RIP と OSPF による経路制御

加藤 朗（東京大学）

1998 年 12 月 15 日

Internet Week 98　**国立京都国際会館**

（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

このチュートリアルでは、インターネットにおけるルーティング制御のうち、IGP として使用されている RIP と OSPF について取り上げます。昨年は、同じ内容を 3 時間コースで説明しましたが、とても時間が足りないことから、今年は 1 日コースとしました。

このチュートリアルで取り上げるのは、次の内容です。

- 経路制御概論
  プログラミング実習などとは異なり、大学の教育体制もあまり整っていませんので、ルーティングの概論を説明します。
- RIP と RIP2
  比較的単純なルーティングプロトコルとして、RIP と RIP2 を取り上げて解説します。
- OSPF
  OSPF の解説を行い、それをどうやって動かすかを解説します。

2 経路制御概論

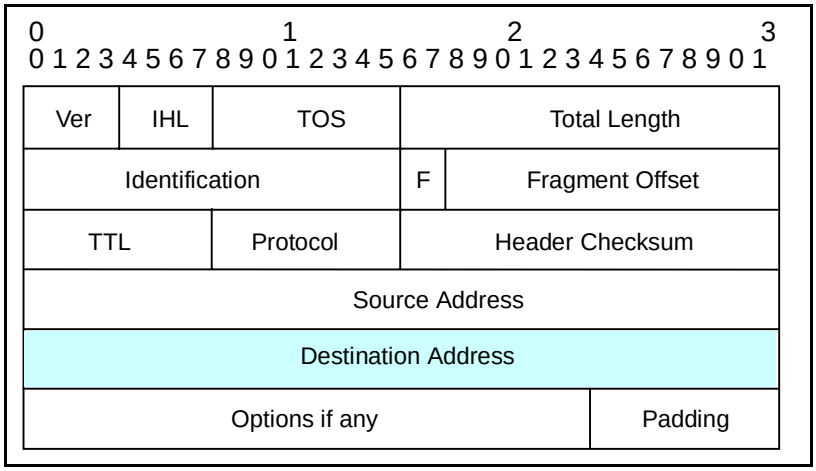
2.1 ルーティングとは

最初に、「ルーティングとは何か」をまとめておきましょう。

インターネットのルーティングは、それぞれのルータにおいて、正しい方向にパケットを送るという作業を繰り返すことによって、高い確率で相手にパケットを届けることです。ルータでは、あらかじめ経路テーブルを管理しておき、受け取ったパケットの「宛先アドレス」を経路テーブルに照らし合わせて、適当なインタフェースに出力するという作業を行います。その経路テーブルをどのように管理するかということが、後述するルーティングプロトコルの役割です。

電話網や X.25 のように、あらかじめコネクションを張っておくネットワークでは、最初のセッションを張る時にのみ宛先アドレスが必要となります。IP はこれとは異なり、それぞれのパケットにアドレスを持っています。IP パケット毎に独立して経路の選択が行われており、同じ宛先のパケットでも別の経路を通る可能性があります。また、それぞれのルータは独立した経路テーブルを持ち、独立して経路制御を行っています。それぞれのルータの経路テーブルを管理をいかに行うか、ということがルーティングプロトコルの本質です。

なお IP では、基本的には宛先アドレスだけを見て経路制御を行っています。パケットの中継にあたって、ルータはバージョン番号や長さのチェックを行い、TTL を操作してチェックサムを再計算するといった作業を行いますが、経路を決定するために必要なのは、宛先アドレスだけです。他のフィールドは参照しませんから、行きと帰りで異なる経路を通ることがあります。ユーザにとっては、相手からパケットが帰ってきて初めて役に立ちますが、IP のパケット毎のルーティングでは、行きと帰りは全く独立した事柄です。



Ethernet ブリッジでは、Ethernet フレームの先頭に宛先アドレスがありますから、先頭を見れば中継作業を開始することができます。IP ルータはパケットを全部受け取ってから中継作業を行いますが、現在では、宛先アドレスを受け取った途端に中継作業を開始するルータもあります。

2.2 古典的な IP の経路制御

古典的な経路には、次の 3 種類があります。

- ホスト経路
  32 ビットの IP アドレス全体を指定して、特定のホストを宛先とする経路です。
- ネットワーク経路
  アドレスからネットワーク部を抽出して、そのネットワークを宛先とする経路です。
- デフォルト経路
  ホスト経路やネットワーク経路に合致するものがない時に用いられるアドレスです。

ルーティングを決定するには、32 ビットのアドレス全体を比較してホスト経路を検索し、一致しなければ、ネットワーク部を抽出してからネットワーク経路を検索し、それでも一致しなければデフォルト経路を使用するという方法（アルゴリズム）が採られていました。デフォルト経路がなければ、ICMP Network unreachable を返します。ネットワーク部の抽出には、アドレ

スの先頭数ビットを見て、どのクラスに属するアドレスかを決定していました。

1985 年 8 月に RFC950 によってサブネットが導入され、クラス B アドレスの有効活用が試みられました。共通する長さのサブネットマスクを決めておき、宛先アドレスのネットワーク部が、自らが属するネットワーク部と一致した場合には、サブネット経路を検索して送り先が決められるものです。サブネットの大きさは、1 つのネットワークにおいて一定でなければなりませんでした。

2.3 現在の経路制御～ CIDR

1993 年 9 月に RFC1517 によって Classless InterDomain Routing（CIDR：サイダー）が導入され、可変長サブネットが使えるようになりました。これは、経路に必ずネットマスクやマスク長を付随させることにして、8 ビット境界のクラスを廃止して、任意の長さのネットワークアドレスを使えるようにするものです。同時に、ほとんど使われていなかった不連続なネットマスクも廃止され、ネットマスクは長さだけで表現されるようになりました。例を挙げましょう。

- 203.178.136.0/24　　クラス C1 個
- 203.178.136.0/25　　上の前半のみ
- 203.178.136.0/23　　クラス C2 個分

1988 年くらいまでは、ネットワークアドレスが 2 バイトで済むために、クラス B アドレスの使用が推奨されていました。日本では 133.1 ～ 133.254 などが、その時期に割り当てられたものです。その後、インターネットの発展に伴って、クラス B のアドレス空間が枯渇する恐れが出てきました。そのため、複数のクラス C アドレスを割り当てることが行われましたが、今度は、経路数が急増してしまいました。

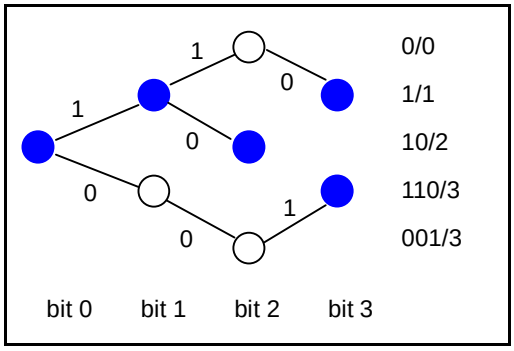
連続する経路を1つの経路にまとめて、たとえば131.112.0.0/16と131.113.0.0/16 を 131.112.0.0/15 にまとめることを考えます。つまり、経路数を削減するために、任意のビット長のネットワークアドレスを扱う必要が出てきたのです。これに伴って、経路テーブルにネットマスクやマスク長を格納できるように、ルーティングプロトコルが改良され、OSPF は OSPF2 に、RIP は RIP2 に、BGP3 は BGP4 になりました。これにより、任意の大きさのサブネットを作成することができる環境が整い、アドレスの枯渇問題にひとつの目途がつきました。

なお、ネットワークアドレスをまとめるためには、連続するネットワークが同じ方向になければなりません。経路の集約の効率を上げるためには、ISP に対してネットワークアドレスのブロックを割り振り（allocate）、ISP の顧客にはそのブロックの一部を割り当てる（assgin）という方法が採られるよ

うになっています。このため、ISP を変更するとネットワークアドレスを変更しなければいけないという不都合が発生しますが、このようにしなければ経路テーブルが溢れてしまい、たちまち、インターネットは動かなくなってしまうでしょう。

CIDR による経路検索は、次のように行われます。全ての経路にはマスク長が付随しますから、宛先アドレスと経路のネットワーク部を取り出したものを比較して、一致した場合にそれを候補とします。複数の経路情報と一致した場合には、最もマスク長が長いものを採用します。最も長いものを探し出すためには、経路情報を単純なテーブルとして記憶すると、何度も探索しなければならないため、実際には二分木によるテーブルの探索を行います。

具体的には、下図のように二分木を作成し、経路が存在した場合（図の●）に、それを経路として採用します。最大長のものを探し出すために、さまざまな工夫がなされますが、最大で 32 回の比較を行えば探索自体は終了しますから、速度的な問題はありません。実際には、二分木を縮退したものが使用されており、高速化のため、かなり複雑なアルゴリズムが使用されることもあります。



1994 年 3 月の IETF によって CIDR の導入が決定され、かなりの経路が集約されました。経路数の増加は、ルータのメモリや CPU に負荷をかけるだけでなく、安定しない経路が発生するなどの問題を引き起こします。そのため、CIDR によって集約されたアドレスを促進するために、ISP に割り当てられたアドレスブロックに移行する場合はホスト数によるアドレス割り当ての制限を緩和したり（APNIC）、19 ビットよりも長いネットワーク部を持つ経路情報を無視する（Sprint Net）などの方策が採られました。その後も、経路数はほぼ線形に増加し続けており、インターネット全体では約 60,000 の経路が存在しています。

CIDR の効果はインターネットにおける経路数の減少だけではありません。企業ネットワークにおいては、可変長ネットマスク（Variable Length Subnet Masks：VLSM）によって、IP アドレス空間を有効に活用することができます。サブネットの大きさを自由に決められるので、セグメント毎に相応しい大きさのネットマスクを利用することができるのです。VLSM を実現するためには、CIDR 対応の経路制御方式を持ったネットワーク機器と、CIDR

に対応した経路制御プロトコル、さらにその概念を理解した管理者が必要となります。

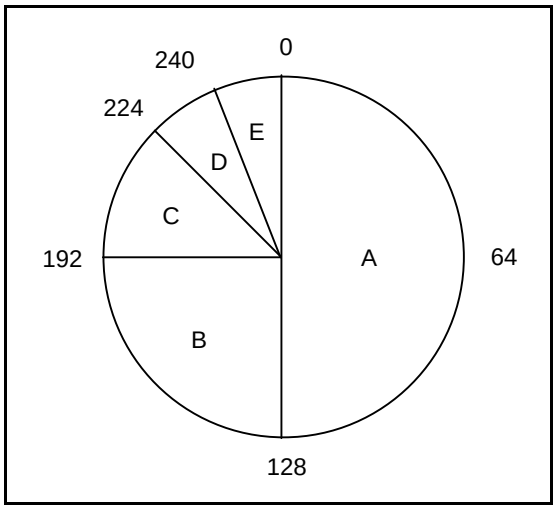
経路制御は、古典的なクラス別ネットワーク単位で経路を管理していた時代から、それに加えてサブネット単位での経路を管理していた RFC950 の時代、そして、任意のビット数を持ったネットワーク部によって経路を管理する CIDR の時代へと変わってきたと言うことができるでしょう。現在では、ISP に割り当てられた CIDR ブロック単位で経路情報を集約してからインターネットに広告することが普通になっています。通常の企業ネットワークでは経路の集約を気にする必要はほとんどありませんが、1,000 を超えるサブネットを持っているような巨大なネットワークでは、経路の集約を考慮すべきかもしれません。特に、経路制御プロトコルとして RIP2 を使用している場合には、集約することは必須でしょう。

2.4 アドレス枯渇

経路制御の話題から多少離れてしまいますが、IP アドレスの枯渇について触れておきましょう。

アドレスが足りなくなってくると、アドレスの管理を行っているレジストリは、必要最小限の割り当てを行おうとしますので、より多くのアドレスを貰おうとする攻防戦が発生します。現在は、社内にはプライベートアドレスを使用し、NAT によってインターネットに接続する形態が増えてきましたので、大きな企業でも多くのグローバル IP アドレスを必要とすることは少なくなってきました。反面、ソースアドレスがプライベートアドレスとなっているパケットが時々観測されるようになり、問題になっています。

IP アドレス空間を図示すると、次のようになります。クラス A が半分を占めていることが分かります。



そのため、クラス A のアドレスを返還していただき、それを分割して割り当てる作業が試験的に行われていますが、究極の解決策は IPv6 の普及を待つことでしょう。

2.5 経路制御の実際

ここで、経路制御の実際を見てみましょう。IP のパケットでは、宛先アドレスだけを見て経路が決定されますから、「経路テーブルをいかに管理するか」ということが最も本質的です。昔は、管理者が設定した静的で安定な経路テーブルが使われていました。ただし、静的な方法では、トポロジの変化に追随できませんし、管理者のミスもカバーできません。

そこで、経路制御プロトコルを導入して動的な経路制御が実施されるようになりました。経路制御プロトコルを導入すると、経路情報を交換するためのバンド幅や CPU 能力、管理者の教育に要するコストなど、さまざまなオーバーヘッドが生じます。しかし、ネットワークトポロジの変化に自動的に追随します。これによって、インターネットの動的な発展が可能になったと言えるでしょう。

2.6 経路制御プロトコル

動的な経路制御は、経路情報が変化した場合に、それぞれのルータの経路テーブルを自動的に書き換えることが本質です。そのためにルータ同士で経路情報を交換しますが、そのために使用されるのが経路制御プロトコルです。

経路制御プロトコルの代表的な方法のひとつが「Distance Vector（Bellman-Ford）型」と呼ばれるものです。それぞれのルータが知っている経路テーブルを交換しあうことを繰り返して、全体として正しい経路テーブルを作り出します。経路テーブルには、宛先までの距離を表すメトリック（metric）という値が格納されていて、より小さい値を持つ経路が選択されます。Distance Vector 型の経路制御プロトコルでは、経路情報を選択的に広告したり受理したりすることによって、ポリシーに基づく経路制御を実現することができます。後述する RIP は、Distance Vector 型の経路制御プロトコルです。

もう一つの代表的な方法は、「Link State 型」と呼ばれるもので、ネットワークトポロジのデータベースを作って、それを全てのルータで共用します。Dijkstra のアルゴリズムを使用して、自ルータを根とする Spanning Tree を作成し、それによって経路テーブルを作成します。データベースのコピーを同期する方法と、フィルタ（ポリシー）を使えないのが問題となりますが、Distance Vector型の経路制御に比べて収束やループの解消が速いと言われています。後述する OSPF や IS-IS などは、Link State 型の経路制御プロトコルです。

3 RIP と RIP2

3.1 RIP の概要

RIP は、長くドキュメントなしで使用されていましたが、RC1058 で定義されました。RFC2400 では「Historic」プロトコルである旨が示されていますが、実際には、まだかなり多く使われています。BSD Unix の routed によって実装されたことから、広く普及しているものです。RIP は UDP ポート 520 番と、ローカルネットワークにおける IP ブロードキャストを使用します。現在のプロトコルでは、必要のないホスト（たとえば IP を使わないホスト）にも負荷をかけることから、ブロードキャストはほとんど使用されません。

RIP は Distance Vector 型の経路制御プロトコルです。1 から 16 までのメトリックで経路の質を表し、小さな数ほど、「近い」ことを表します。メトリック 16 は無限大、すなわち届かないことを示しますから、大規模なネットワークでは使えません。交換する経路情報のうち、宛先 0.0.0.0 はデフォルト経路を表し、全ての宛先にマッチすることになっています。

RIP では、ルータが自分が持つ経路テーブル全体を、ブロードキャストによって 30 秒に 1 回アナウンスします。ブロードキャストを使用していますが、そのアナウンスが相手（隣接ルータ）に届かない場合も考えられます。そのため、180 秒間待った上で、更新されない経路をダウンしていると判断し、メトリックを 16 に変更します。これを「ホールドダウン」と言います。さらに、120 秒の間に経路がアナウンスされてこない場合には、その経路が消失したものとして経路テーブルから消去します。

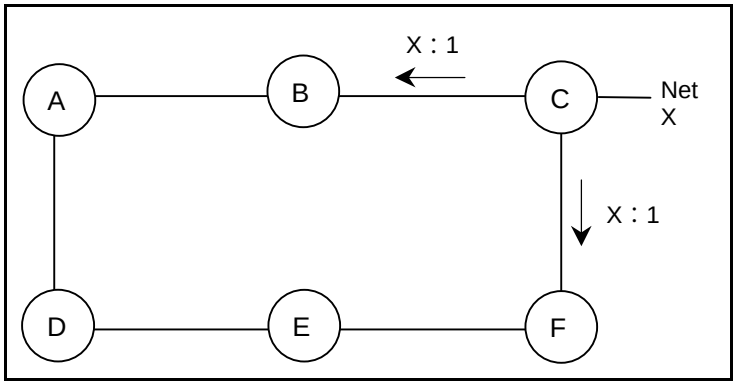
RIP のメトリックは 1 から 15 までしか使えませんから、リンクの帯域を表現するには範囲が足りません。そのため、リンクの帯域にかかわらず、ホップ毎にメトリック 1 を加えるのが普通です。複数の隣接ルータから同じ値のメトリックが通知された場合、RIP ではどちらか一方を選択します。一方からのアナウンスが途絶えがちの場合には、他方を選択する、あるいは両方を使って負荷分散を図るという賢い選択をする実装もあるようですが、ほとんどの実装では、既に経路テーブルに存在している経路を優先します。

経路テーブルに先に存在した経路を使うというこの実装では、ネットワーク機器が起動した順序によって、安定後の経路に幾つかのパターンが存在することになります。これは、経路デバッグを行う時に非常に面倒になり、好ましいものとは言えません。より現代的な経路情報プロトコルである OSPF や BGP4 では、到着順によらず、1 つのパターンに収束するように工夫されています。

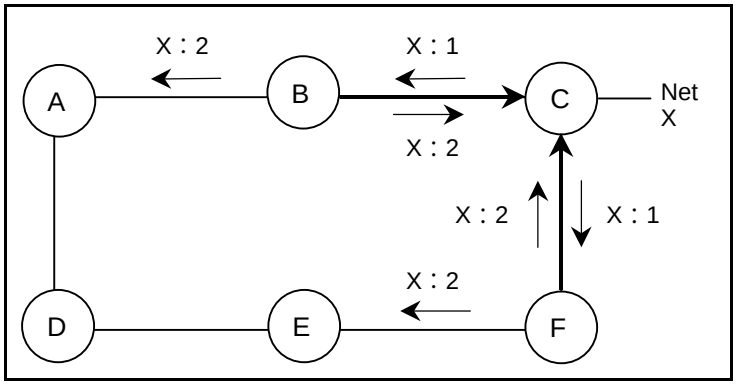
3.2 経路の伝搬

RIP では、30 秒ごとに経路テーブルを隣接するルータに伝搬します。したがって、RIP によって経路が伝搬する方向は、トラフィックと逆方向になります。ルータ C から、ネットワーク X に至る経路が伝搬していく模様を示します。

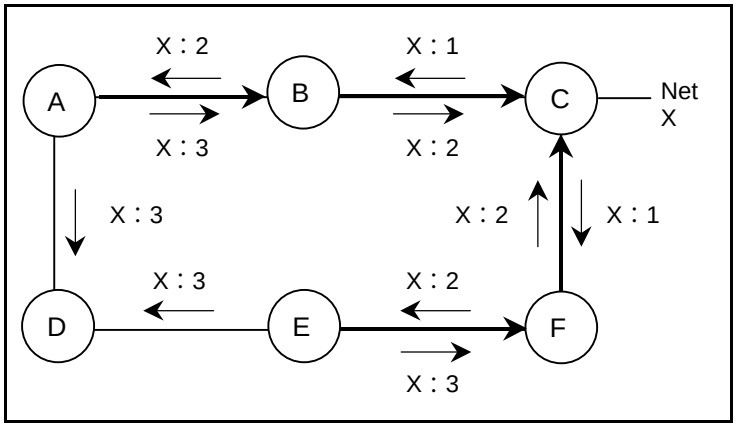
1. ルータ C から X に関する情報がルータ B と F に伝えられます。



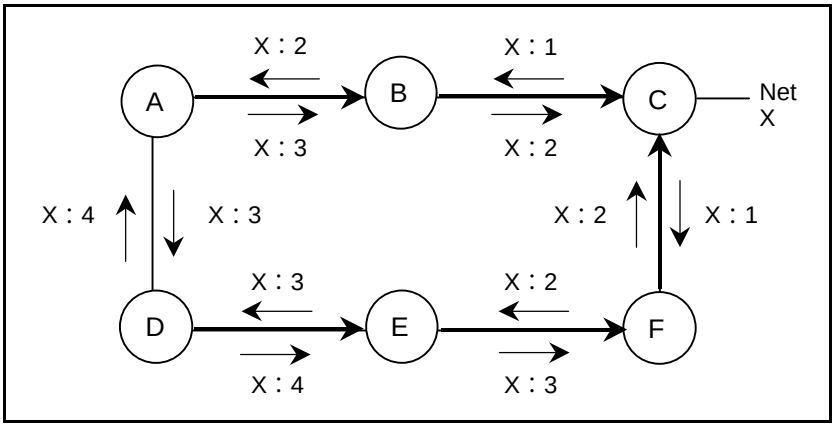
2. 前項に加えて、ルータ B と F から、メトリック値を増した X に関する情報が、ルータ A/C/E に伝えられます。ルータ C では、伝えられた情報のメトリック値を比較し、受け取ったものを捨てます。



3. 前項に加えて、ルータ A と E から、メトリック値を増した X に関する情報が、ルータ B/D/F に伝えられます。ルータ B/F ではその情報を捨てます。



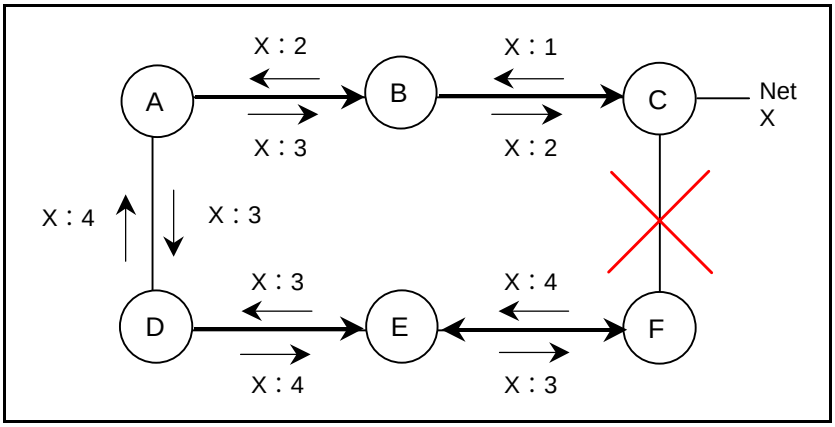
4. 前項に加えて、ルータ D から、メトリック値を増した X に関する情報が、ルータ A/E に伝えられます。ルータ A/E はその情報を捨てます。以上で、X に至る経路が全体に行き渡り、安定したことになります。



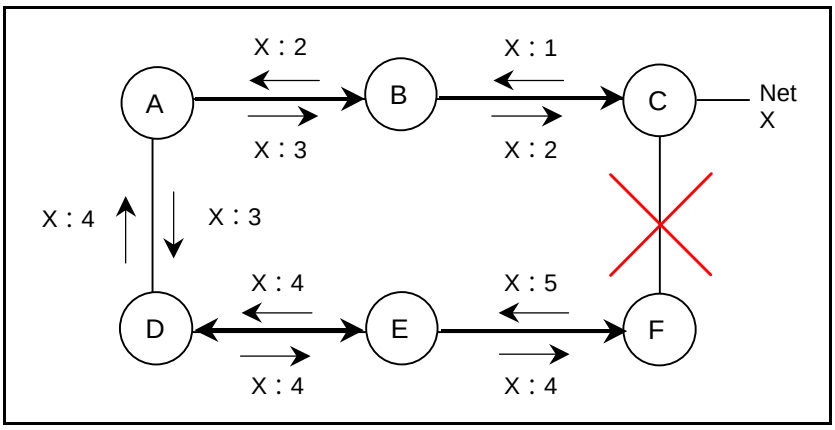
RIP は 30 秒ごとに経路テーブルを伝達しますから、D が X への経路を知るまでは、最大で 90 秒、平均 45 秒もの時間がかかります。ネットワークが大きい場合には、この時間が大変な問題となります。

次に、同じ例で、C と F の間のリンクがダウンした場合を考えてみましょう。

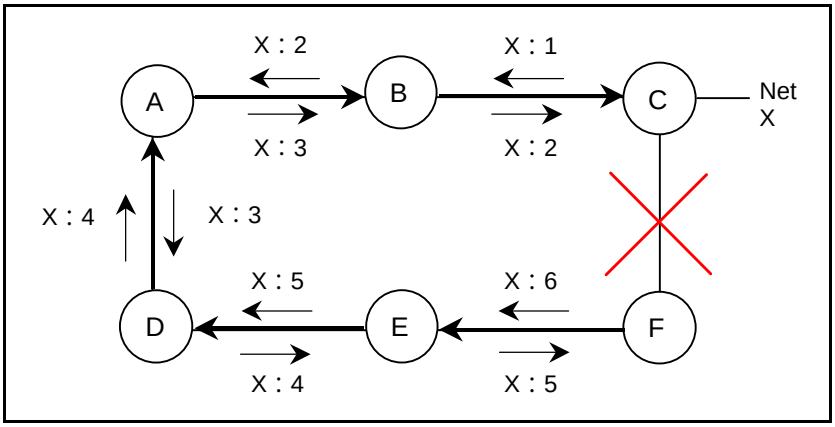
ルータ F は、ルータ C からのアナウンスが途絶えてから 180 秒後に、X に至る経路のメトリックを 16 にして、到達できないことを示します。さらに、ルータ E からアナウンスされる X への経路（メトリック 3）を採用して、X への経路をルータ E に向けます。F における X への経路のメトリックは 4 になります。



これは、E と F の間で、X 行きの経路がループしている状態です。パケットは TTL 値が尽きるまでピンポンしてから破棄されます。



ルータ E は、D から X への経路（メトリック 4）と、F から X への経路（メトリック 4）を比較しどちらかを採用します。F から伝えられる X に至る経路のメトリック値が 5 に増加します。その時点で、E は X 宛の経路として D を選択します。D が X への経路として E を選択している場合には、D と E の間で X 行きの経路がループしている状態となります。

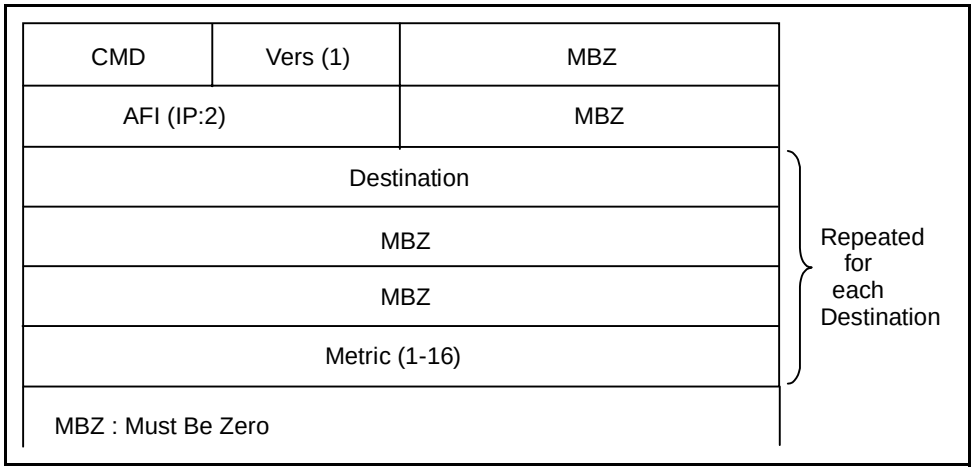


前項と同様のやりとりによって、E/F から X に至る経路のメトリック値が増加し、D は X に至る経路として A を選択します。

このように、リンクのダウンはやがてネットワーク全体に伝達されますが、経路ループのない状態に至るには、かなり長い時間がかかってしまいます。

3.3 RIP パケット

RIP パケットの構造を示します。経路情報毎に、9 バイト目以降のフィールド構造が繰り返されます。0 の部分が多いのは、XNS の構造をそのまま使用したためです。



メトリックフィールドは、32 ビットの大きさを持っていますが、1 から 16 の値しかとれないことに注意してください。また、RIP は UDP パケットを使いますから、フラグメンテーションを避けるために、全体で 512 バイトが上限となり、1 つのパケットで 25 経路分しか送ることができません。

CMD フィールドには、RIP コマンドが納められます。値の意味は次のとおりです。

- 1 ―リクエスト
  隣接するルータに経路情報を要求します。Address Field Identifier（AFI）を 2 に、Destination にアドレスをセットし、Metric を 0 にすると、そのアドレスに対する経路情報が返されます。同様に、AFI を 0、Metric を 16 とすると、経路テーブル全体が返されます。ほとんど使われませんが、ルータが起動したときに、経路テーブル全体を要求する実装があります。
- 2 ―レスポンス
  リクエストに対する応答です。30 秒毎に送られる更新メッセージも、このフォーマットとなります。

3.4 RIP の改良

歴史的なプロトコルであるだけに、RIP に対するさまざまな改良案が提案され、実装されています。以下にそれらを幾つか挙げます。

- メトリック 16
  メトリック 16 の経路を受け取った場合には、経路がなくなったという重要な情報ですから、180 秒間のホールドダウンタイムを省略して、すぐに 120 秒間の GC タイマを起動します。
- Triggered Update
  経路変更が通知されてきた場合、30 秒を待たずにアナウンスを行い、変化を素早く伝搬しようとします。
- Split Horizon
  経路が送られてきた方向（インタフェース）には経路情報を送らないことによって、オーバーヘッドを低減し、また、誤った経路選択を防ごうとします。
- Poisoned Reverse
  Split Horizon と併用し、経路が向いている方向にはメトリック 16 の経路を送出することで、特定の経路へのアナウンスを抑止します。オーバーヘッドは増加しますが、ループの解消が速い場合があります。
- 複数の経路候補を保持
  定常状態では最小メトリックのものを使いますが、ホールドダウン状態が発生した時に直ちに次善の経路を設定することによって、経路が切り替わるまでの待ち時間を節約することができます。ほとんどの routed や Gated に実装されています。

3.5 RIP とサブネット

RIP はネットマスクを伝搬しませんが、サブネットの時代には、それぞれのルータがサブネットの大きさを知っていたので、次の方法でサブネット経路とホスト経路を区別することができました。

- 宛先がルータと同一のネットワーク
  宛先アドレスをサブネット部とホスト部に分け、ホスト部が 0 でなければホスト経路、ホスト部が 0 であればサブネット経路。
- 宛先がルータとは別のネットワーク
  宛先アドレスのホスト部が 0 でなければホスト経路、ホスト部が 0 であればネットワーク経路。

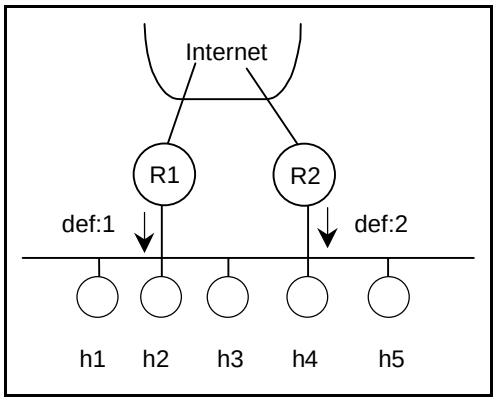
3.6 RIP の問題点

これまでに述べてきた RIP の問題点をまとめておきましょう。

- 1 つのパケットで 25 経路しか搬送することができないので、経路数が増加するとルータの負荷が増大し、さらにルータの入出力キューがオーバーフローする可能性があります。
- メトリックが 1 から 16 までなので、大きなネットワークには適応できません。
- 30 秒に 1 回の情報伝搬が基本なので、経路の収束が遅く、大きなネットワークには適応できません。
- ネットマスクを搬送できないので、CIDR に対応できません。

これらのことから、RIP は直径が数ホップの小さなネットワークで利用するか、デフォルト経路だけを通知するために使用するのが相応しいでしょう。ルータからデフォルト経路だけを通知し、ホストはそれを受信するだけという設定を行い、ルータの存在の通知を RIP で行うこともできます。IPv6 では、同様の機能が Router Advatisement として組み込まれています。

複数のルータからメトリック値の異なるデフォルト経路をアナウンスすると、メインのルータが落ちた場合に、自動的にバックアップルータに迂回することも実現できます。



ただし、このようなバックアップ用途には、現在では、Hot Standby Router Protocol（HSRP）なども用いられています。

3.7 RIP2

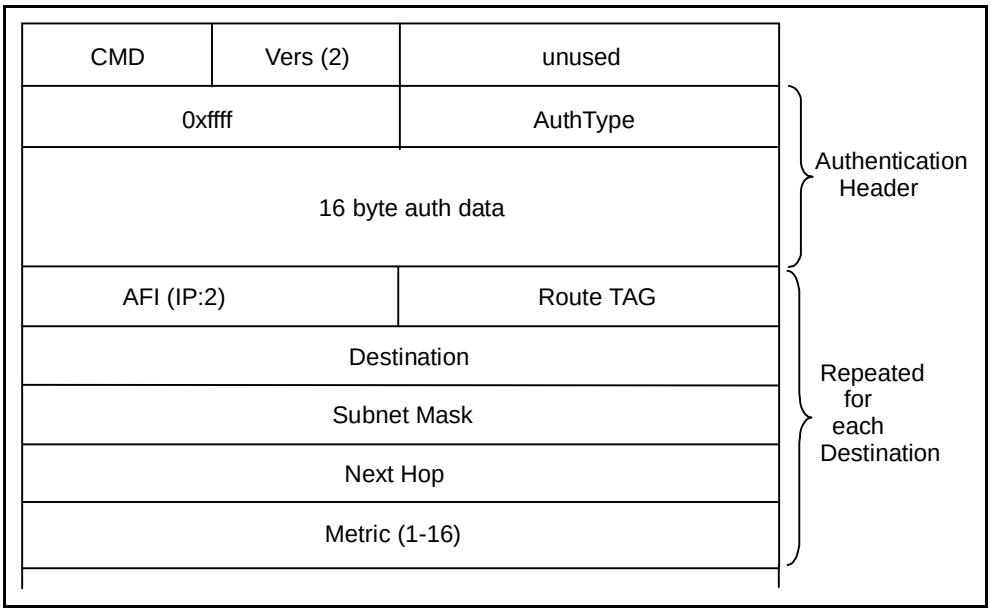
RIP2 は、RIP の改良版として RFC2453 によって定義されています。CIDR に対応したことが最大の改良点です。また、ブロードキャストではなく 224.0.0.3 のマルチキャストを使用することも大きな改良でしょう。メトリックの制限や、アルゴリズムは変更されていません。

また RIP2 では、1 経路分のフィールドを使って認証を行う機能が追加されています。AFI が 0xffff の場合に、16 バイトの暗号化されていないパスワードを格納することで、遠方からの誤った経路情報を排除するしくみが組み込まれています。

RIP2 では、RIP で使われていなかったフィールドを用いて、次のような情報を伝達するようになっています。

- ネットマスク（4 バイト）
  これにより CIDR に対応します。
- ネクストホップアドレス（4 バイト）
  自ルータとは異なるルータへの送信を指示できます。
- タグ（2 バイト）
  BGP と連携する時などに用いられます。

RIP2 のパケット構造を示します。認証を行わない場合には、オーセンティケーションヘッダの部分が省かれます。



RIP2 は、gated や、Cisco/Bay Networks のルータなどに実装され、使用するための土台が揃っています。しかし、現実には、OSPF が普及して安定して稼働していること、ホストで使用するには RIP1 や DHCP によるルータアドレスの通知で十分なことなどから、あまり使われていないようです。

4 OSPF

Open Shortest Path First（OSPF）は、RFC2328（STD0054）で定義されている Link State 型のルーティングプロトコルです。複雑なプロトコルであり、8 年間に及ぶ数回の改訂を経た仕様書は、210 ページを超えます。

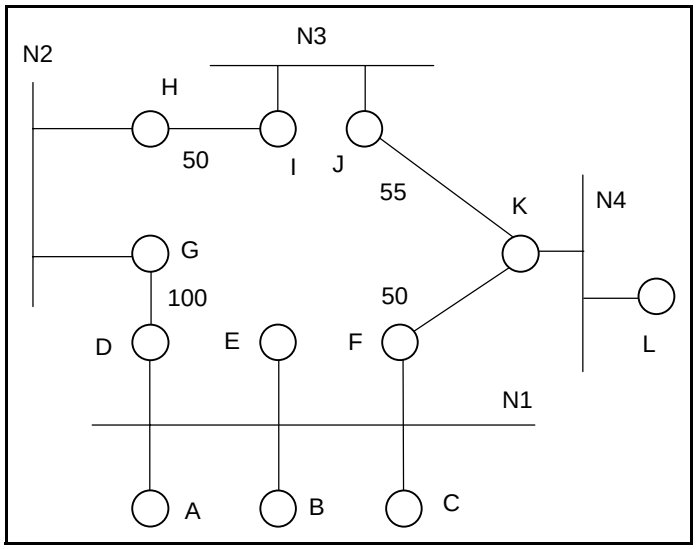
OSPF は Link State 型のルーティングプロトコルで、それぞれのルータとネットワークを Link State Advertisement（LSA）から成るデータベースに登録し、それらをルータ間で同期することで経路制御を行います。エリアによる階層構造を持つこと、外部から与えられる経路情報（通常は BGP）を扱えることなども OSPF の特徴です。OSPF は、IP プロトコル番号 89 を使用し、独自の再送プロトコルを持っています。同一ケーブル内ではマルチキャストを使用して情報を伝達しますが、再送はユニキャストで行われます。また、単純なパスワードと、秘密鍵と MD5 を使用したセキュリティ機能も持っています。

4.1 Link State 型経路制御

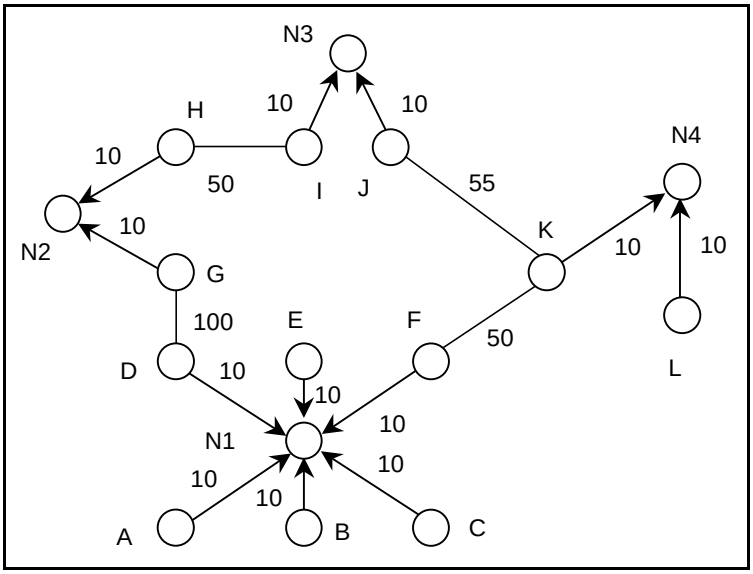
Link State 型の経路制御では、たくさんの LSA からトポロジデータベースを作成し、それぞれのルータで同一のトポロジデータベースを共有し、そこから経路を計算することになります。同一のデータベースをルータ間で保持するためには、信頼性のある通信と、LSA の変化をできるだけ短時間で伝達するしくみが必要となります。OSPF は IP 層を直接使うので、これらを全て含むものとなります。また、データベースから経路を計算するには、ループが発生しないように、全てのルータで共通のアルゴリズムを使用しなければなりません。共通のデータベースを参照するため、ループは短時間で解消しますが、ポリシーの実現は困難です。

4.2 経路の計算方法

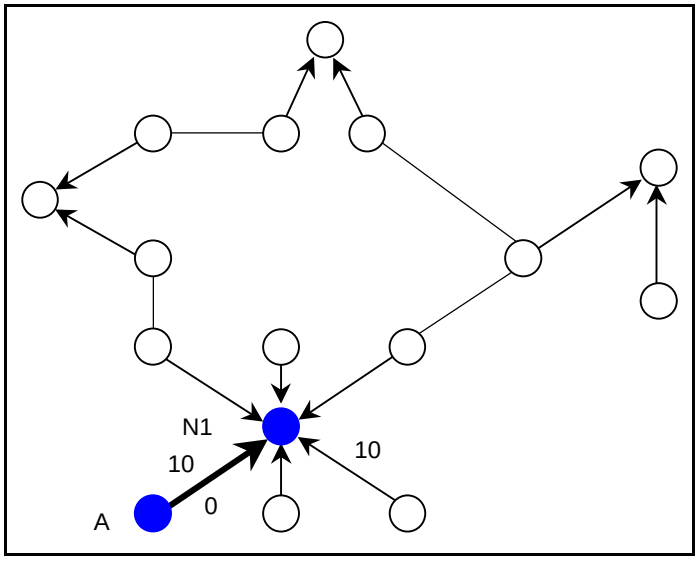
OSPF で経路を計算するには、Dijkstra のアルゴリズムを使用します。例を挙げて、計算方法を示します。N1 ～ 4 は Ehternet などのマルチアクセスネットワーク、A ～ L はホストまたはルータ（ノード）、ルータ間を結ぶ線（回線）に付けられた数字はその回線のコストを示すメトリックです。図中には示していませんが、マルチアクセスネットワークのメトリックは 10 とします。



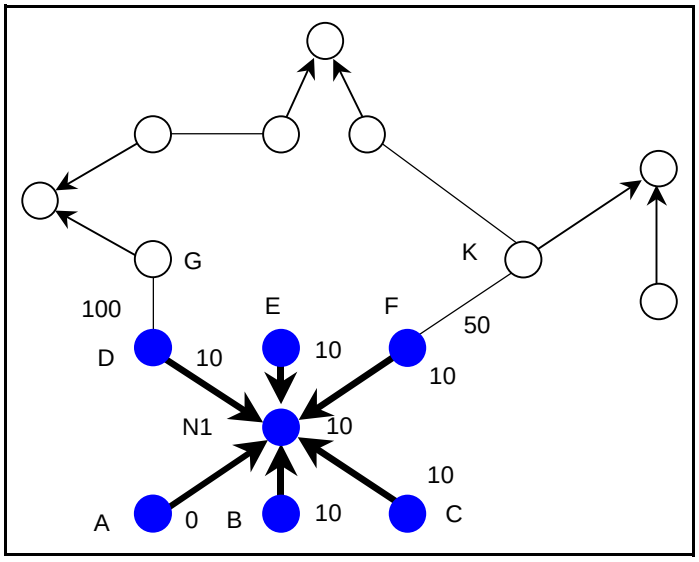
まず、マルチアクセスネットワークも 1 つのノードとして表し、その出力側にコストを与えながら、ネットワーク全体のトポロジをグラフ化します。ノード A における経路テーブルを得るために、A に対する SPF 木を計算します。



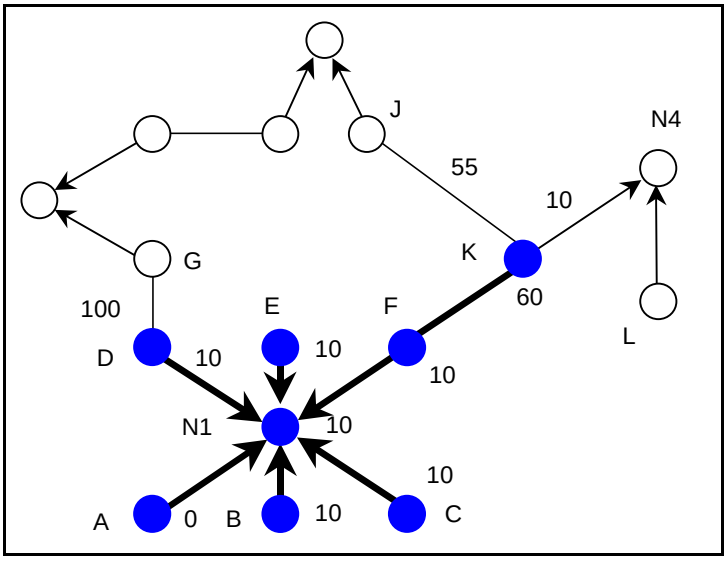
A からは、コストは 10 で N1 に至ることができます。



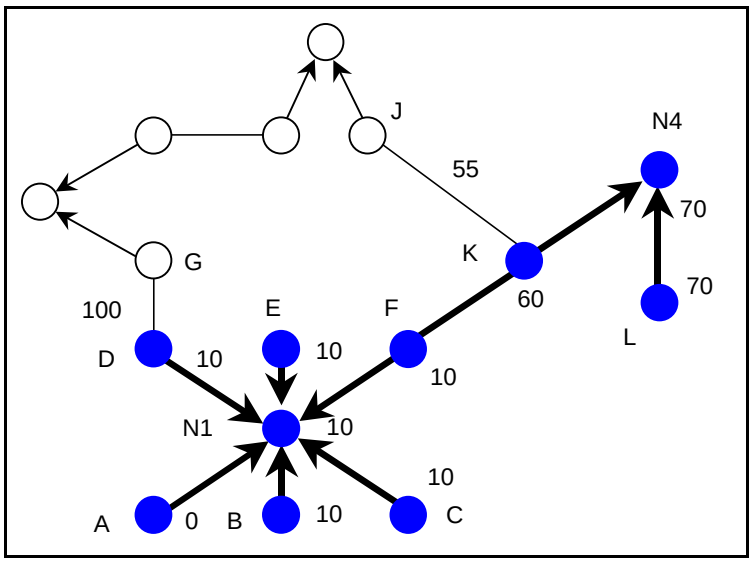
N1 に隣接するノードへのコストを計算します。マルチアクセスネットワークからの入力はコスト 0 ですから、コスト 10 で B/C/D/E/F に至ることができます。次に探索する候補は、G と K になり、それぞれのコストは 110 と 60 ですから、コストの小さい K を先に計算していきます。



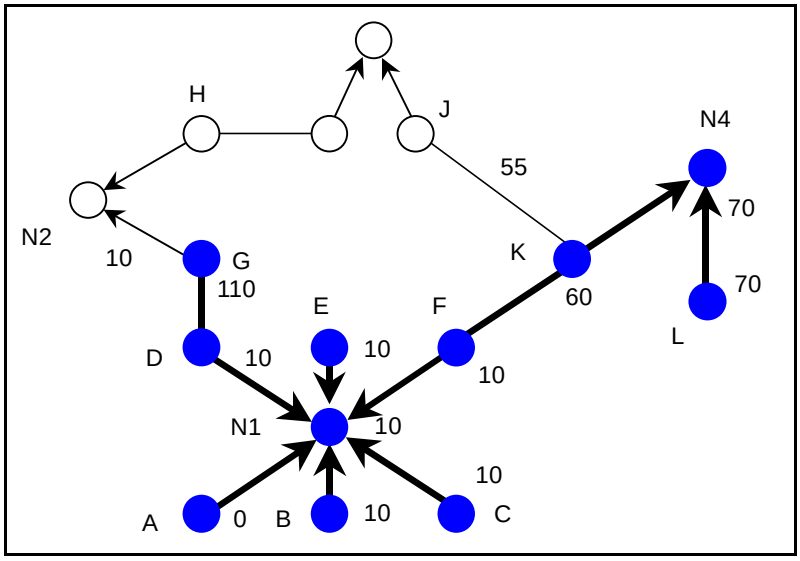
SPF 木に K を加えます。次の候補は、前のステップで後回しにされた G（コスト 110）と、J（コスト 115）、N4（コスト 70）です。最小コストの N4 を選択します。



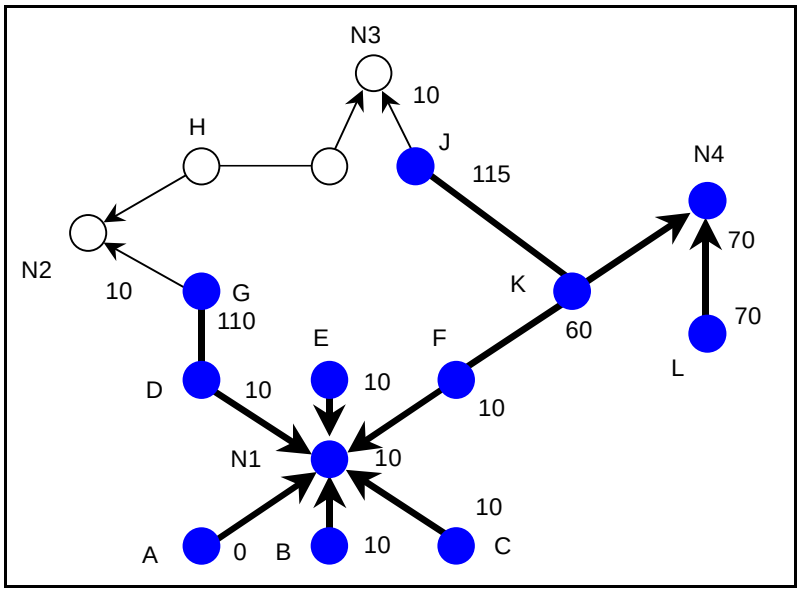
SPF 木に N4 と、そこからコスト 0 で到達できる L を追加します。N4/L からは先に枝がありませんから、前のステップで後回しにした G（コスト 110）と J（コスト 115）のうち、コストが小さい G を選択します。



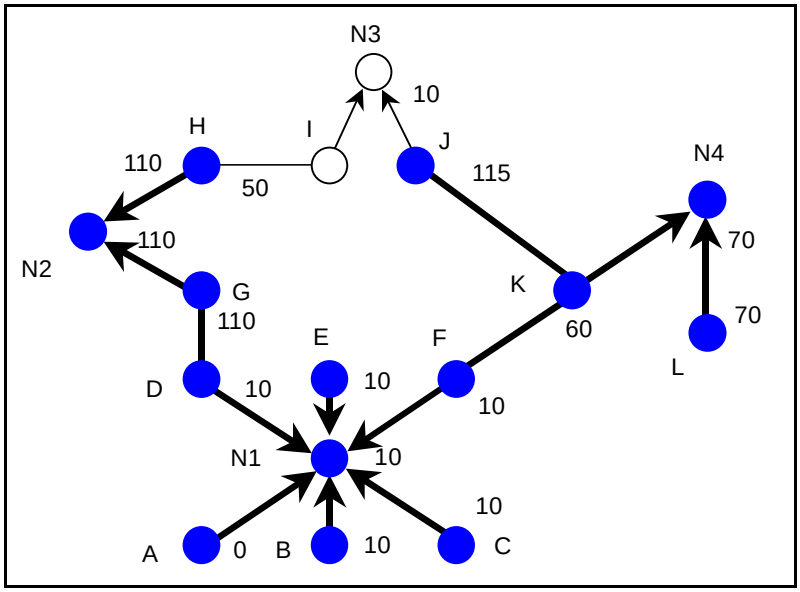
SPF 木に G を加えます。次の候補は、前のステップから残っている J（コスト 115）と、N2（コスト 120）です。J を選択します。



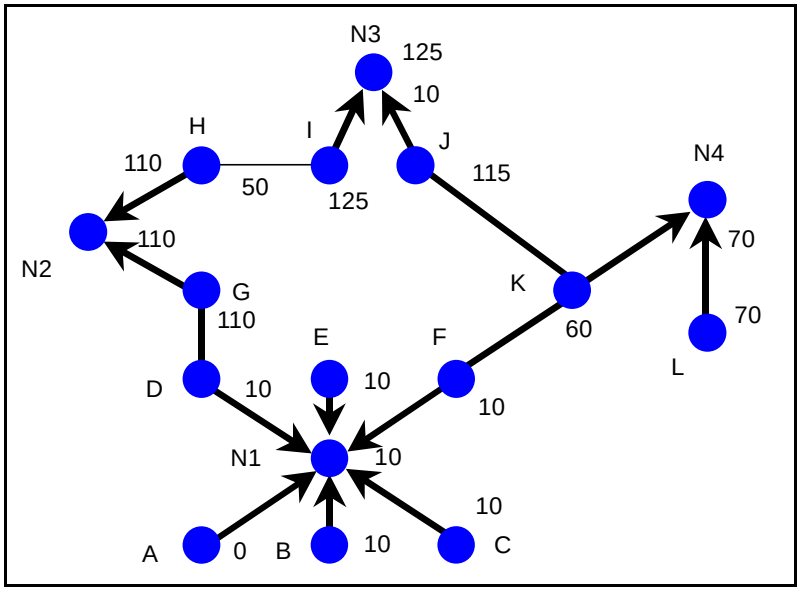
SPF 木に J を加えます。次の候補は、前のステップから残っている N2（コスト 120）と、N3（コスト 125）ですから、N2 を選択します。



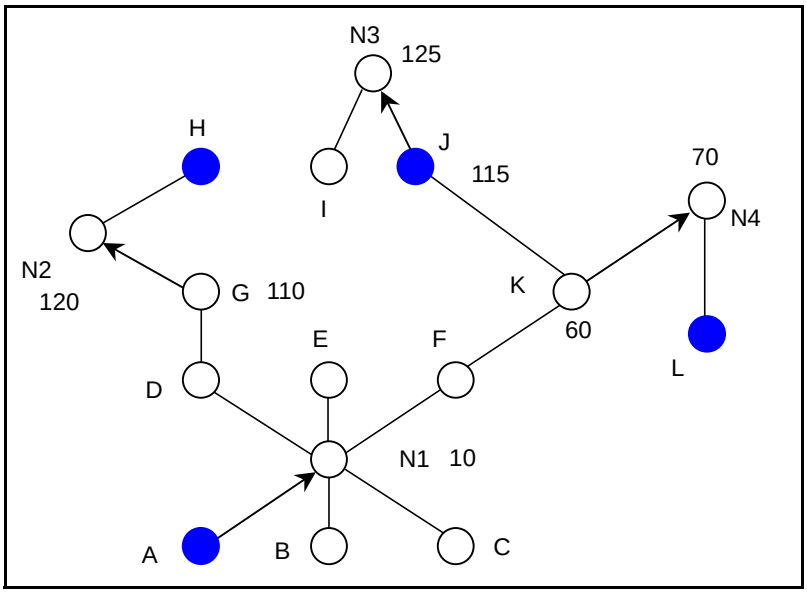
SPF 木に N2 と、そこからコスト 0 で到達できる H を追加します。次の候補は、前のステップから残っている N3（コスト 125）と、I（コスト 160）です。N3 を選択します。



SPF 木に N3 と、そこからコスト 0 の I を追加します。次の候補は、前のステップから残っている I（コスト 160）ですが、より小さいコスト 125 で到達できることが分かったので、候補から捨てます。



これで、SPF 木が完成しました。



このように、最小のコストで到達できるノードを SPF 木に順次取り込んでいき、それぞれのノードに至る最小のコストを計算していくことで、完全な SPF 木を作ることができました。

この SPF 木から、A における経路テーブルを作ると、次のようになります。

表 1：経路テーブルの例

宛先	ネット / ホスト	次のホップ	コスト
G	ホスト	D	110
K	ホスト	F	60
J	ホスト	F	115
N1	ネット	A	10
N2	ネット	D	120
N3	ネット	F	125
N4	ネット	F	70

OSPF では、まずネットワークの地図を作り（作り方は後述）、全ノードにその地図をコピーしておきます。各ノードでは、ここまでに述べてきた計算方法で、自分を根とする SPF 木を計算し、経路テーブルを作成します。全ノードが同じ地図から経路を計算しますから、地図を同期させることが最も重要な作業になります。

4.3 データベースの構成

OSPF では、5 種類の LSA の集合でネットワークトポロジを記述します。それぞれのルータは 32 ビットのルータ ID を持ち、ネットワークはそのネットワークに属する指定（代表）ルータ（Designated Router：DR）のアドレスを使って表現します。

LSA の種別は以下のとおりです。全ての Type-1 と Type-2 の LSA が揃うと、ネットワークのトポロジが分かることになります。

- Type-1：ルータを表します
  それぞれのルータが生成し、自らが接続されているネットワークをリストアップします。
- Type-2：ネットワークを表します
  DR が生成し、Type-1 とは逆に、ネットワークに接続されているルータをリストアップします。
- Type-3：サマリ情報を表します
- Type-4：AS 境界ルータを表します
- Type-5：AS 外部経路を表します

4.4 LSA の伝搬

全てのルータ間で LSA をやりとりするには、「ルータ台数の 2 乗回」ものやりとりが必要になります。これでは効率が悪いので、ネットワークに 1 台の代表（DR）と、そのバックアップ（BDR）を選出します。その上で、AllSPFRouters　という全てのルータが属するマルチキャストアドレス（224.0.0.5）と、AllDRouters という DR と BDR のみが属するマルチキャストアドレス（224.0.0.6）を定義します。DR/BDR　からの情報伝達には AllSPFRouters 宛に送り、各ルータから DR/BDR への応答には AllDRouters 宛に送信します。再送が必要な場合には、ユニキャストを使用します。全ての情報は、一旦 DR/BDR を経由して、ネットワーク上の全てのルータに伝搬します。

OSPF では、全てのルータが一意の Router ID を持っています。もちろん、Router ID はドメイン内で一意でなければいけませんから、通常はルータが持っている IP アドレスのいずれかを使用します。Router ID が変動しては困りますから、なるべく安定したインタフェースか、ダミーのソフトウェアインタフェースを作成して、それを Router ID として使用するのが良いでしょう。

4.5 データベースの同期

OSPF を喋るルータが起動してから、同期がとれて安定した状態になるまでの模様を、順を追って大まかに見て行きましょう。

- Down：未活性の状態
  まず、Hello パケットを AllSPFRouters 宛てに定期的にアナウンスします。Hello パケットには次のような情報が含まれており、通信の双方向や相互の状態を確認できるようになっています。
  - 自分の Router ID
  - Area ID
  - ネットマスク
  - Hello Inerval
    Hello を送る間隔（秒）。デフォルトは 10 秒です。ネットワーク内で共通でなければなりません。
  - Router Dead Interval
    Hello が途絶えてから、ダウンと判断するまでの秒数です。あるルータがダウンした場合、必要に応じて新しい DR/BDR が選出されます。ネットワーク内で共通でなければなりません。
  - Router Priority
    DR/BDR になり易さを表します。この値が同じ場合は、Rouer ID が大きなものが DR/BDR になります。0 では DR/BDR になりません。

- DR/BDR の IP アドレス
  既に DR/BDR が選出されている場合に、それらの IP アドレスを示します。DR/BDR が不明な場合には、0.0.0.0 にします。
- 既知のルータの Router ID リスト
  それまでに受信した Hello パケットの送り主のルータのリストを入れておきます。

- Init：初期状態
  受信した Hello パケットの中に、自らが含まれていたならば、その Hello パケットを送信したルータとの間で、双方向の通信が確認されたことになります。

- 2-Way：双方通信が確認された状態
  必要に応じて、DR と BDR を選出します。DR/BDR にならなかったルータは Full 状態に移行し、DR または BDR になったルータは ExStart に移行します。既に DR/BDR が選出されている場合には、Router Priority が大きいルータが後から加わっても、再選出は行いません。

- ExStart：DD の初期交換
  Master と Slave を決定し、シーケンス番号を交換します。Router ID が大きいものが Master になります。

- Exchange：Database Description（DD）の交換
  Master の主導によって、データベースのカタログ情報（DD)、すなわち LSA ヘッダを交換します。受信した LSA ヘッダを見て、後でリクエストする LSA を決定します。DD パケットには、シーケンス番号、最後に受信したシーケンス番号、複数の LSA ヘッダなどが含まれています。

- Loading：LSA の交換
  DD を見て、必要な LSA を要求する LS Request パケットと、要求された LSA を提供する LS Update パケットを使って、LSA を交換します。複数の LSA を 1 つのパケットで交換することができます。LSA の指定には、LS Type、LS ID、アナウンスしたルータの ID が使われます。

- Full：同期が取れた状態

シリアルリンクを通して OSPF のやりとりを行う場合は、リンクの両端でデータベースの同期を行う必要があります。双方が DR として動作しますから、Hello パケットに含まれる Priority は意味を持ちません。

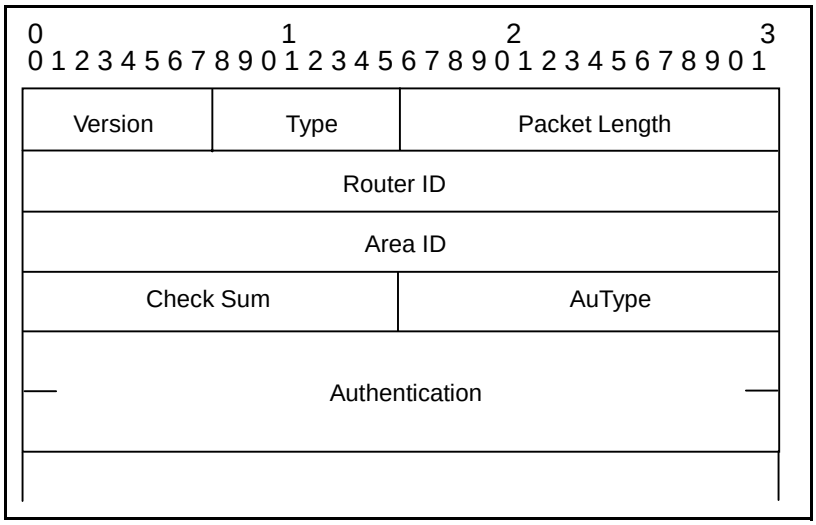
ATM や Frame Relay などの Non Broadcast Multi Access Network（NBMA：マルチキャスト通信が行えないネットワーク）では、あらかじめ各ルータに DR/BDR の候補を設定しておく必要があります。HelloInterval は 30、RouterDeadInterval は 120 と、やや控えめな値がデフォルト値となっています。なお、NBMA では、ダウンしているルータにも Hello が送られます。

4.6 OSPF パケット

ここから、主要な OSPF パケットの構造を見ていきましょう。

4.6.1 OSPF パケットヘッダ

全てのOSPFパケットの先頭に、24バイトのパケットヘッダが付けられます。

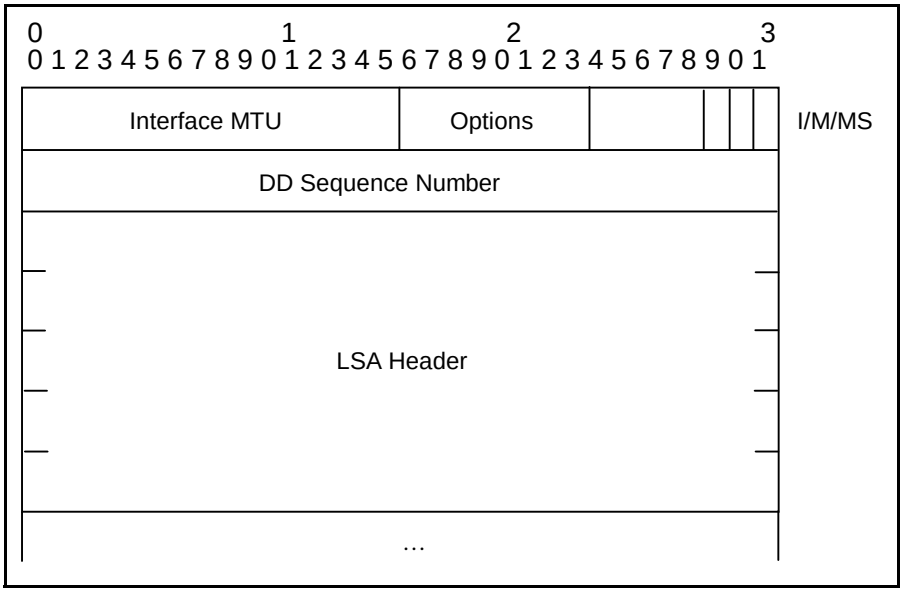


Type の意味は次のとおりです。

- 1：Hello
- 2：Database Description
- 3：LS Request
- 4：LS Update
- 5：LS Acknowledgment

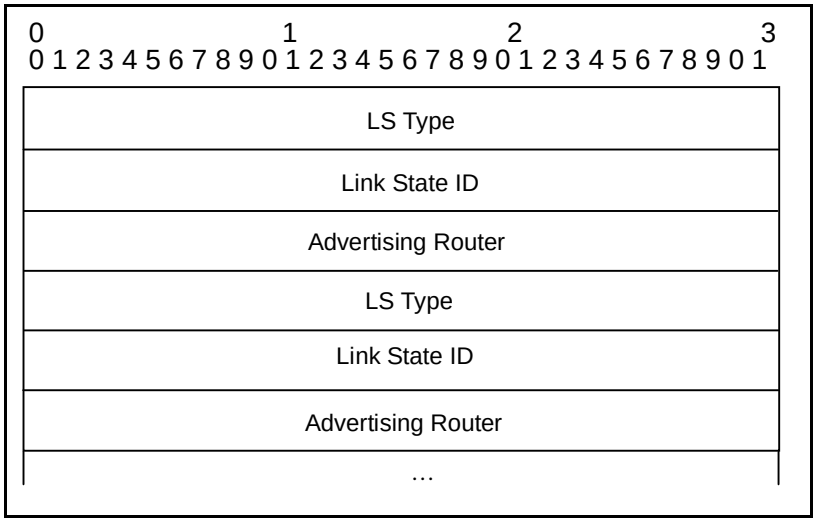
4.6.2 Database Description（DD）パケット

複数の LSA ヘッダを収容して、LSA のリストを通知します。



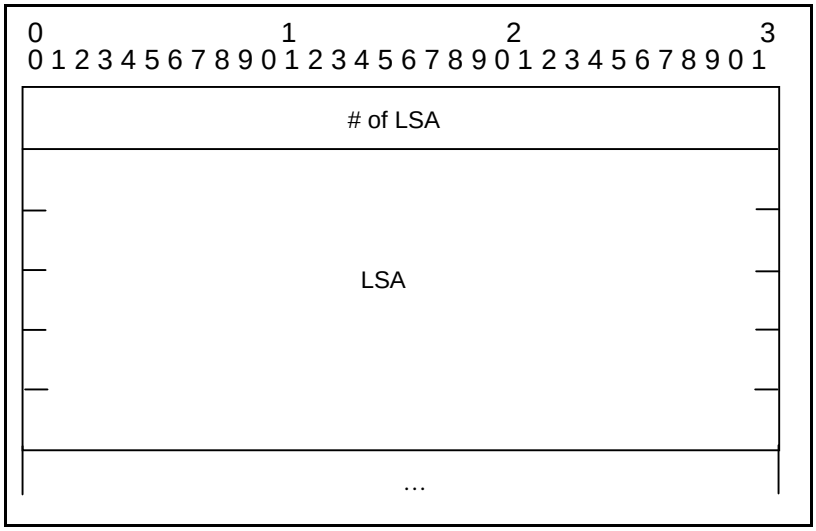
4.6.3 LS Request パケット

LSA の送信を要求します。複数のリクエストが収容されます。



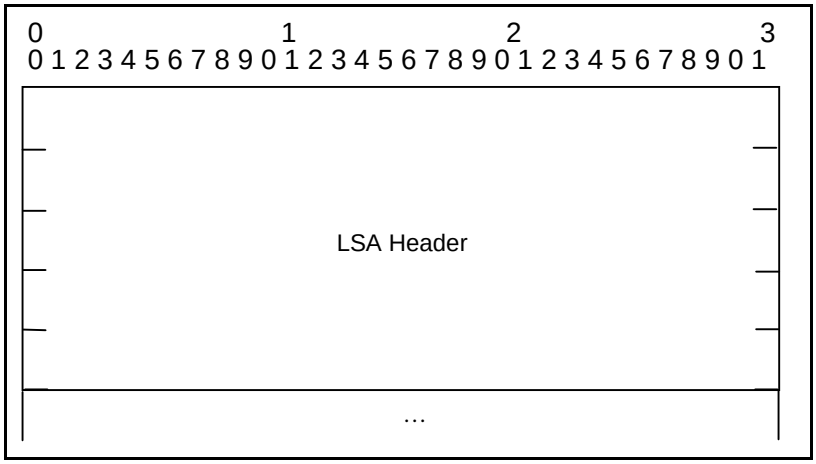
4.6.4 LS Update **パケット**

LSA が更新されたことを通知します。複数の LSA が収容されます。



4.6.5 LS Ack **パケット**

LSA の受信確認を行います。LS Update パケットを受信し損なった場合には送られません。



4.7 LSA

全ての LSA には、次の内容が含まれています。

- LSA 種別
- LSA Age
- Link state ID （LSID）
- シーケンス番号
  0x80000000 から 0x7fffffff。オーバーフロー時は特別な処理が規定されています。
- LS の長さ
- チェックサム
- Expire Time

ルータを表す LSA（Type 1）は、各 OSPF ルータが生成します。LSID には、この LSA を生成した Router ID が格納されます。ルータが接続しているそれぞれのネットワーク毎に、次の内容が含まれます。

- Link ID：隣接リンクの ID など
- Link Data ：リンクの IP アドレスなど
- コスト
  Type of Service 毎に複数のコストを格納できますが、TOS=0 以外は使われていません。

マルチアクセスネットワーク（2 台以上のルータが存在すること）を表す LSA（Type 2）は、そのネットワークの DR が生成します。LSID には、この LSA を生成した DR の IP アドレスが格納されます。ネットワークのネットマスクと、接続されているルータの Router ID のリストが含まれます。

4.8 メトリック

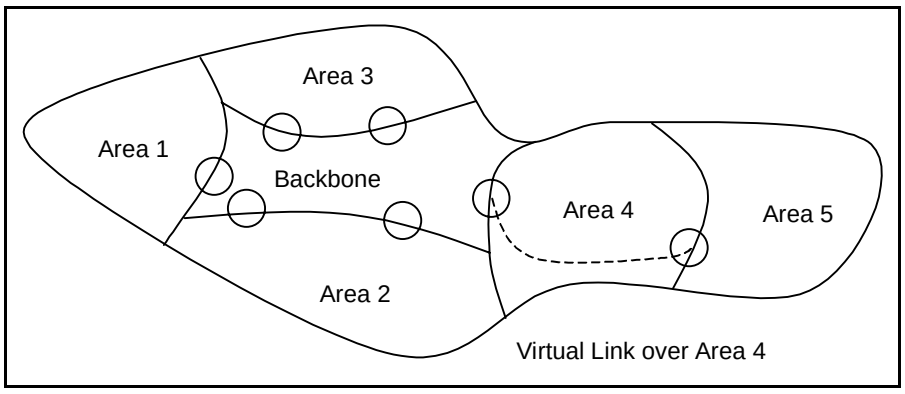
OSPF のコスト計算に使用するメトリックは、リンクの出力側に対して設定され、入力のコストは 0 と考えます。Ethernet ではルータ毎に、シリアルリンクでは方向別に値を設定します。メトリック値は 1 から 65535 の正数であり、「10^7／リンクの帯域（bps）」が推奨値とされていますが、あくまでも参考値です。

なお、同じコストの異なる経路が存在する場合は、ルータの実装によりますが、パケット毎に負荷分散を図ることができます。

4.9 エリア

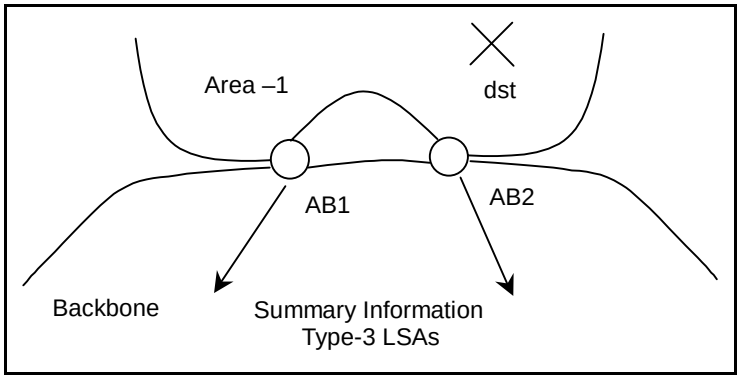
OSPF では、経路計算を簡易化するために、階層化の概念が導入されています。AS（経路ドメイン）をエリアに分割して、それぞれに 32 ビットのエリア ID を付けます。それぞれのネットワークは単一のエリアに所属し、ルータは複数のエリアに所属することもあります。あるエリアの内部の経路は、厳密に Dijkstra 法によって経路計算を行い、エリア外部にはサマリ情報のみを通知することで、経路計算のオーバーヘッドを低減します。

エリアの中で、エリア ID が 0 のものを、バックボーンエリアと呼びます。あるエリアから別のエリアに至るには、必ずバックボーンエリアを経由します。すなわち、バックボーンエリアは、全てのエリアに隣接していなければなりません。図のように、バックボーンエリアに隣接しないエリアが存在する場合には、バーチャルリンクを使用してバックボーンエリアを拡張します。



これにより、OSPF におけるルータは、次の 4 種類に分けられることになります。

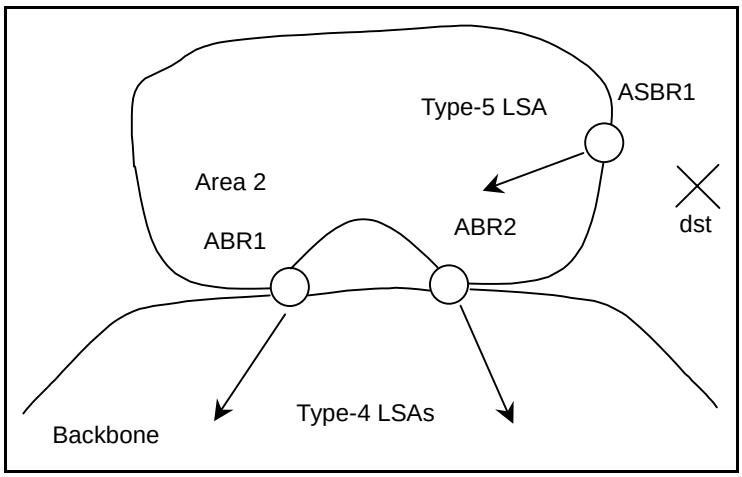
- 内部ルータ
  単一のエリアのみに所属するルータです。
- バックボーンルータ
  バックボーンエリアに所属する、内部ルータ、またはエリア境界ルータです。
- エリア境界ルータ
  バックボーンと他のエリアを接続するルータです。サマリ情報として、LSID にエリア内の各宛先を LSID として持つ Type-3 LSA を生成します。すなわち、バックボーンエリアおよび他のエリアには、エリア境界ルータのいずれかを選択するための LSA のみが通知され、エリア内部の構造は伝達されません。



また、エリア境界ルータは、エリアの内部に存在する AS 境界ルータへのサマリ情報として、AS 境界ルータの Router ID とそこまでのコストを持つ Type-4 LSA を生成します。

- AS 境界ルータ
  AS と他の AS を接続するルータで、AS 外部の宛先毎に、そのアドレスを LSID とする Type-5 LSA を生成します。Type-5 LSA のメトリック値には 2 種類あります。1 つは、外部から知らされたメトリック値と内部のメトリック値が比較可能で、メトリックの合計が小さい経路を選択するための Type-1 メトリックです。もう 1 つは、外部から知らされたメトリックが支配的で、内部メトリックを考慮しない Type-2 メトリックです。Type-1 メトリックと Type-2 メトリックを混在することができますが、ほとんどの場合は、どちらか一方のみが使用されます。

  AS 境界ルータが生成した Type-5 LSA は、（stub エリアを除く）他のエリアに伝達されます。また Type-5 LSA を生成したルータの情報は、エリア境界ルータにて Type-4 LSA、すなわちサマリ情報として他のエリアに通知されます。



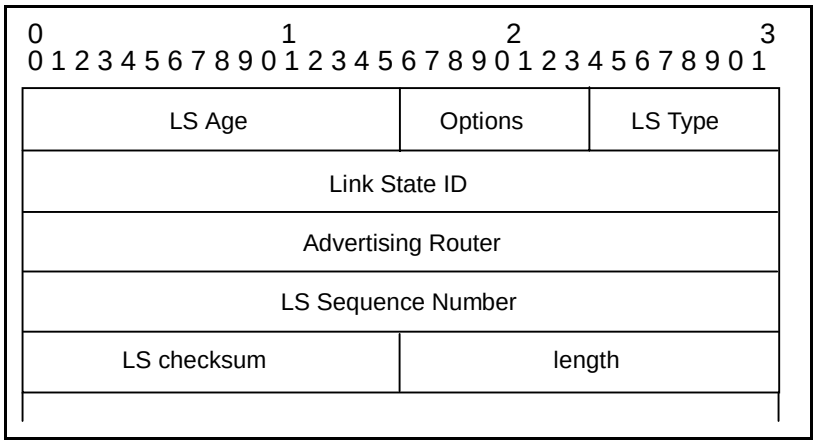
OSPF では、エリア内部のルータ全てが合意した場合に、エリアをスタブエリアとすることができます。スタブエリアでは、外部経路を取り扱わず、デフォルト経路を使用するので、Type-5 LSA は伝達しません。

4.10 LSA の構造

ここから、LSA の構造を解説していきましょう。

4.10.1 LSA Header

全ての LSA には、次のヘッダが共通して付けられます。

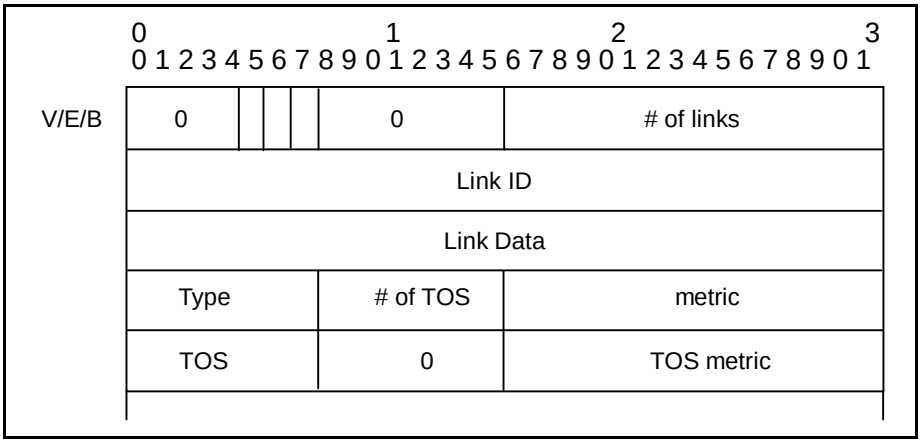


LS Type には次の値が格納されます。

- Type-1：Router LSA
- Type-2：Network LSA
- Type-3：Summary LSA （IP network）
- Type-4：Summary LSA （AS 境界ルータ）
- Type-5：AS-External LSA

4.10.2　Type-1 LSA

ルータを表す Type-1 LSA は、次の構造を持っています。



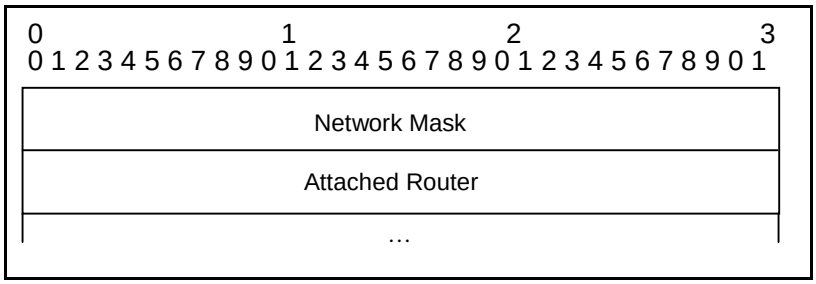
type フィールドの値と意味は、次のとおりです。

- 1：point-to-point リンク
  隣の Router ID
- 2：transit network（OSPF ルータが複数存在するネットワーク）
  DR のアドレス
- 3：stub network（OSPF ルータが 1 つしか存在しないネットワーク）
  ネットのネットワーク部
- 4：virtual link
  隣の Router ID

Link Data の値には、自分のアドレスが使われます。bit v は virtual link の一方である場合、bit E は AS 境界ルータである場合、bit B はエリア境界ルータである場合に、それぞれ 1 となります。

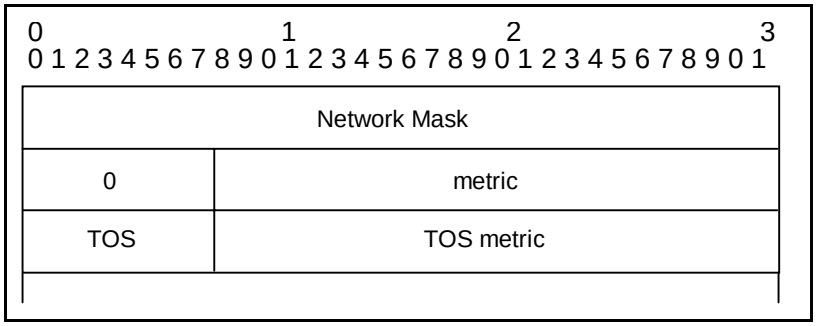
4.10.3 Type-2 LSA

ネットワークを表す Type-2 LSA は次の構造を持っています。Type-2 LSA は DR が生成し、そのネットワークに接続しているルータを表しています。



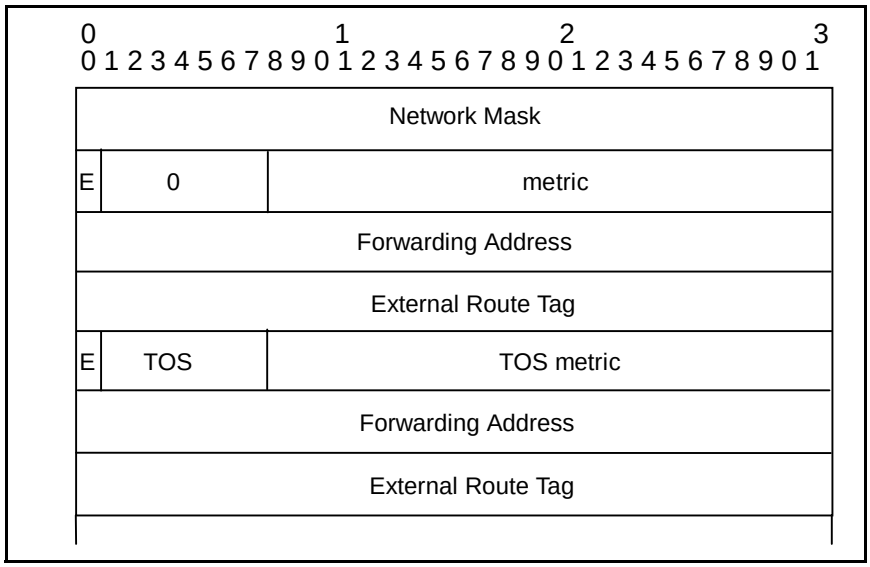
4.10.4 Type-3 LSA/Type-4 LSA

サマリ情報を表す Type-3 LSA と Type-4 LSA は、次の構造を持っています。



4.10.5 Type-5 LSA

AS 外部の経路を表す Type-5 LSA は次の構造を持っています。メトリックにおいて、E が 0 の場合は type-1 メトリックを、E が 1 の場合は type-2 メトリックを意味します。スタブエリアには伝送されません。



4.11 認証

ルーティングプロトコルは、インターネットの経路制御を決定するものです。ルータにはホストと同程度のセキュリティが求められるので、何らかの認証機能が必要になります。秘密の情報ではありませんので暗号化は必要ありません。平文パスワードによるパスワード認証では、Reply 攻撃に弱いので好ましくありません。そのため、暗号化シーケンス番号を使用し、MD5 の署名を付加するなどの暗号化された認証が準備されています。

4.12 OSPF と CIDR

RFC1583（CIDR）が出るまでの間は、マスクが異なる同一宛先を表現できなかったため、CIDR に完全に対応していませんでした。RFC1583 によって、LSID としてホスト部のビットを立てて良いことになり、CIDR に対応できるように改良されました。

4.13 Point to Multi Point

最近の OSPF では、ATM やフレームリレーのように、ブロードキャストができないネットワークを抽象化するために、Point to Multipoint を扱う方法が導入されています。従来は、Point to Point が多数あるような抽象化を行いましたが、Point to Multipoint を使うことで、よりスマートな抽象化が行われています。実際には、ルータのドキュメントをよく読み、テストを行ってから使うのが良いでしょう。

5 OSPF の運用

Cisco のルータにおける OSPF の設定例を示します。アドレスの一部を文字「#」で隠してあります。

5.1 インタフェースの設定

```
interface ethernet 0/0
 ip address 192.168.1.1 255.255.255.0
 ip ospf authentication-key himitsu ←"service password-encryption" を指定すると、暗号化されます。
 ip ospf cost 10 ←デフォルト値ですから不要です。
 ip ospf priority 1 ←デフォルト値ですから不要です。
 ip ospf hello-interval 10 ←デフォルト値ですから不要です。
 ip ospf dead-interval 40 ←デフォルト値ですから不要です。
```

5.2 OSPF プロトコルの設定

```
router ospf 100 ← OSPF の利用を宣言
 network 192.168.1.1 0.0.0.0 area 0.0.0.0 ←エリア 0。
 area 0.0.0.0 authentication ←エリア 0 では認証を使用。
```

5.3 インタフェースの確認

```
% show  ospf interface
Ethernet0/0/0 is up, line protocol is up
  OSPF not enabled on this interface
Fddi5/0.0 is up, line protocol is up
  Internet Address 203.178.137.##/29, Area 0.0.0.0
  Process ID 2500, Router ID 203.178.136.#, Network...
  Transimit Delay is 1 sec, State BDR, Priority 1
  Designated Router (ID) 203.178.136.#, Interface
  Backup Designated router (ID) 203.178.136.#, Inte...
  Timer intervals configured, Hello 10, Dead 40, W..
    Hello due in 00:00:07
  Neighbor Count is 1, Adjacent neighbor count is ..
    Adjacent with neighbor 20.178.136.#  (Designat..
```

5.4 Neighbor の確認

```
% show ospf neighbor
Neighbor ID  Pri State   Dead Time Address
203.178.136.#  1 FULL/DR 00:00:35 203.178.137.##
150.##.0.1     1 FULL/ - 00:00:35 203.178.136.###
203.178.141.## 1 FULL/ - 00:00:38 203.178.141.##
203.178.136.#  1 FULL/ - 00:00:32 203.173.141.##
203.178.141.## 1 FULL/ - 00:00:36 203.178.141.##
203.178.141.## 1 FULL/ - 00:00:30 203.178.141.##

% show ospf neighbor
Neighbor ID   Pri State        Dead Time Address
203.178.136.###  1 2WAY/DROTHER 00:00:55 203.178.136.###
203.178.137.##   1 FULL/BDR     00:00:56 203.178.136.###
203.178.136.###  1 2WAY/DROTHER 00:00:59 203.178.136.###
203.178.136.###  1 2WAY/DROTHER 00:00:55 203.178.136.###
203.178.136.###  1 FULL/DR      00:00:59 203.178.136.###
203.178.136.###  1 2WAY/DROTHER 00:00:55 203.178.136.###

% show ospf neighbor detail
Neighbor     203.178.136.###,     interface     address
203.178.136.###
  In the area 0.0.0.0 via interface Serial1/0
  Neighbor priority is 1, State is FULL
  Options 2
  Dead timer due in 0:00:59
Neighbor     203.178.142.###,     interface     addrsss
203.178.142.###
  In the area 203.178.136.### via interface Serial1/3
  Neighbor priority is 1, State is FULL
  Options 2
  Dead timer due in 0:00:59
  Link State retransmission due in 00:00:04
      LSA in retransmission queue 2
```

OSPF で経路情報をやりとりしたいルータ同士が Neighbor にならない場合、その原因として、エリア ID、ネットマスク、タイミングパラメータ、認証方式、認証、などのパラメータが一致していないことが考えられます。状況を確認するには、次のコマンドを実行して、デバッグメッセージを表示させます。

```
# debug ip ospf adj
```

5.5 Database の確認

```
% show ip ospf database database-summary
Area ID         Rtr Net  SumNet SumASR Subtotal
0.0.0.0          61  17    46     48    172
203.178.136.### 2    0    97     70    169
AS External                          4452
Total            63  17   143    118   4793

% show ip ospf database
                Router Link States (Area 0.0.0.0)
Link ID         ADV Router     Age Seq#     Checksum Link
133.27.2.##     133.27.2.###   340 0x80000A44 0x11F9  1
150.65.0.##     150.65.0.###   999 0x8000613E 0xEEF4  5
202.249.3.##    202.249.3.##   622 0x800015B4 0x53C0  1
202.249.3.##    202.249.3.##   666 0x80001638 0x2760  1
202.178.136.## 203.178.136.#  961 0x80004965 0x65D6  3
202.178.136.## 203.178.136.# 1201 0x8000144A 0x35EB  2
202.178.136.## 203.178.136.#  690 0x800054DA 0x7682  3
202.178.136.## 203.178.136.#  949 0x800018D2 0x1CCB 11

% show ip ospf database network 203.178.137.##
 Routing Bit Set on this LSA
 LS age: 854
 Options: (No TOS-capability)
 LS Type: Network Links
 Link State ID: 203.178.137.## (address of DR)
 Advertising Router: 203.178.136.#
 LS Seq Number: 8000001E
 Checksum: 0x36B2
 Length: 32
 Network Mask: /29
```

```
        Attached Router: 203.178.136.###
        Attached Router: 203.178.136.###

% show ip ospf database database-summary
% show ip ospf database
% show ip ospf database router <LSID>
% show ip ospf database network <LSID>
% show ip ospf database summary <LSID>
% show ip ospf database asbr-summary <LSID>
% show ip ospf database external <LSID>
```

5.6 OSPF 設定ミスの例

OSPF の設定ミスの例をいくつか紹介しましょう。

- バックボーンに隣接しないエリアを作ってしまう場合があります。バーチャルリンクやエリア分割を見直します。なお、1 つのエリアに含まれるルータ数は、数十にしておいた方が良いでしょう。
- ほとんどのルータは､OSPF でインターネットのフルルートを扱う能力がありません。
- 同じ Router ID を複数のルータに割り振った場合には、経路計算を誤り、ループが発生するなどの症状や、シーケンス番号が急激に上昇する症状が現れます。Router ID を訂正して、1 時間ほど待つと正常状態に復旧します。

5.7 Tips

運用上の Tips をいくつか紹介しておきます。

- 単一エリアに属するルータにおいて、全てのインタフェースを同じエリアにするには、次のコマンドを実行します。

```
router ospf 100
 network 0.0.0.0 255.255.255.255 area <AREA>
```

- 他にルータのないネットワークでは、Hello を止めるために、次のコマンドを実行します。

```
router ospf 100
  passive-interface ethernet 0/0
```

- OSPF 以外の経路を導入するには、次のコマンドを実行します。

```
router ospf 100
 redistribute rip route-map rip-ospf subnet
access-list 1 permit any
route-map rip-ospf permit 10
 match ip address 1
 set metric 1000
 set metric-type type-1
```

- OSPF 以外の経路を再導入するには、次のコマンドを実行します。

```
clear ip ospf redistribute
```

6 まとめ

最後に、RIP と OSPF の特徴を、もう 1 度まとめておきましょう。

- RIP
  限られた場面では、依然として有効なルーティングプロトコルです。ただし、大きなネットワークで使ってはいけません。
- OSPF
  プロトコルは複雑ですが、最近の実装は概ね安定しています。運用はそれほど難しくありませんが、デバッグはちょっと大変です。

7 参考文献

- 「Interconnections」Radia Perlman
  Addison-Wesley, ISBN 0-201-56332-1
  （和訳「Interconnections」加藤 朗監訳、ソフトバンク ISBN 4-89052-712-5/品切れか）
- 「Internet Routing Architectures」Bassam Halabi
  Cisco Press, ISBN 1-56205-652-2
- 「Routing in the Internet」Cristian Huitema
  Prentice Hall, ISBN 0131321927
- 「Managing IP networks with Cisco Routers」Ballew et al
  O’Reilly, ISBN 1565923200
- 「OSPF: Anatomy of an Internet Routing Protocol」John Moy
  Addison-Wesley, ISBN 0201634724

- 「Cisco Router OSPF - Design and Implementation Guide -」
  William R. Parkhurst
  McGraw Hill, ISBN 0070486263
- 「OSPF Network Design Solutions」 Thomas M., II Thomas
  Cisco Systems, ISBN 1578700469