性が生じるため、基本的にはルータ ID にはループバックアドレスを設定すべきでしょう。ループバックアドレスをルータ ID に設定することで、ダウンすることなく安定して運用できるようになるわけです。

2.3.1 OSPF による網設計

網設計においては、最初に要望条件を整理し、ポリシーを策定するようにします。次に、いくつかの例を示します。

- 基本機能の実現
  - 静的状態での接続性
  - 迂回機能の実現
- コストの低減
  - 回線数や回線量の削減
  - バックアップ回線も、1 対 1 のアクトスタンバイではなく、N 対 1 によるロードバランスとする。
- 信頼性の向上
  - ノード障害
  - リンク障害
  - 機種レベルでの冗長化
  - 技術レベルの冗長化
  - ビル障害

- 保守運用性の向上
  - 物理構成や論理構成が単純
  - 地域ごとやサービスごとに分離可能
  - 移行が容易
- 将来性
  - ビル数、ノード数、ユーザ数の増大対応
  - サービス種類の増大対応

OSPF では、各エリア内のすべてのルータで共通のトポロジカルデータベースが作成されます。また、図 5 のように、バックボーンエリア以外のエリアは、バックボーンエリア（area 0）にぶらさがるような形になります。エリアの設計においては、まずバックボーンエリアを検討（または構築）した後に、各エリアを検討（または構築）していくようにします。1 台のエリア境界ルータが担当するエリアは、なるべく 2 つまでとします。また、バックボーンエリア以外の 1 つのエリアには、複数のエリア境界ルータを設置して、冗長性を確保するようにします。さらに、当然のことですが、経路は各エリア境界ルータによって集約するようにします。

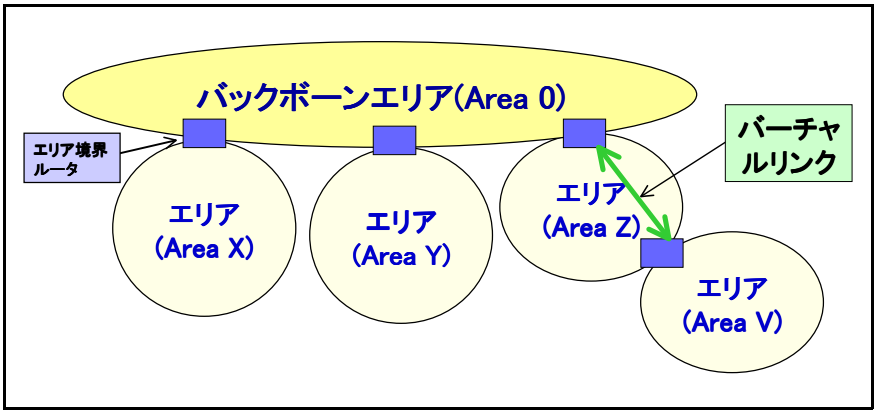

図 5：エリア設計

なお、大規模ネットワークでは、バーチャルリンクに期待してネットワークを設計すべきではありません。これは、バーチャルリンクによって、設計が複雑になるだけでなく冗長性も確保しづらくなるためです。たとえば、図 6 のようなネットワークでは、エリア V をエリア 0 以外にしてしまうとルータ B が 3 つのエリアを担当することになってしまいます。これを避けるために、エリア V をエリア 0 とすると、エリア 0 が大きくなってしまい、規模対応性に対する効果がほとんど得られなくなってしまいます。このため、バーチャルリンクは、パッチングや網変更等での緊急措置対応のためにのみ利用するようにします。

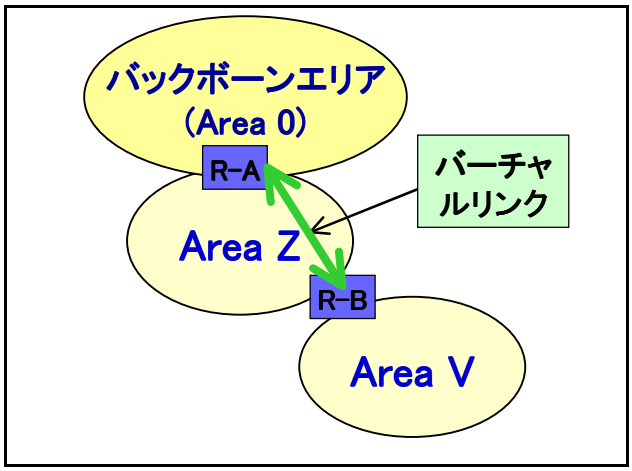

図 6：バーチャルリンクの利用

DR（Designated Router：指名ルータ）は、イーサネット等のマルチアクセスネットワーク上に必ず 1 つ存在します。また、DR がダウンしたときのためのバックアップとして BDR（Backup Designated Router：バックアップ指名ルータ）も存在します。それ以外のルータは DROTHER となります。DR 以外の各ルータは、DR と情報を交換します。DR となったルータの負荷は結構高くなるため、処理能力が高いルータや、他の処理があまり発生しないルータを DR とするようにします。また、1 つのルータが複数のネットワークの DR とならないように注意する必要もあります。

DR によって、隣接する各ルータは DR に 1 度リンクステートを送るだけで、そのリンクステートが DR によって他のすべての隣接ルータに送られるようになります。これによって、図 7 のように、より少ないトラフィックで、各ルータが相互にリンクステートを交換できることになります。

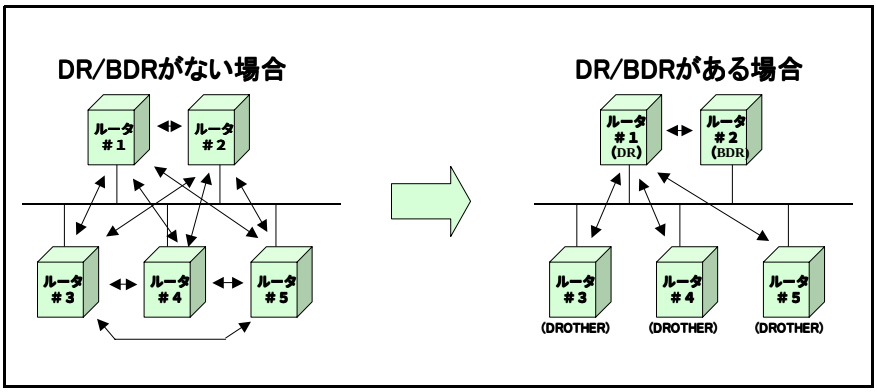

図 7：DR の利用

OSPF では、DR へのなりやすさが設定できます。たとえば、Cisco Systems 社のルータでは、次のコマンドによってDRへのなりやすさを設定します。

```
ip ospf priority *
```

このときには、高い値が指定されたルータほど、DR になりやすくなります。ただし、対象となるネットワーク上に DR や BDR が既に存在していたときには、それ以上に高い値を指定したとしても、DROTHER にしかなりません。このため、実際には最初に起動した 2 つのルータがそれぞれ DR と BDR となってしまうため、ネットワークを新たに起動するとき等には、優先度を高くしたルータから起動するようにする必要があります。また、優先度としてゼロ（0）を指定したルータは DR や BDR には選ばれないため、負荷を増加させたくないルータに対してはゼロを指定するようにします。

2.3.2 ルーティングテーブルの作成

大規模ネットワークにおいて、OSPF がどのような影響を与えるのかを知るためには OSPF のプロトコルについて理解する必要があります。既に示したように OSPF では、図 8 のような手順で実行されるリンクステートアルゴリズムが利用されています。

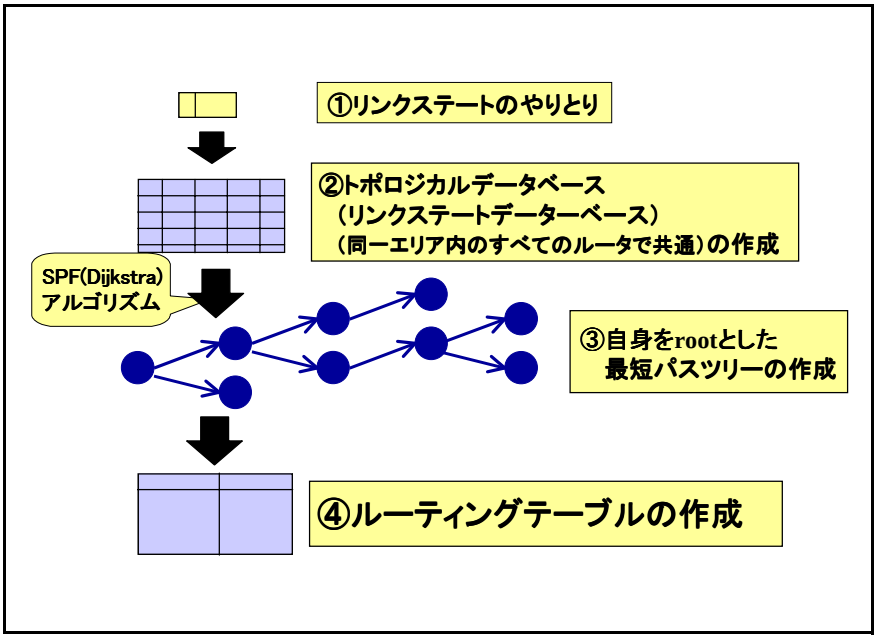

図 8：リンクステートアルゴリズムでの処理内容

図 8 に示したトポロジカルデータベースは、最近では「リンクステートデータベース」と呼ばれますが、今回はイメージがつかみやすいように「トポロジカルデータベース」と呼ぶことにします。これはルータとネットワークによって構成された有向グラフです。次に、トポロジカルデータベースの作成の仕方を説明します。

まず、複数のルータが存在しているマルチアクセスネットワークでは、図 9 のように、ルータからネットワークへ、また逆にネットワークからルータへ、つまり、双方向につなぎます。

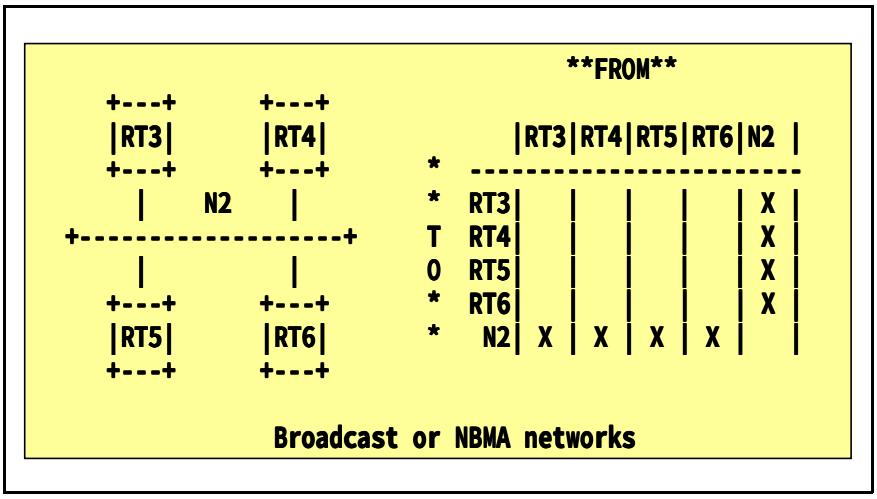

図 9：複数のルータが存在しているマルチアクセスネットワークでの記述例

また、ルータが 1 台しかないスタブなネットワークでは、図 10 のように、ルータからネットワークへの片方向だけが記述されます。

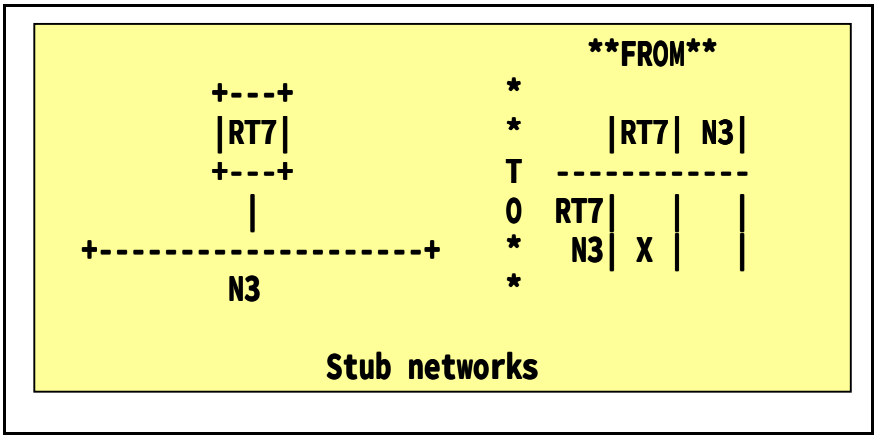

図 10：スタブネットワークでのトポロジカルデータベースの記述例

さらに、2 台のルータがポイントツーポイントで接続されているときには、そのルータ間が図 11 のように双方向に接続されます。このとき、unnumbered の場合はルータのみ、numbered の場合はそのインタフェースが各ルータにスタブなネットワークとして接続されているように記述されます。

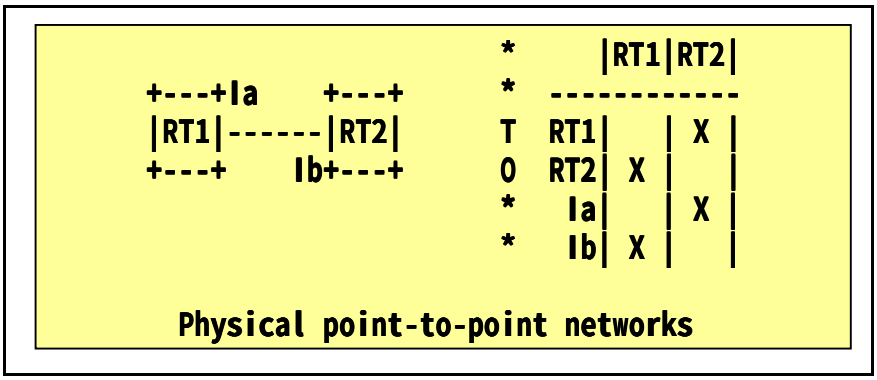

図 11：ポイントツーポイントでのトポロジカルデータベースの記述例

実際のトポロジカルデータベースでは、インタフェースの出力側でのコストが図 12 のように記述されます。このとき、ネットワークからルータに向かうもののコストは、常にゼロ (0) となります。

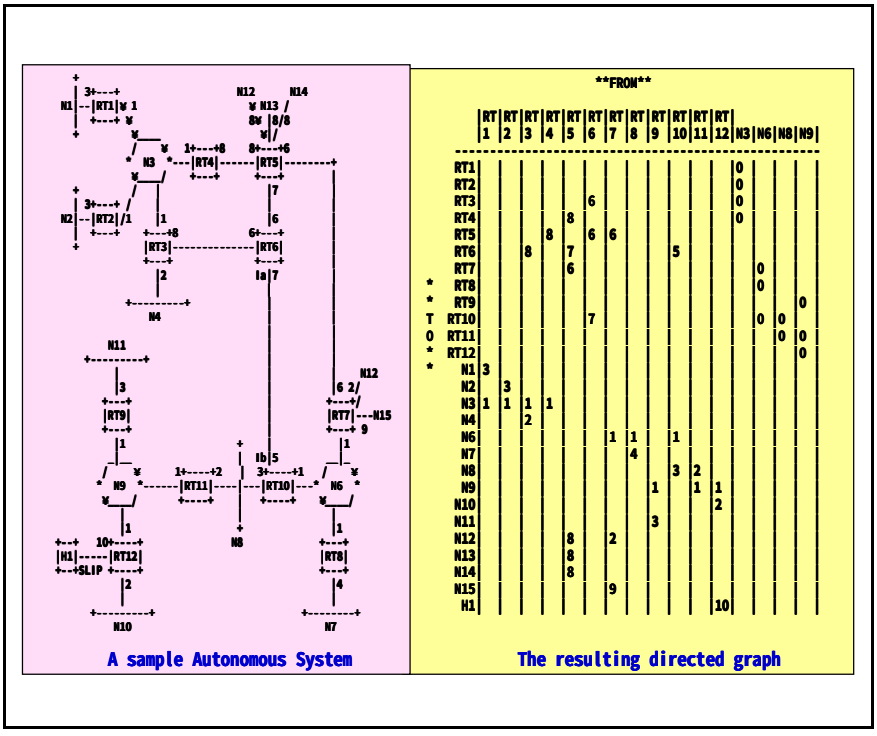

図 12：トポロジカルデータベースの記述例

以上のようなトポロジカルデータベースは、ルータ LSA や、ネットワーク LSA によって作成されますが、次に、その LSA について説明します。

まず、LSA の前に、OSPF のパケットにどのようなものがあるかですが、OSPF では、表 1 に示す 5 種類のパケットが利用されています。

表 1：OSPF のパケットの種類

Type	パケット名
1	HELLO
2	Database Description
3	Link-state Request
4	Link-state Update
5	Link-state Acknowledgment

Type1 の HELLO は、neighbor の検出・維持や、DR や BDR の決定等に利用されます。また、neighbor がダウンしていないかどうかを確認するために、すべてのルータによって 10 秒ごとに HELLO が送信されます。

Type2 の Database Description は、ルータが新たにネットワークに参加したときに、DRのデータベースとの違いをチェックするために利用されます。このとき、LS age（Link-state が作成されてからの時間）を比較して、どちらが最新の情報を保持しているかが判断されます。そして、自分が保持している情報が古い、もしくは情報を持っていない場合は、Type3 の Link-state Request を送信して最新情報を入手します。

Type5 の Link-state Acknowledgment については、次に示す Link-state Update を受信したときに確認のために利用されます。

最後に最も重要な OSPF パケットである Type4 の Link-state Update について説明します。今まで「OSPF ではリンクステートを交換する」と言ってきましたが、Type4 の Link-state Update がそれにあたります。1 つの Link-state Update パケットは、OSPF のヘッダと、それに続く複数の LSA（Link-state Advertisement）によって構成されています。表 2 に、LSA の種類を示します。

表 2：LSA の種類

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

まず、Type1 のルータ LSA はルータの接続情報で、すべてのルータで生成されます。そのルータにどのようなリンクがついているかという情報を持ち、エリア内に伝わります。これによって、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しているかがわかります。

また、Type2 のネットワーク LSA は、ネットワークに接続しているルータのリストで、DR によって作成されます。これもエリア内に伝わります。

Type3 と Type4 のサマリ LSA は、エリア境界ルータによって生成されます。AS 内ではあるがエリアの外にある経路（つまりエリア間経路）の情報を持ちます。このうち Type3 はネットワークへの経路であり、Type4 は AS 境界ルータへの経路となります。

Type5 の AS external LSA は、AS 境界ルータによって生成され、他の AS への経路が記述されます。たとえば、再分配（redistribute）された経路はこれにあたります。

なお、ここで言う AS とは OSPF のルーティングドメインという意味で、一般的に使用される AS とは意味が違うので注意してください。

これらの 5 種類の LSA によって OSPF 内でリンクの状態情報が受け渡されます。ただし、スタブエリアでは Type5 による AS external LSA が受け渡されないため、スタブエリアに AS 境界ルータを設置することはできません。

このような制限を回避するために、RFC1587『The OSPF NSSA Option』では NSSA（Not So Stubby Area：準スタブエリア）を規定しています。NSSA では、Type7 LSA という新たな LSA を利用することで、スタブエリアに AS 境界ルータを配置できるようにしています。この Type7 LSA は、NSSA の AS 境界ルータでしか生成されず、その NSSA 内でしか流れません。そして、NSSA から他のエリアに移るときに、エリア境界ルータによって Type7 LSA を Type5 LSA に変更します。

ここまでに説明したような LSA を交換することによって、先に説明した方法によりトポロジカルデータベースが作成されます。そして、各ルータは、トポロジカルデータベースを基に、SPF（Dijkstra）アルゴリズムによって自分自身をルートとした最短パスツリーを作成します。

たとえば、図 12 に示したトポロジカルデータベースの記述例を基に作成したルータ RT6 の最短パスツリーは、図 13 のようになります。

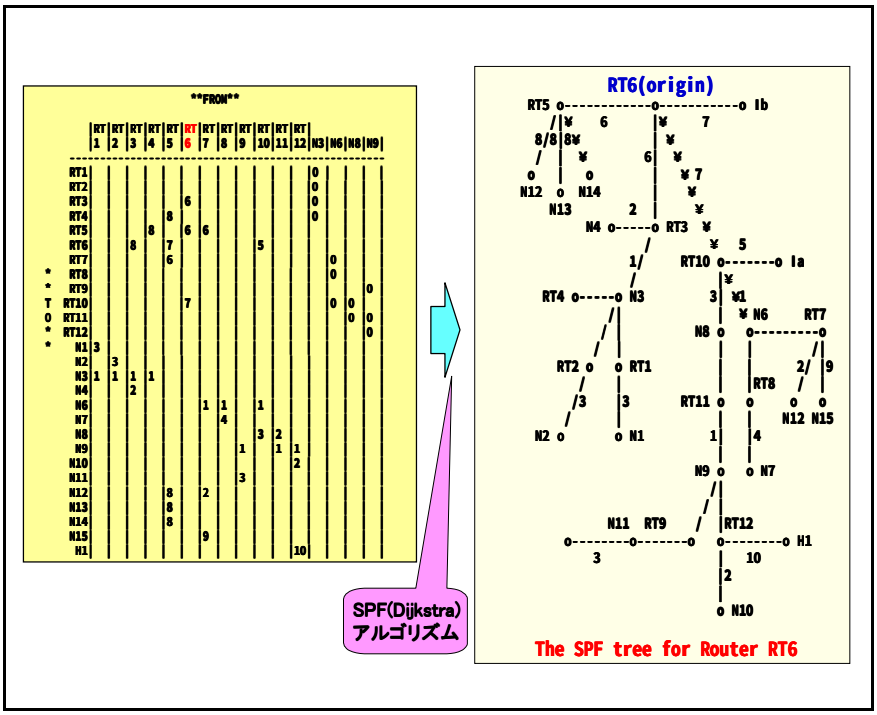

図 13：最短パスツリーの作成例

SPF アルゴリズムでは、自身のルータに近いノードから候補に加え、すべての候補の中からコストが最小なものを確定します。そして、確定したノードの次のノードを候補として加え、そしてまた、すべての候補の中からコストが最小なものを確定します。それを繰り返していきます。

たとえば、RT6 での動作を例で示します。まず、トポロジカルデータベースで、RT6 の次のノードには RT3、RT5、RT10 がありますので、それらを候補のリストに入れます。そして、この 3 つのうちコストが最小値の 6 である RT3 と RT5 を確定します。次に、RT3 の次のノードである RT1、RT2、RT4 と、RT5 の次のノードである RT4 を候補のリストに加えます。RT3 経由の RT4 と、RT5 経由の RT4 は区別されています。そして、これらの候補の中より、コストが最小であるノードとして、RT10、RT1、RT2、そして RT3 経由の RT4 を確定します。RT4 は RT3 経由が確定されたので、RT5 経由の RT4 は候補から外します。そして、候補のリストアップ

と確定を繰り返すことによって、RT6 をルートとする最短パスツリーが作成されます。

このようにして作成した最短パスツリーから、図 14 のようなルーティングテーブルを作成します。

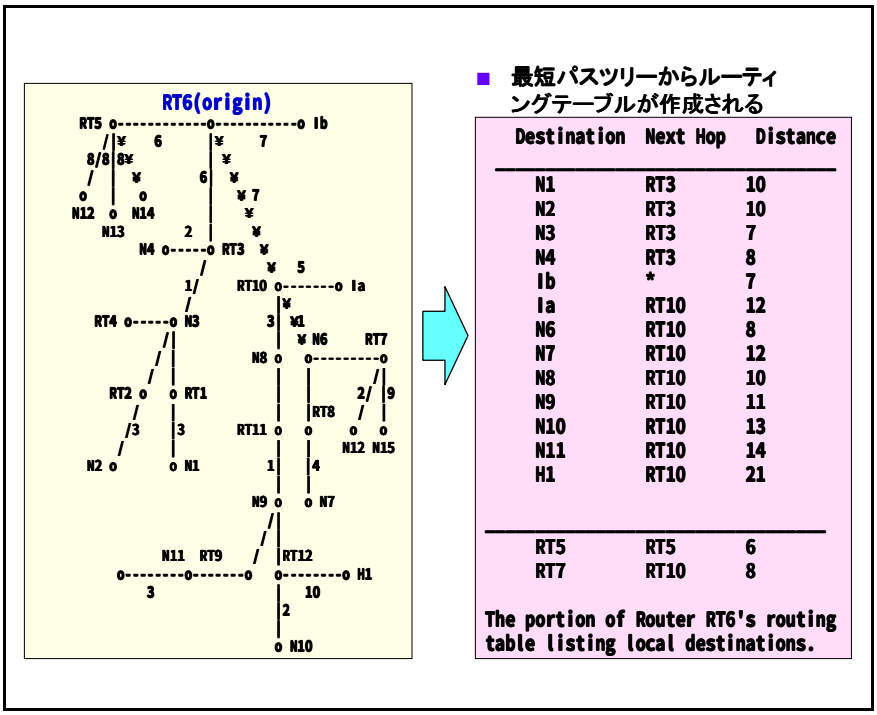

図 14：ルーティングテーブルの作成例

この例はエリアが 1 つの場合でしたが、エリアが複数ある場合は、ルータ LSA とネットワーク LSA だけでなく、サマリ LSA や AS external LSA も加味してトポロジカルデータベースが作成されるため、もう少し複雑になります。しかし、基本的には同様にトポロジカルデータベースが作成され、それを基に SPF アルゴリズムによって最短パスツリーが決定されます。

2.3.3 OSPF の負荷について

前項で説明したルーティングテーブルの作成では、まず、最短パスツリーを作成するために利用される SPF アルゴリズムの負荷が問題となります。

前項の説明で、候補リストはあらかじめコストの低い順に並べてあるとします。すると、コストが最小なノードの選択は、単に先頭のノードを選ぶだけの処理となります。しかし、新たな候補を候補リストに加える際に、低い順に並べるというルールを守りながら、候補を挿入する必要があります。

候補リストのしかるべき位置に挿入する処理の回数は、候補リストに載っているノード数を m とすると O(log(m)) となります。また、よく考えればわかると思いますが、すべてのリンクは必ず一度ずつ調べられています。よって、そのエリアの全リンクの数を l とすると O(l*log(m)) の負荷がかかります。ここで、エリア内のノードの数を n とすると、m は n を超えることがなく、おおよそ一次関数的な関係と言えるため、O(l *log(n)) であると言えます。以上より、あるエリアでの SPF アルゴリズムにかかる負荷は、エリア内のリンク数に比例し、ノード数の log に比例すると言えます。よって、同じノード数であっても、かかる負荷はネットワーク構成にかなり左右されることになります。

さらに、OSPF による実際の網設計で問題となるのは、SPF アルゴリズムだけではありません。経験によると、むしろリンクステートの交換が大きな負荷となっているように見えます。また、十分にメモリを用意していても不安定な状態となってしまうこともよくあります。結局、実装に依存している部分も多く、どこをどうすれば絶対安定するかというはっきりした答えは、誰にも言えないと言うのが実状です。

2.3.4 大規模ネットワークにおける OSPF 設計

以上に述べたように、どの程度の規模のネットワークまで OSPF が耐えられるかは、ルータの機種やメモリ、ネットワーク構成、安定度等によって異なるため、一概には言えません。また、大規模なネットワークを検証することは実際に困難です。よって、基本的に、経験則に頼ることになります。また、J. Moy 著『OSPF Anatomy of Internet Routing Protocol』や Bassam Halabi 著『OSPF DESIGN GUIDE』等も参考になります。

しかしながら、たとえば、よく聞かれる質問に「1 つのエリア内に設置できるルータの台数」があります。このような質問は「一概には言えない」というのが決まり文句ですし、実際その通りですが、それではつまらないので講演者の経験から言うと、Cisco7513 RSP4 256M くらいのルータであれば、100 台程度は十分安定して稼動できると思います。ただし、一切責任は持てません。今までの説明のとおり、ネットワーク構成によってかなり左右されます。

実際は、ルータを増やしていって、たとえば、どこかのリンクをシャットダウンしたとき等に、増設する前に比べてコンバージェンス時間（CPU が落ち着くまでの時間）が明らかに大きくなったら、そろそろ限界だと思うべきでしょう。これは、それなりに注意して運用していれば必ずわかります。たとえば、トラフィックが非常に多く、ただでさえ負荷の重いルータに注意したり、性能の低いルータに注意したりすることが必要になります。

また、そろそろ限界だと思ったら、機器をアップグレードする、網変更する等の、何かしらの対策（5 章、6 章の事例参照）をとる必要があります。

また、その他にも OSPF を利用した大規模ネットワークを設計する上でのいくつかの Tips を示します。

- リンク数

  多数のリンクが存在するネットワーク構成はあまり良いとは言えません。たとえば、ATM スイッチを使い PVC をフルメッシュで張ったりするよりは、マルチアクセスのスイッチング機器を利用したほうが、規模対応性に関してははるかに良いと言えます。

- メモリ

  メモリが多いに越したことはありませんが、メモリ量が十分であるからといって安心してはいけません。メモリ量が足りていても、不安定になることはよくあります。

- DR と BDR

  DR は高負荷となるため、処理能力の高いルータ、または他の処理負荷があまりかからないルータを選択する必要があります。また、1 つのルータが複数のネットワークの DR とならないように注意する必要があります。

- ループバックアドレス

  ルータ ID を安定させるために、ループバックアドレスを設定するようにします。

- エリア

  エリア 0 を中心として、ネットワークを拡張していくようにします。また、冗長性を確保するために、1 つのエリアに複数のエリア境界ルータを設置するようにします。また、各エリア境界ルータが担当するエリアはなるべく 2 つまでとします。さらに、バーチャルリンクに依存した設計は避けるようにします。

- 経路数

  できる限り経路が集約できるように、IP アドレスを割り当てるようにします。

- デフォルトルート

  デフォルトルートを適切に利用するようにします。また、BGP からの経路を OSPF に再分配しないようにします。

また、このような項目を考慮してネットワークを設計するのに加えて、さらに重要なことは、危ないと感じたときにどのように対処するかです。これには、まずは機器の性能をアップグレードすることがあります。Cisco7513 なら、RSP2 から RSP4 に変えると劇的に変わります。また、大容量のルータに交換して処理能力を上げることに加え、ノード数やリンク数を減らすのも 1 つの手です。しかしながら、上記の対処が困難なことや、それだけでは不十分なことがあるかと思います。その場合は、5 章や 6 章に挙げる工夫をすれば良いことになります。static-to-bgp（5 章）や、confederation（6 章）に限らず、OSPF プロセス分け、IS-IS 化、virtual link の積極利用や、confederation ではなく OSPF を分けて別のプロトコルで結ぶ等々、自分のネットワークに合った工夫を行えば良いわけです。

2.4 IS-IS

米国の大手 ISP では、OSPF ではなく IS-IS（Intermediate System-to-Intermediate System）を利用しているところが多いようです。また、IS-IS は、OSPF よりもスケーラビリティがあるとも言われています。この IS-IS には、次のような特徴があります。

- リンクステートアルゴリズムを利用した OSI スタックのルーティングプロトコルである。
- レベル 1 とレベル 2 の、2 つの階層がある。
- コストベースのルーティングプロトコルである。
- NSAP アドレスを使用している。

また、IS-IS には DR の仕組みも存在し、VLSM にも対応しています。

図 15 に、ネットワークの構成例を示します。

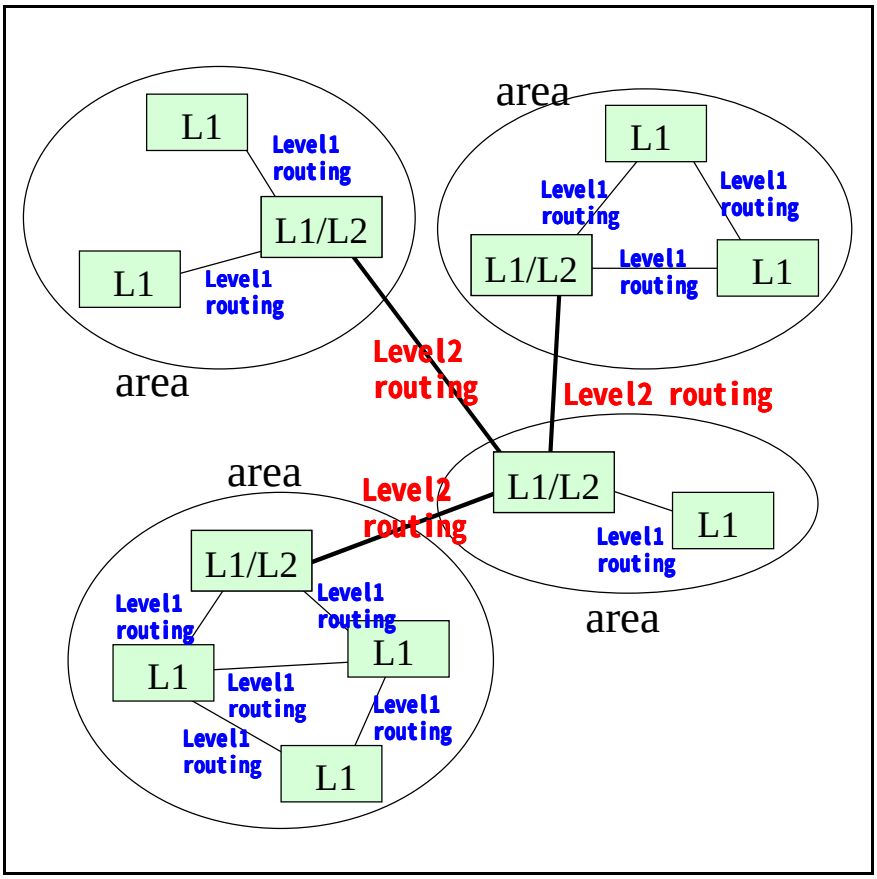

図 15：IS-IS でのネットワーク構成例

この構成例からわかるように、IS-IS にもエリアの概念があります。ただし、各エリアの境界は、OSPF とは異なりリンクとなっています。また、エリア内のルーティングを「レベル 1 ルーティング」と呼び、エリア間のルーティングを「レベル 2 ルーティング」と呼んでいます。そして、各ルーティングを担当するルータをそれぞれ、「レベル 1 ルータ」と「レベル 2 ルータ」と呼びます。また、レベル 2 ルータには自エリア内でレベル 1 ルータとしても機能するものもあり、このようなルータは「レベル 1/ レベル 2 ルータ」と呼ばれています。

表 3 に、TCP/IP と OSI での違いを示します。

表 3：TCP/IP と OSI の違い

プロトコルスタック	TCP/IP	OSI
アドレスや伝送の仕組み	IP	CLNS
アドレス	IP アドレス	NSAP アドレス

ここに示した CLNS（Connectionless Network Service）とは、IP と同様に OSI でのアドレスや伝送の仕組みです。また、NSAP（Network Service Access Point）アドレスは、CLNS で利用されるアドレスです。

IS-IS の LSP（Link State PDU）は、OSI のノード間のやりとりとして認識され、CLNS によって実行されます。このため、各ルータは、OSPF でのルータ ID と同じ役割を果たし、NSAP アドレスで表現される NET を持つ必要があります。そして、IS-IS では、次のような順序でルーティングテーブルが作成されます。

1. CLNS によって IS-IS のやりとりを実行し、データベースを作成する。
2. NET に基づいたツリーを作成する。
3. IP および CLNS のルーティングテーブルを作成する。

IS-IS では、SPF アルゴリズムによる最短パスツリーの作成が、レベル 1 ルーティングとレベル 2 ルーティングで個別に実施されます。さらに、最短パスツリーは、OSPF ではルータ ID を基に作成されるのに対して、IS-IS では NET を基に作成されます。ただし、このような違いは、IS-IS と OSPF 間での本質的な違いとはなりません。

また、IS-IS では、各エリアの境界がルータ間となりますが、図 16 のように、エリア間のルーティングを担当するレベル 2 ルータを OSPF のエリア境界ルータと考えることで、両者の間に本質的な違いはなくなると考えられます。

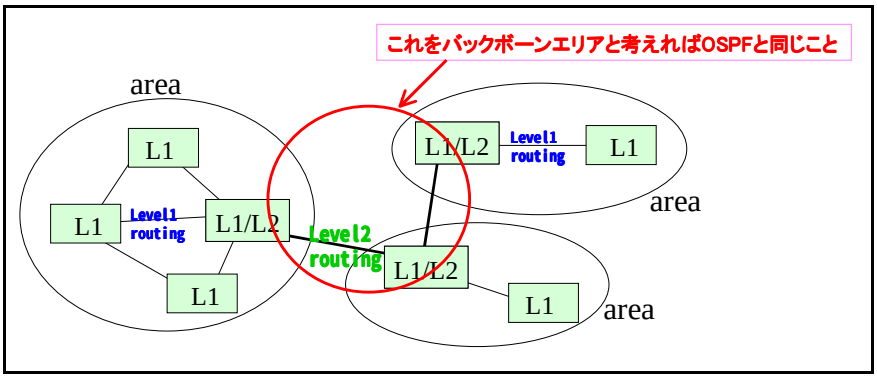

図 16：IS-IS と OSPF

このようなことから、もちろん実装の点では何らかの差があるかもしれませんが、少なくともプロトコル的に比べると、IS-IS と OSPF では規模対応性に大きな違いは見あたりません。実際、米国 ISP でも、「IS-IS のほうが規模対応性がある」という理由ではなく、「以前から IS-IS を利用し、使い慣れている」という理由から、現在も IS-IS を利用しているところもあります。

これに対して、日本では、IS-IS に精通している人も少なく、ドキュメントもほとんどありません。また、現在のネットワーク構成を IS-IS に変更することも困難です。IS-IS による大きなメリットが見あたるわけではないので、これまでと同様に OSPF を IGP として利用していったほうが良いと言えるでしょう。

3 BGP によるシステム設計論

まず、OSPF（Open Shortest Path First）と BGP（Boarder Gateway Protocol）を比較してみます。

表 4：OSPF と BGP の比較

OSPF	BGP
リンクステートアルゴリズムを利用したプロトコルであり、状態変更ごとに LSA が連鎖伝播される。	パスベクターアルゴリズムを利用したプロトコルであり、状態変更ごとに UPDATE が連鎖伝播される。
トポロジの管理に主眼が置かれ、エリア内で共通のトポロジカルデータベースが作成される。それを基に各ルータが個別にパスツリーを作成する。	プリフィックス（ネットワーク）の有効 / 無効とパス属性に着目し、受領した UPDATE が各 AS やルータのポリシーに基づいて処理され、以遠伝播される。
あるネットワーク（ルータ）の状態の変更によって、全ルータのパスツリー再作成を発生する。また、30 分ごとにリフレッシュが実施される。	あるネットワークの状態の変更は、基本的にそのプリフィックスだけの問題でしかない。また、リフレッシュも実施されない。
基本的には、OSPF を起動しているすべての隣接ルータと経路情報を交換する。	明示的に定義された隣接ルータとのみ経路情報を交換する。
経路ごとにはポリシーを付加できない。	経路ごとにポリシーを付加できる。

BGP には、AS（Autonomous System）という概念があります。AS とは、単一のルーティングポリシーが適用される範囲であり、1 つの ISP と考えることができます。また、AS は、1 〜 65535 の 16 ビットで表される番号空間を持っています。たとえば、BIGLOBE は AS2518 であり、OCN は AS4713 です。このような AS 番号のうち、64512 〜 65535 はプライベート AS として利用することができます。

まず、図 17 のような階層型経路制御について考えてみます。

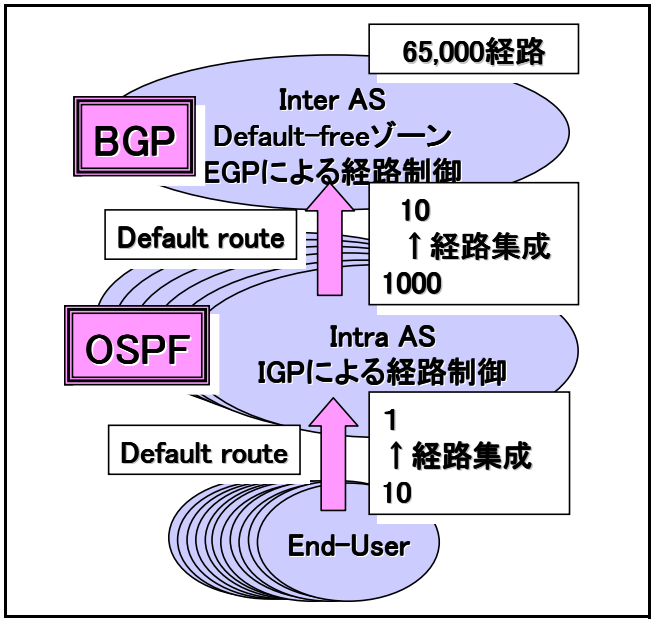

図 17：インターネットでの階層的な経路制御

このような階層型経路制御は、デフォルトでは制御されることなく、すべての経路情報を保持する Inter AS と、IGP によって経路制御される Intra AS と、End User の 3 階層に分かれます。また、このような階層では、各境界で経路が集約されています。このような経路制御における、もっとも単純な BGP の導入は図 18 のようなものとなります。

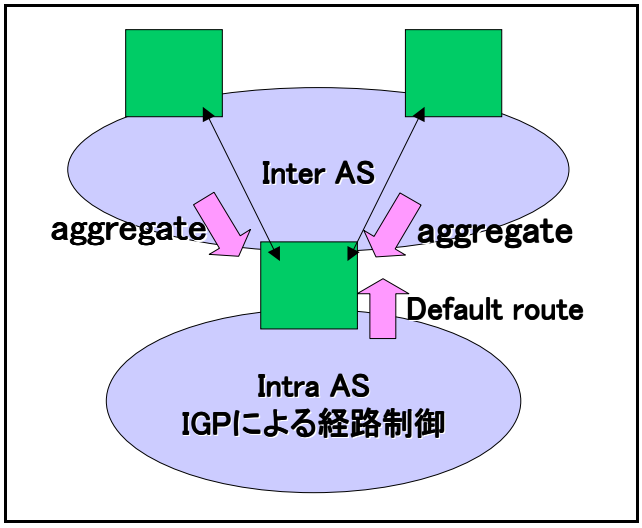

図 18：単純な BGP の導入

図 18 は単純で分かりやすいモデルではありますが、この境界ルータが single point of failure（単一障害点）となるという意味で問題があります。そこで、境界ルータを 2 つ以上持ち、複数箇所で他の AS と接続したいという要求が出てきます。境界ルータを 2 つとするモデルが図 19 です。このとき、デフォルトルートは 2 つ存在することになりますが、IGP によって近いほうの境界ルータが採用されます。そして、各境界ルータの間では、iBGP（internal BGP）を確立して、経路情報を同期させるようにします。

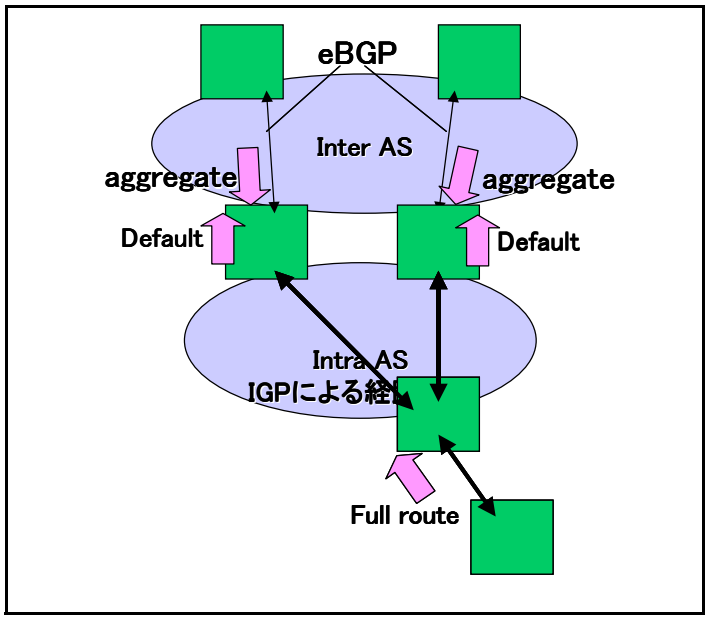

図 19：2 つの境界ルータによる問題解決

3.1 iBGP や eBGP 利用での問題点

2 つ以上境界ルータがある場合、一般的にはある特定の対地に対して、どちらの境界ルータからも到達可能です。どちらを主に利用するかに関しては、BGP のパス属性によって定まるか、自分自身のポリシーで定めるかのいずれかによって決まり、その情報は iBGP によって AS 内で同期が取れています。また、iBGP は境界ルータが隣接しておらず、間に BGP を起動していないルータがある場合にも確立可能ですので、図 20 のような問題が起こりえます。

図 20 において、緑（大きな□）は BGP を起動している境界ルータ、黄色（小さな□）は BGP を起動していないルータです。たとえば、左側の黄色ルータから net N 宛てにパケットを転送しようとする場合、黄色ルータはフルルートを持たないので net N を知らず、デフォルトに従って直近の（左の）境界ルータに転送します。しかし、左の境界ルータにおいては BGP によって、net N に対しては右の境界ルータからパケットを外に出そうとしており、黄色ルータにパケットを戻そうとします。この場合、net N に対しては永遠にパケットが到達せず、経路制御としては問題があります。

このような状況に陥らないように、Cisco Systems 社のルータでは、IGP において知らない経路は、BGP で他から知り得ても、フォワーディングテーブルに反映しないようにデフォルトでは動作します。このしくみを IGP Synchronization（IGP 同期）と呼びます。

しかし、この IGP Synchronization のしくみは、結局 IGP でもフルルートを持たなければならないということを意味し、規模対応性において著しい不利となります。

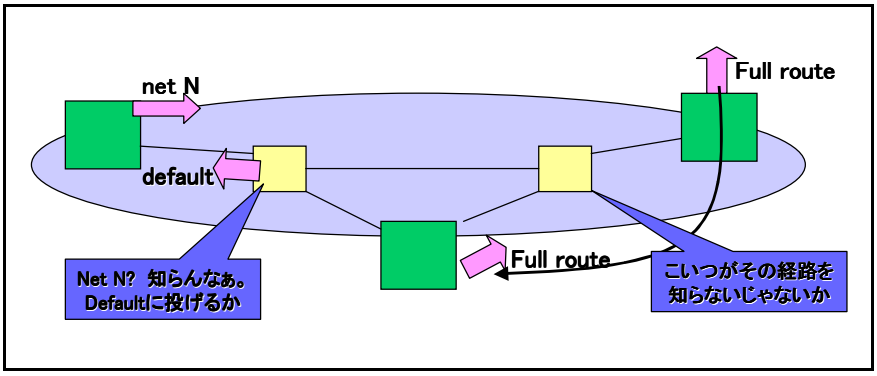

図 20：iBGP 内でのフルルートの問題

この問題を解決させるためには、Cisco Systems 社のルータが持っている IGP synchronization を適用させないようにするコマンドである、no synchronization を利用するとともに、iBGP はすべて隣接して配置するようにネットワークを設計します。つまり、AS の中を、BGP を起動している境界ルータのみからなるトランジット層と、それ以外のアクセス層に分けるという設計となります。

これによって AS の中も階層的な設計が導入されることになります。

また、ある iBGP が得た経路は同位の他 iBGP に再伝播されないため、すべてのノードをメッシュ状に接続しなければなりません。たとえば、五つの境界ルータが存在していたときには 10 個の接続が必要であり、境界

ルータが 10 個となったときには 45 もの接続が必要となります。この問題は、次の 2 つの方法で解決できます。

- iBGP ルートリフレクタ
- BGP コンフェデレーション

このうち iBGP ルートリフレクタによる解決方法では、各 iBGP をリフレクタとリフレクタクライアントの 2 種類に分け、リフレクタが得た経路をクライアントに再分配するようにします。また、BGP コンフェデレーションによる解決方法では、AS 内をさらに小さなサブ AS に分割し、各サブ AS 間を eBGP（external BGP）として接続するようにします。これによって、すべての iBGP に対して接続する必要がなくなります。

次に、BGP のスケーラビリティに対する問題解決方法の例を示します。

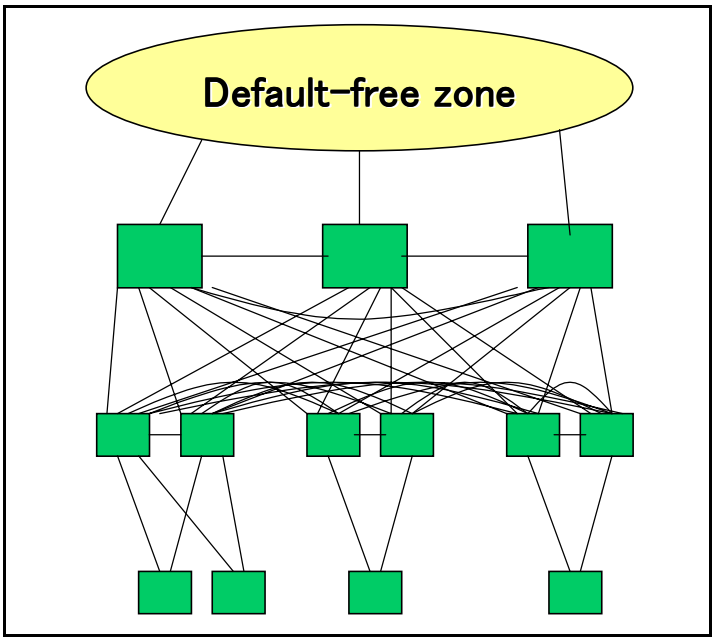

図 21：AS 内での BGP のスケーラビリティに関する実際の問題

ここでは、次のような問題が存在しているものとします。

- 複数の対外接続が存在する。
- 地域や POP ごとに BGP 接続加入者が存在し、それぞれに BGP ノードが必要となる。
- POP にコアルータを 2 台用意して、冗長性を確保する必要がある。
- BGP 加入者が増加することで、BGP 加入者収容ルータが増加する。

このような問題は、POP コアルータに対するものと、加入者集線ルータに対するものの 2 種類のルートリフレクタを階層化することで対応できます。また、地域ごとや POP ごとにサブ AS を設定し、BGP 加入者の主要ルータ間に iBGP を設定するコンフェデレーションによっても、この問題に対応することができます。

次に、eBGP のスケーラビリティに関係する項目を示します。

- 経路数
- ピア数
- ルートフラッピング
- ポリシー変更の反映

このうち、経路数は現在の最大経路数が約 65,000 と膨大な数となっているため、要求されるメモリ量が 64 メガバイトでは不十分であり、128 メガバイトを確保する必要が出てきています。また、IX（Internet eXchange）で多数のピアと接続したときには、必要とされるメモリ量が増加してしまいます。

不安定なリンク等によって経路広告が不安定となる「ルートフラッピング」が発生すると、経路の更新や消去が連続して発生するため、CPU 資源が浪費されてしまいます。この問題は、不安定となっている経路を一定時間ルーティングテーブルから削除することで対処できます。この方法を、flap dampening（フラップダンプニング、経路抑制）と呼びます。

また、eBGP でのポリシー変更では、リセットによるピアのクリアが必要となり、アップストリームからフルルートすべてを受けなおす負荷が生じます。この問題に対して Cisco Systems 社では、セッションのクリアなしに、経路に対して新しいポリシーを反映できるようにしています。これを soft-reconfiguration と呼びます。

3.2 ルーティングポリシーの実現

既に示したように、BGP ではルーティングポリシーが設定できます。BGP では、プリフィックスとパス属性によって経路情報が記述されます。このパス属性の値を調整することで経路選択を制御し、ルーティングポリシーを反映できます。このときのルーティングポリシーとは、複数ピアを持つ AS とのトラフィックの交換方法、セキュリティのための経路のフィルタリング、複数のアップストリームに対するトラフィックバランス等となります。

実際には、次のパス属性の値を変更することで、ルーティングポリシーを実現できます。

- Local Preference
- AS_PATH
- MED
- コミュニティ

Local Preference を使用することで、ネットワーク設計者の意図に従った優先順位を付けられます。たとえば、図 22 のようなネットワークでは、通常は AS_PATH が短い AS20002 が利用されますが、Local Preference によって AS20004 を優先させることができます。

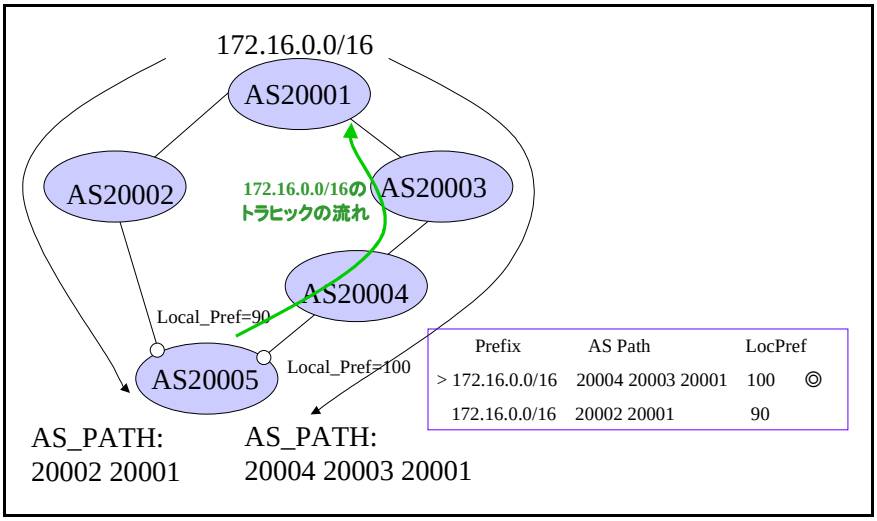

図 22：Local Preference の利用

また、as-path prepend というコマンドを使って AS_PATH の長さを実際よりも長く見せかけることで、選択されるパスを制御することも可能です。たとえば、図 23 の例では、as-path prepend によって自らの AS を追加し、AS20002 を優先させています。

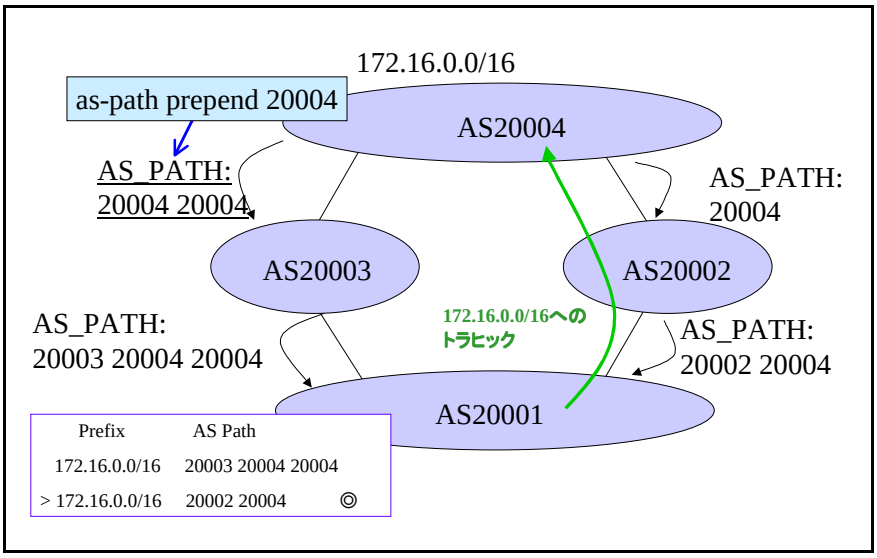

図 23：AS_PATH の利用

MED を利用することで、同一 AS に対して隣接する複数ピアの優先度を指定できます。たとえば、図 24 の例では、大阪ルータに対して 100、東京ルータに対して 200 がそれぞれ指定されているため、MED の指定値が小さい大阪ルータが優先されるようになります。

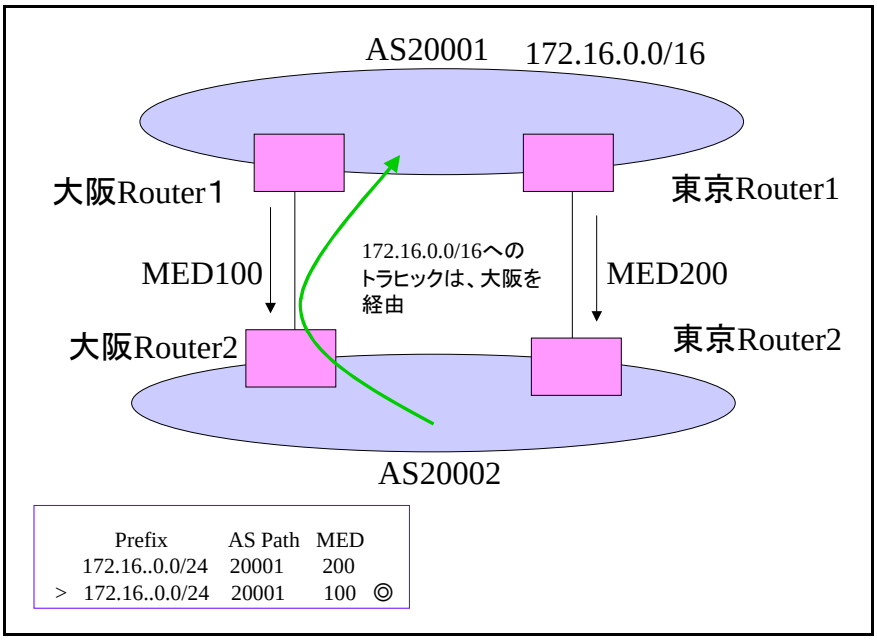

図 24：MED の利用

ここまでに示した方法とは異なり、コミュニティは数値自体にはプロトコル上の意味はありませんが、その数値を受け取った AS やルータに何らかの作用を及ぼす 32 ビットの整数値です。通常、コミュニティは、上位 16 ビットが対象 AS を表し、下位 16 ビットがその AS での動作を指示している「New Format」と呼ばれる書式で記述されます。

たとえば、MCI（現 C&W ）の場合、下位 16 ビットにあらかじめ指定されたいくつかの数値を持つコミュニティ属性値（たとえば、3561:70）を持った経路情報に関して、Local Preference をその数値に指定します。これによって、コミュニティを付加した元の AS への戻りのトラフィックの制御が可能となっています。この他、ISP によっては AS prepend を付加するコミュニティ属性値を準備する等、さまざまなパターンがあるようです。

このようなパス属性値は、同一プレフィックスの経路情報が複数存在しているときに、最適経路を決定するために利用されます。たとえば、Cisco Systems 社の製品では、次のような順序で最適経路が決定されます。

1. Local Preference が大きい。
2. AS_PATH が短い。
3. MED が小さい。
4. IGP 上での次ホップが近い。
5. BGP のルータ ID が小さい。

このように、BGP では経路情報に付加する属性値によってトラフィックのコントロールが可能ですが、一般的な ISP やエンドユーザサイトの場合、サーバよりクライアントのほうが多いため、サーバからのコンテンツを受け取る方向、つまり外部から受け取るトラフィックのコントロールのほうが重要であることに注意してください。

4 文献紹介：Jessica Yu による Internet Draft『Scalable Routing Design Principles：規模対応性の高い経路制御設計の指針』

ここでは、Jessica Yu によって記述された『Scalable Routing Design Principles：規模対応性の高い経路制御設計の指針』（Internet Draft：draft-yu-routing-scaling-01）を紹介します。本書では、大規模ネットワークの経路制御システムにおける問題点が概説され、設計上の指針が示されています。また、本書の邦訳は、（株）インターネットイニシアティブの近藤邦昭氏とともに友近剛史と前村昌紀が行い、*http://www.janog.gr.jp/doc/draft-yu-routing-scaling-01-j.txt* として公開しています。

まず、本書では、経路制御設計の一般的な目的として、次の項目を挙げています。

- スケーラビリティが高いこと。
- 冗長性があり強靱であること。
- 障害発生等に対する収束時間が妥当であること。
- 経路情報が完全であること。
- 経路制御ポリシーが実用的で管理可能なこと。

そして、本書では現在の大規模ネットワークの特徴として、次のような米国 Tier1 の現状が示されています。

- 数百ノード、数千ユーザのほとんどが BGP 接続
- 冗長性を確保するための複雑なトポロジー
- 現状 65,000 にも及ぶフルルートの伝播
- 集線ルータへの数百ユーザの接続

このような大規模ネットワークでの一般的な問題点は、ルータの資源消費、BGP 処理上での問題、IGP 処理に関するものとされています。このうち、ルータの資源消費に関する問題として、経路数の過多や不安定なネットワークによるフラッピングが挙げられています。また、BGP の処理では、デフォルトフリーな状態となっている IX とのピアによって経路過多となりメモリを圧迫することと、プリフィックスのフィルタリング、顧客集線ルータでのピア過多が問題とされています。さらに、IGP では、大規模ネットワークでの巨大なトポロジカルデータベースが問題とされています。

また、IGP に関する問題としては、次の項目が挙げられています。

- トポロジカルデータベースの肥大化
- フラッディング
- 経路計算の複雑化
- 過負荷の悪循環

このうち一斉広告であるフラッディングは､既に示したように 1 台のルータの状態変化がすべてのルータへの LSA 伝播となるものです。また、過負荷によって HELLO パケットを受け損なったときに LSA が発生し、復旧したときにも再度 LSA が発生するため、過負荷の悪循環が発生するとしています。

さらに、本書では、BGP に関する問題として次の項目が挙げられています。

- iBGP のフルメッシュ化の問題
- フラッピングによる更新と失効の繰り返し
- プリフィックスフィルタリング

このような大規模ネットワークでの問題点を解決し、スケーラビリティを確保するために、本書では次の項目を挙げています。

- 上下分割による階層構造化
- 左右分割による区画化
- 適切なトレードオフの設定
- 経路制御処理の負担の軽減
- スケーラブルな経路制御ポリシーの実装
- ルートサーバを利用した out-of-band 経路処理

大規模ネットワークでは、単一階層でフルメッシュな構成は拡張性を損ないます。このため、本書では、階層を Transit Core Network と Access Network の 2 階層に分けることで、構造が単純化し管理しやすくなるとしています。具体的には、OSPF のバックボーンエリアとその他のエリア、IS-IS のレベル 1 とレベル 2、iBGP でのルートリフレクタの階層化等の実装方法が示されています。

また、本書では、階層構造化での 2 層目を区画化することによってもスケーラビリティを確保できるとしています。これによって、問題や障害を局所化でき、セグメントごとに処置できるようになります。また、区画化によって経路も集成されます。具体的には、BGP でのコンフェデレーションによる IGP ドメインの分割等の実装方法が示されています。

さらに、本書では、過度の冗長性によってスケーラビリティが損なわれるとしています。また、収束性と安定性の間にもトレードオフが存在することも示しています。このようなトレードオフを適切に設定することで、スケーラビリティが確保できます。

経路制御処理の負担は、次の方法で軽減できるとされています。

- ルートサーバを利用した out-of-band 経路制御
- 経路情報の削減

このうち経路情報は、適切な集約や要約を実施し、できる限りデフォルトルートを利用し、過度な冗長構成としないことで削減できるとしています。また、スケーラブルな経路制御ポリシーは、次のような方法で実装するようにすべきだとしています。

- 要件を満たす範囲で、可能な限りポリシーを簡素にする。
- 間違いが起こりやすい手作業を避け、可能な限り自動化する。
- 経路制御の完全性のためにプリフィックスによる経路フィルタリングを実施することを除き、プリフィックスごとのポリシーは可能な限り避ける。
- 例外を作ることを避ける。
- 可能であれば out-of-band による経路制御ポリシープロセスを利用する。

また、out-of-band 経路処理では、ルーティングとフォワーディングというルータが持つ 2 つの機能のうち、経路選択、ポリシー処理、ルーティングテーブルの完全性の維持等、ルーティングについてはルートサーバを利用するようにすべきだとしています。そして、ルートサーバによって作成されたルーティングテーブルをルータに供給し、経路制御を実施する必要がなくなったルータがパケットの転送を担当することで、スケーラビリティを確保できるとしています。

これまでに示してきた内容が記述されている本書は、経路制御のスケーラビリティについて非常に良くまとめられている名著だと思います。ただし、BGP 加入者数の多さ等、米国 Tier1 と日本の環境の違いには注意する必要があります。また、ルートサーバの導入はアイディアとしては良いものの、実際に利用する上では安定性等の点で問題が多く、一般的でないことにも注意が必要です。階層構造化や区画化等は以下に紹介するとおり、既に実施している問題解決方法ですが、この方法が適切であることを改めて認識させてくれます。

5 事例紹介：スタティック経路の BGP への再分配

OCN（AS4713）では、OSPF で扱う経路数が非常に増えていました。また、OCN のサービスの性格から、経路の 90 パーセント以上が、顧客へのスタティックな経路をOSPFに再分配したexternal経路となっていました。また、JPNIC から割り振られるアドレスブロックは、その 80 パーセントを消費しなければ次のアドレスブロックを割り振られないこともあり、効率よくアドレス集約ができていませんでした。疑似環境で検証してみたところ、OSPF では external 経路と言えども、ある程度以上の external 経路が流れると、メモリは十分足りているにもかかわらず、ネットワークが一度不安定になると、なかなか安定した状態に戻らないことが確認できました。このようなことから、OSPF の経路数増大に対応しなければならないと考えていました。

次に、ネットワークの特徴と条件について述べます。まず、バックボーンネットワークに関しては完全に２重化されており、それを有効に使用するため、さまざまな箇所でトラフィックのロードバランス（ロードシェアリング）がされています。また、なるべくサービスを停止せずに設定を変更する必要があります。最後に、これはかなり重要なことですが、運用の手順の変更を極力少なくしなければなりませんでした。このような条件を踏まえながら、OSPF の経路数増大に対処する必要がありました。

OSPF の経路数削減の方法としては次のようなものが考えられます。

- OSPF をリンク部分で分割する。

  つまりコンフェデレーション等ということです。しかし、図 25 のようなネットワークでは、BGP 自体ではロードバランスしないため、1 つ手前のルータでバランスさせる必要があります。このため、特別で複雑な設定になってしまい、設定や運用の大幅な変更が必要となります。適切にネットワークが運用されていくためには、あまり特別な設定や複雑な設定をするべきではありません。ネットワークだけでなく、運用のスケーラビリティがとれなくなるからです。よって、この方法は見送りました。

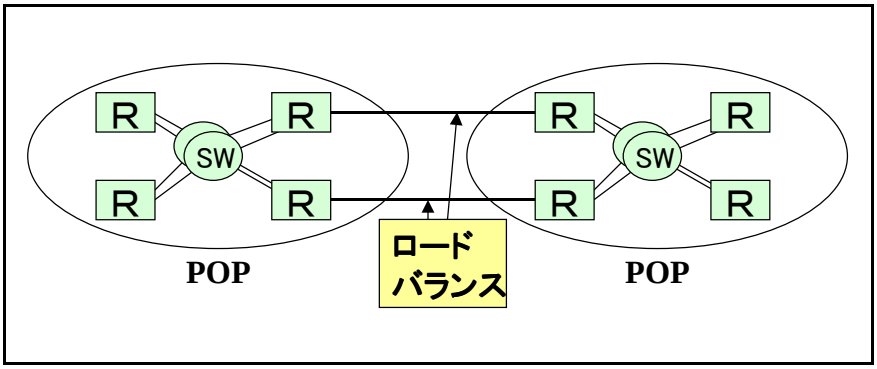

図 25：OSPF のリンク部分での分割

- OSPF から IS-IS に変更する。

  これは設計、運用ノウハウが不足しており、実際の効果も不明であったため、やはり見送りました。

その他、いろいろありましたが、どれも決め手に欠きました。ある日考えついたのが次の方法です。

- スタティックな経路を直接 iBGP に再分配する。

この 3 番目の方法は、static 経路を OSPF ではなく直接 iBGP に再分配する方法です。ループバックインタフェース等の内部情報については OSPF を利用し、外部経路に対して BGP を利用します。つまり、OSPF はトポロジーだけ理解してもらうということです。この方法ですと、ロードバランスや、サービスの継続、運用変更の少なさ等の条件を満たし、OSPF の経路数も削減することができます。BGP は OSPF に比べ、経路数に関するスケーラビリティも高いため、この方法でいけると思いました。後でわかったのですが、この考え方はアメリカの大手 ISP でも採られている方法のようです。ただ、このときは講演者が考えついたものでした。この方法の前提としては、次のようなものが挙げられます。

- iBGP セッションは当然（元々）ループバックアドレス同士である。
- ルータのループバックアドレス等の internal 経路については、当然（元々）OSPF に流れている。
- スタティックを設定しているルータでも BGP を話す。

しくみとしては、図 26 のようになっています。

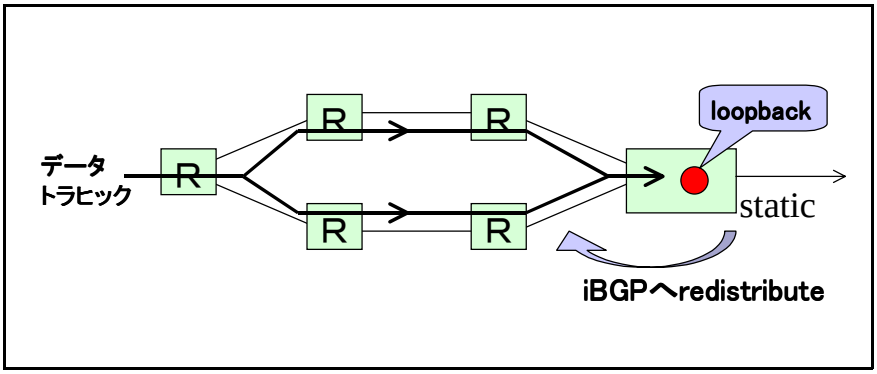

図 26：スタティック経路の BGP への再分配

図 26 を説明すると、まず、ある経路にパケットが行くためには、BGP next-hop である、再分配（redistribute）したルータのループバックアドレスに向かおうとします。その後、その next-hop へ向けて、OSPF で作成されたルーティングテーブルを参照します。つまり、再帰的にルーティングテーブルをルックアップ（recursive lookup）します。よって、OSPF のおかげでロードバランスするわけです。

また、図 27 のように、このような経路に no-export のコミュニティ（community）を付けることによって、specific な経路が AS 外部に流れないようにしています。

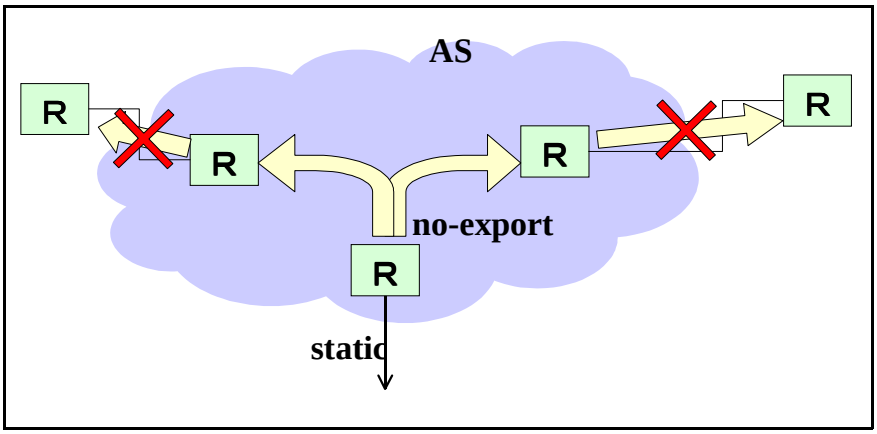

図 27：no-export によるトラフィックの制御

また、BGP を話すルータが多くなることにより、ルートリフレクタの負荷が高まることが懸念されますが、図 28 のようにルートリフレクタを階層化することで、各リフレクタの負荷を軽減しました。また、フルルートを必要としない箇所ではフィルタリングを実施しました。

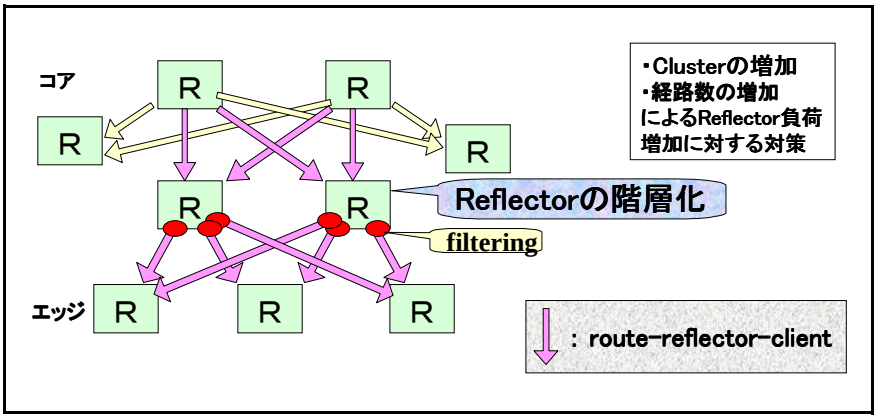

図 28：ルートリフレクタによる階層化

実際にこのような方法を用いることによって、AS4713 の内部ルーティングの安定性が増しました。また、トラフィックの流れ方を変えず、サービス断もほとんどなく、さらに運用手順もほとんど変化がなく実行できました。

最後に結論ですが、一般的にも、static 経路はもはや IGP ではなく iBGP に流し、IGP はトポロジーの情報を持つだけでよいと言えると思いますので、皆さんもよかったらこの方法を参考にしてください。

6 事例紹介：コンフェデレーションの応用

ルートリフレクタと同様にコンフェデレーションも、iBGP のフルメッシュ化を解決するための手法として利用されます。BGP のコンフェデレーションでは、異なる AS ナンバーを持つ複数の BGP スピーカを、外部からは単一の AS 番号として見せかけることができます。この例では、この機能を IGP のスケーラビリティに利用しています。図 29 のように、1 つの AS をサブ AS に分割し、サブ AS ごとに異なる IGP プロセスとして起動するようにします。

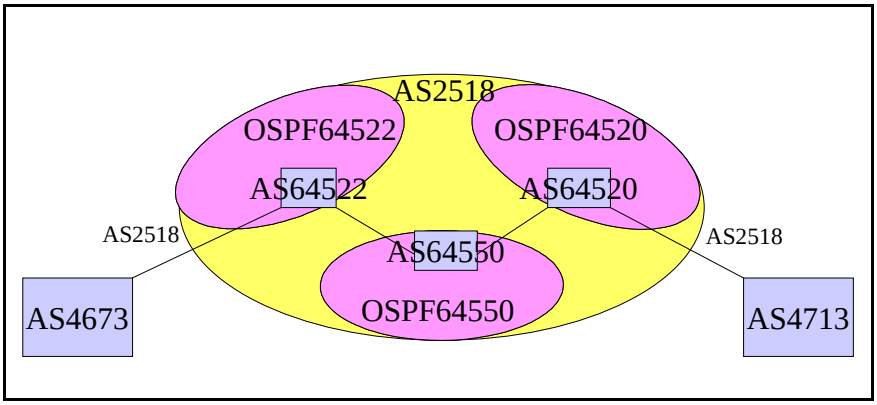

図 29：コンフェデレーションによる IGP のスケーラビリティの確保

実際には、次のように記述してコンフェデレーションを実行します。

```
router bgp 64551
      bgp confederation identifier 2518
      bgp confederation peers 64520 64521 ...
```

ここで、1 行目に記述した 64551 はプライベート AS であり、2 行目に記述した 2518 は外部に示す AS です。また、64520 や 64521 は、コンフェデレーションによるサブ AS です。

BIGLOBE では、OSPF プロセスの肥大化によって、アクセスポイントに設置した小さなルータのメモリ不足や、経路制御不全が懸念されていました。このため、OSPF プロセスを小さくし、肥大化し始めたときには分割できるネットワーク構成に移行することにしました。

コンフェデレーション内での経路については、サブ AS 間は eBGP とし、サブ AS 内では iBGP を利用することができます。また、Local Preference、MED、NexHop は、iBGP としての扱いとなり、サブ AS を跨って保存できます。さらに、コンフェデレーション内のサブ AS は、AS_PATH では認識されますが、ホップ数の評価には利用されません。

BIGLOBE での経路制御設計では、サブ AS ボーダルータにおける BGP 広告は OSPF を再分配（redistribute）して生成し、サブ AS ごとに集成しています。AS としての集成経路は中央のルータ 2 台で生成し、それ以外の細かな経路情報は対外接続ルータでフィルタリングするようにしています。

Internet Routing Architecture におけるコンフェデレーションの推奨デザインは、中央の集権的サブ AS に対してスタブなサブ AS が接続されるようなデザインですが、それとは大きく異なる BIGLOBE のサブ AS デザインでも問題なく動作しています。また、対外接続ルータでは 1 台のルータに 1 つの AS 番号を割り当て、OSPF を起動せずに BGP の処理だけに専念させるようにしています。

BIGLOBEのネットワークは、このコンフェデレーション化によってOSPFの動作が安定し、それ以降、OSPF の問題に悩まされることはなくなりました。

ただし、コンフェデレーションによる IGP の分割において、次のような点が不便だと思われます。

- コンフェデレーション内でサブ AS がホップ数の評価対象とならないため、他の属性で制御する必要がある。
- サブ AS ごとにルーティングポリシーが必要となる。
- サブ AS ごとにデフォルトルートが必要となる。

実際の移行は、次の順序で実施しました。

1. サブ AS の境界ルータとなるルータに BGP を設定する。
2. OSPF プロセスを分割する。
3. 同時に、OSPF による経路を再分配する。

このようにコンフェデレーションを利用してみて、iBGP のフルメッシュ解消方法としては、ルータリフレクタよりも容易なものと思えました。そして、このようなコンフェデレーションの利用によって、IGP のスケーラビリティが確保できました。