ルートリフレクタの開発

Development of Route Reflector

遠藤淳一*
Junichi Endo

概　要　大規模なネットワークを構築・運用している通信事業者やISP（Internet Service Provider）は，網の拡張性や安定性の確保が必須である。通信事業者のIP-VPN網やISPのインターネット網ではBGP（Border Gateway Protocol）を使用しているが，AS（Autonomous System）内で使用するiBGP（internal BGP）は，規模拡張性に問題を持つ。この問題の対処のために，これらの網ではルートリフレクタを導入している。本稿では，大規模ネットワークに対して十分な規模拡張性を持ち，また，高速性能を備えたルートリフレクタの開発について報告する。

1. はじめに

大規模なネットワークを構築・運用している通信事業者やISP（Internet Service Provider）は，網の拡張性や安定性の確保が必須である。通信事業者のIP-VPN網やISPのインターネット網で使用されているBGP（Border Gateway Protocol）プロトコルは，大規模ネットワークにおいては数十万にも及ぶ経路を交換する必要がある。また，AS（Autonomous System）内で使用されるiBGP（internal BGP）は，プロトコル仕様上，あるiBGPピアから受信した経路情報を他のiBGPピアに送信してはならない。すなわち，iBGPを使って全ての経路情報を共有するためには，iBGPセッションをフルメッシュに接続することが必要となる。このため，経路交換する相手の数が増えてくると経路交換負荷が飛躍的に増大し，通信網全体のパフォーマンス低下させてしまう（**図1**：iBGPフルメッシュ問題）。

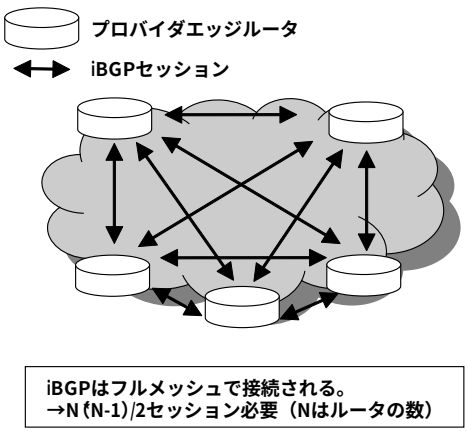

図1　iBGPフルメッシュ問題
Network congestion due to iBGP full mesh problem

この問題を対処するために，IP-VPN網やインターネット網では，既存のBGPルータが持つ機能を利用して，経路情報を集約／再配布するルートリフレクタを導入している。ルートリフレクタに接続し，リフレクタから経路を再配布されるiBGPピアをクライアントピアと呼ぶ。ルートリフレクタはiBGPピアから経路を受信した場合，ピアのタイプによって次の動作を行う[1)]。

1）非クライアントピアからの経路
　全てのクライアントピアに経路反射する
2）クライアントピアからの経路
　全ての非クライアントピアに加え，クライアントピアに経路反射する

この動作により，ルートリフレクタとクライアントピアはiBGPフルメッシュと同等の経路情報共有を可能とする。ルートリフレクタを導入することで，各iBGPピアはフルメッシュ接続の代わりにルートリフレクタを中心としたスター型の接続構成をとることができる。その結果，各ルータは保持すべきBGPセッションの数を減らすことができ，各ルータにかかる負荷を軽減することが可能となる（**図2**：ルートリフレクタの導入）。しかしながら，網の拡大に伴う経路数の増大や接続されるルータの増加により，既存のBGPルータの流用という手法では拡張性・安定性において対処が難しくなっている。

本稿では，大規模ネットワークに対して十分な規模拡張性を持ち，また，高速性能を備えたルートリフレクタの開発について，大規模ネットワークに対応した評価環境，性能向上をポイントとして報告する。

2. 市場分析

前項で述べた通り，ルートリフレクタは大規模ネットワークでの使用を想定している。ネットワークの規模は，その網で使用している経路数で表すことができる。2002年度末時点では，全世界に接続されているインターネットの規模（経路数）が十

* 研究開発本部 ファイテルネットワーク研究所

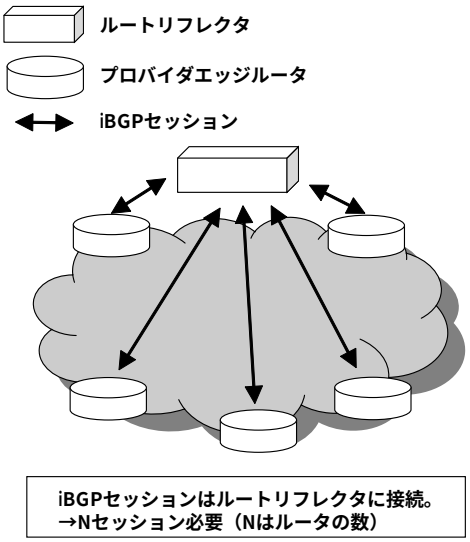

図2　ルートリフレクタの導入
New network configuration by introduction of a route reflector

数万経路であるのに対し，IP-VPN網（1事業者）の規模はすでにインターネットの規模を超え，20万経路に達する見込みである。IP-VPN網においては，ユーザ毎の経路情報を別々に管理しており，インターネットで実施されている経路集約技術を利用することができない。このため，接続ユーザ数の増加に従って経路数も増大する。例えば，あるIP-VPN網に200ユーザが接続しているとして，各ユーザが1,000経路使用している場合，IP-VPN網の経路数は20万経路（200×1,000）となる。現状，ルートリフレクタとして使用されているルータは20万経路の経路情報を処理するのが上限であり，IP-VPN網においては，ルートリフレクタの強化が急務となっている。

市場の課題を整理すると，次の2つが挙げられる。

（1）増大する経路数への対応

（2）収束時間の増大に対する対処

（1）の対応については，使用しているルータをさらに高価なバックボーンルータに置き換えることで対処可能ではあるが，（2）の収束時間の増大に対する対処にならないうえ，設備投資の増大を引き起こしてしまう。これらの課題を解決可能な規模拡張性および高速性能を備えるルートリフレクタを市場に供給することが重要である。

3.　性能評価環境の確立

ルートリフレクタ開発にあたって，第一の課題として挙げられたのは性能評価環境についてである。ルートリフレクタは，従来開発していたルータとは市場の中での位置付けが異なるため，従来とは桁違いの経路情報および接続相手が目標となっている。このため，従来の評価環境ではルートリフレクタの性能評価を行うことができない。すなわち，開発したルートリフレクタが，開発目標を満たしているかどうかの判断を行うことができない。そこで，新規に性能評価環境を確立する必要がある。

例えば「100万経路（50万×2冗長）を100ピアに再配布する場合の経路収束時間」を評価するための環境は，次に挙げるプロトコルの経路情報交換をシミュレーション可能な装置が100台必要となる。

- iBGPによる経路情報交換
  （以下iBGPの機能として必要）
- IP-VPN経路を5,000経路以上生成し，iBGPピアに対して送信できること
- IP-VPN経路にRD（Route Distinguisher），SOO（Site Of Origin），tagなどのアトリビュート情報を付ける事ができること

上記を満たすことのできる装置は，IP-VPN網で使用しているルータ装置となるが，ルータ装置は高額であるため100台購入すると莫大な費用が必要となってしまう。そこで今回，上記を満たしている，当社ルートリフレクタで開発したルーティングソフトウェアをPC上で動作させることを基本方針とし，安価ではあるが大規模ネットワークをシミュレーション可能なシミュレータ装置を開発することとした。PC上で動作させるOSは，本開発のルートリフレクタでも使用しているNetBSDを採用した。NetBSD上に1台分のシミュレータ機能を搭載するだけでは，計100台のPCが必要となり，ルータ装置を並べるよりは安価だが，取り扱いが非常に大変になる。

このため，NetBSD上に複数のシミュレータ機能を搭載できるよう，1物理インタフェース（Fast Ethernet）上に複数のIPアドレスを設定（ifconfig alias 設定）し，また，使用するIPアドレスを指定できるようにルーティングソフトウェアの改良を行った。この改良によって，1台のNetBSD上で複数ルータのシミュレータ機能を動作できるようになった。

使用するPCのハードウェアスペックは，CPUクロック：1 GHz，メモリ容量：1 GBとし，さらに，1台のPC上で動作させるルートリフレクタを10台とすることで，評価対象機であるルートリフレクタに対して十分な性能を確保できる構成とした（本開発のルートリフレクタのCPUクロックは700 MHz，メモリ容量は768 MBである）。これにより，100台分の経路制御は10台のPC（以降，「シミュレータ装置」と記述する）でシミュレーション可能となった。シミュレータ装置構成は**図3**のように表される。

さらに，シミュレータ装置が使用するBGPDでは，シミュレーション用の経路生成や，その経路に付随するアトリビュート情報が設定可能でなければならない。これらの機能をBGPD上に新たに開発し，既存のBGPD設定処理部分を利用することで経路およびアトリビュート情報を設定可能とした。

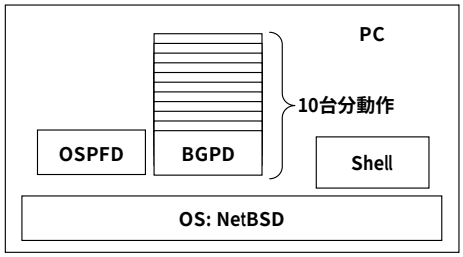

図3　シミュレータ装置構成
Composition of a simulator

```
!
router bgp 7675
 bgp router-id 192.168.1.1
 neighbor 172.16.1.3 remote-as 7675
 neighbor 172.16.1.3 update-source 192.168.1.1
 neighbor 172.16.1.3 timers connect 3
!
 address-family vpnv4 unicast
 neighbor 172.16.1.3 activate
 neighbor 172.16.1.3 send-community extended
 neighbor 172.16.1.3 route-map no-in in
  network 1.0.1.0/24 rd 1:1 tag 1 rt 1:1 soo
65001:1234567891
  network 1.0.2.0/24 rd 1:2 tag 2 rt 1:2 soo
65001:1234567891
  network 1.0.3.0/24 rd 1:3 tag 3 rt 1:3 soo
65001:1234567891
  network 1.0.4.0/24 rd 1:4 tag 4 rt 1:4 soo
65001:1234567891
  network 1.0.5.0/24 rd 1:5 tag 5 rt 1:5 soo
65001:1234567891
```

図4　シミュレータ装置設定例
Example of simulator configuration

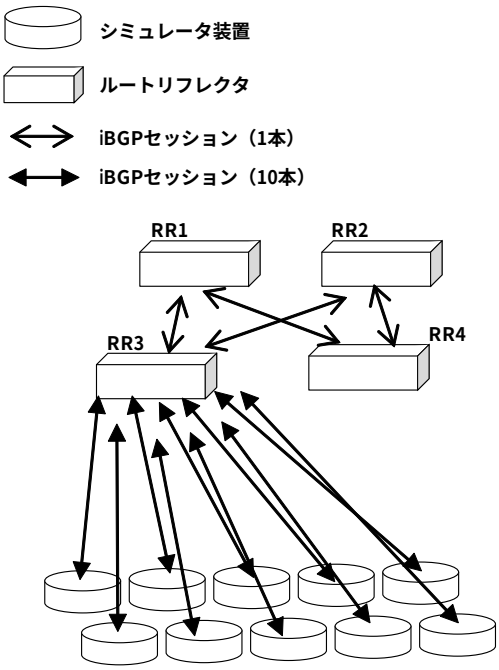

図5　ルートリフレクタ評価環境
Evaluation environment for route reflector

図4は，シミュレータ装置のBGPD設定例である。“ network ”以降が経路及びアトリビュート情報の設定となる。このnetworkコマンドを5,000行書くと，BGPDに5,000経路の経路情報が導入され，ルートリフレクタにその5,000経路を送信することが可能となる。

また，**図5**は今回構築したルートリフレクタ評価環境である。評価対象機はRR3である。RR3は，RR1, RR2とiBGP接続を行い，冗長構成をとり，また，RR1, RR2は，RR3, RR4をそれぞれルートリフレクタクライアントとして設定し，ルートリフレクタの階層化（RR1, 2を“ 上位RR ”,RR3, 4を“ 下位RR ”と呼ぶ）を行っている。図5には示していないが，RR4とシミュレータ装置の間においてもiBGPを接続しており，シミュレータ装置が生成した経路情報は，まず，RR4に展開される。次にRR4から，RR1, RR2にそれぞれ経路情報が転送され，最後にRR3に対してRR1, 2の両ルータから経路情報が渡される構成となっている。RR3としての性能は，RR1, RR2からの経路受信処理，および，受信経路をシミュレータ装置（100台）に反射（フィルタリング設定にてpermitとなった経路のみを反射）する処理の全ての処理の合計時間で表すことができる。

4.　性能向上

BGPプロトコルの一般的な実装では，ピアとのセッションが切れた場合，そのピアから受け取った経路の削除処理（経路情報テーブルからの削除と他ピアへの削除通知）を行うが，その間に再度ピアからの接続要求が来てもそれを拒否し，完全に経路の削除処理を終了してからその接続要求に応える（**図6-1**）。経路数が少ない場合は，経路の削除処理を完了してから接続要求に応えても，経路の削除処理に大した時間を要しないため，大きな問題とはならない。しかし，経路数や送信ピア数が増加するとともに経路の削除処理にかかる時間も増加し，100万経路といった大規模ネットワークにおいては無視できないほどの時間となる。そしてその削除処理が終了するまでは次の接続を行うことができない。実際に，当社の性能評価において，100万経路を受信し80ピアに対して10万経路ずつ送信（合計800万経路送信）した状態からピアとのセッションを切断した場合，再接続するのに約5分間必要であることが判明した。そこで，再接続までの時間を短縮するために削除処理の変更を検討した。

削除処理を分析すると，そのほとんどの時間が削除情報を他ピアへ送信する時間である。削除処理を変更することによって，プロトコルに矛盾が発生してしまってはならない。今回の場合，削除情報を他ピアへ送信する処理を即時に行わなくてもプロトコルに矛盾は発生しないことに着目した。そこで，即時に送信処理を行わなくても良いように，削除された経路情報にトークンを設定し無効化する処理を追加した。また，全経路の無効化を終えた後に削除情報を他ピアへ送信するように変更した。無効化する処理はフラグ操作を行うだけであるので，ピアからの再接続要求に即座に応えることが可能となった（**図6-2**）。

6.　まとめ

本開発によって大規模ネットワークに対応した評価環境を安価に構築することができた。また，この評価環境を使用することで，高性能かつ優れた規模拡張性を持つルートリフレクタを開発した。今回開発したルートリフレクタは，100万経路（50万×2冗長）の経路収束時，**表1**の収束性能を達成した。

表1　収束性能
Convergence time performance

BGPセッション再接続時	reset実行時	電源off／on時
12分	15分	15分

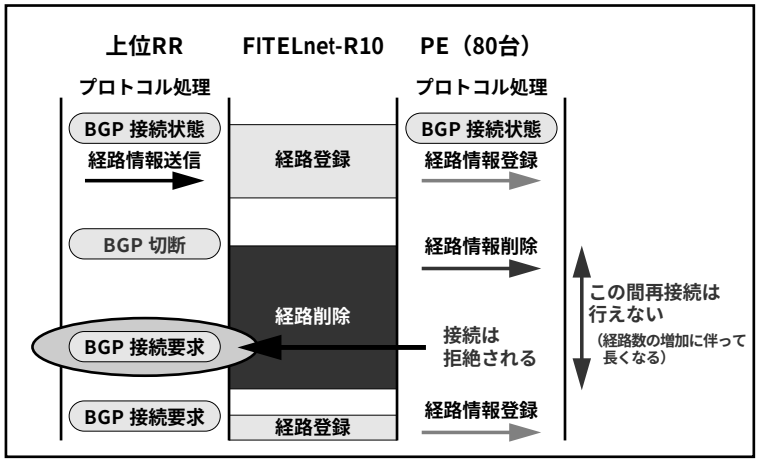

図6-1　一般的な削除処理
General deletion processing

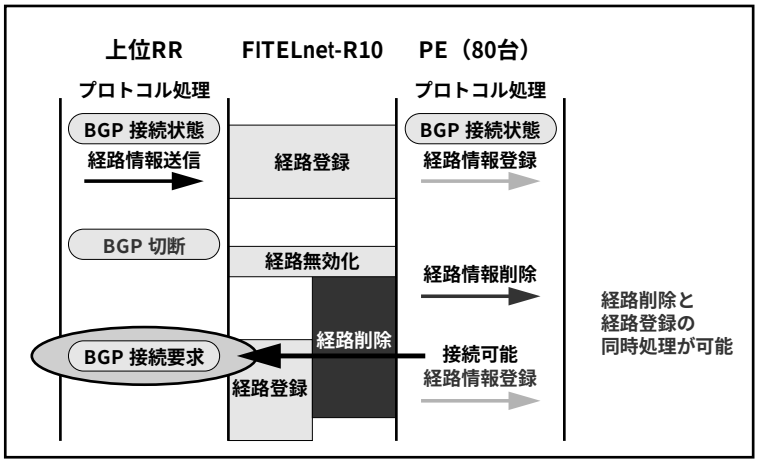

図6-2　変更後の削除処理
Deletion processing after change

図7　評価環境設備
Equipment for evaluation environment

図8　ルートリフレクタ外観
Appearance of FITELnet®-R10

7. おわりに

今回開発したルートリフレクタ用ルーティングソフトウェアは，規模拡張に優れ，かつ，高性能であるため，当社ルータ（メトロエッジルータ）に流用可能である。評価環境についても同様に次期開発時での活用が期待される。これらのルートリフレクタ開発で蓄積された技術を応用することで，先端的な製品を迅速に開発できると確信している。

参考文献

1）Bates, T., Chandra, R. and E. Chen: “BGP Route Reflection An Alternative to Full Mesh IBGP,” RFC 2796, **4**（2000）, 5.