Development of the FITELnet-R10 Route Reflector

by Jun'ichi Endo [*]

ABSTRACT For the carriers and Internet service providers (ISPs) who build and operate large-scale networks, it is essential to have network scalability and stability. Border gateway protocols (BGPs) are used in carrier IP-VPNs and ISP networks, but the internal BGPs (iBGPs) that are used within autonomous systems have a problem in terms of scalability. To address this problem, route reflectors are being introduced in these networks. In this paper we report on the development of the FITELnet-R10, a route reflector that offers adequate scalability for large-scale networks, while providing high-speed performance.

1. INTRODUCTION

For the carriers and Internet service providers (ISPs) who build and operate large-scale networks, it is essential to have network scalability and stability. The BGPs used in carrier IP-VPNs and ISP networks must be capable, in large-scale networks, of exchanging hundreds of thousands of routes. Further, the iBGPs used within autonomous systems must, in terms of protocol specifications, be able to transmit the route information received from one iBGP peer to another. This means that full-mesh connection of iBGP sessions is needed to share route information using an iBGP. Thus as the number of routings to be exchanged increases there is a quantum increase in the routing exchange load, leading to degradation in the performance of the communications network as a whole (see Figure 1).

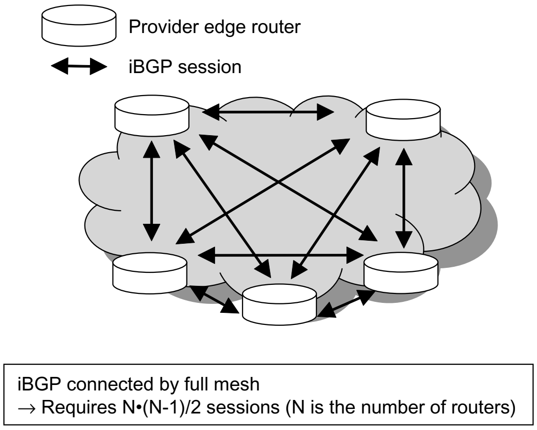
Provider edge router
iBGP session
iBGP connected by full mesh
→ Requires N•(N-1)/2 sessions (N is the number of routers)

Figure 1 iBGP full-mesh problem.

To address this problem, route reflectors are being introduced into IP-VPN networks and Internet networks to aggregate and re-distribute route information taking advantage of the functions of existing BGP routers. An iBGP peer that is connected to the route reflector and is re-distributed by routing from the reflector is known as a client peer. When the route reflector receives a routing from an iBGP peer, it can, depending on the type of peer, carry out the following operations: 1)

1) routing from non-client peers: route reflection to all client peers;
2) routing from client peers: route reflection to client peers in addition to all non-client peers.

By this operation, it is possible for the route reflector and client peers to share route information in a manner equivalent to iBGP full mesh. By the introduction of a route reflector, instead of full-mesh connection of each iBGP peer, it is possible to create a star topology connection centered on the route reflector. As a result it is possible to reduce the number of BGP sessions that must be held by each router, thereby lightening the load on each router (see Figure 2).

Due to the increase in the number of routes and the number of routers connected associated with network expansion, however, it is difficult to address the problems of scalability and stability with any method that continues to make use of existing BGP routers.

In this paper we report on the development of a route reflector that offers adequate scalability for large-scale networks while providing high-speed performance, and focus on an evaluation environment and performance upgrade responsive to large-scale networks.

[*] FITEL-Network Lab., R&D Div.

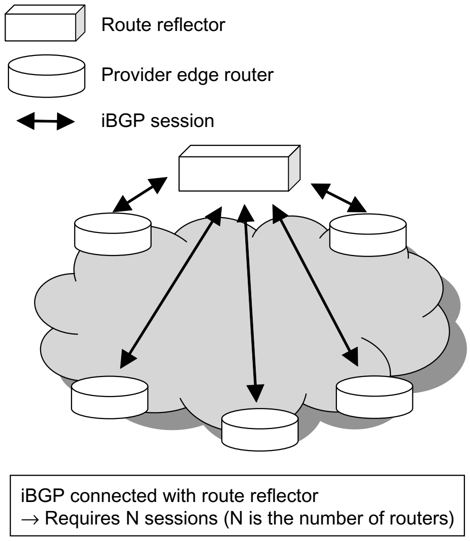
Route reflector
Provider edge router
iBGP session
iBGP connected with route reflector
→ Requires N sessions (N is the number of routers)

Figure 2 Introduction of a route reflector.

2. MARKET ANALYSIS

As was pointed out in the previous section, route reflectors assume the use of large-scale networks, and the scale of a network is defined by the number of routes used in it. As of the end of March 2003, the scale of the Internet worldwide (number of routes) is some 120,000, whereas the scale of an IP-VPN (for a single carrier) is about to exceed 200,000 routes, surpassing the scale of the Internet.

In an IP-VPN, route information for each user is managed individually, and it is not possible to use the route aggregation technology that is implemented on the Internet. For this reason the number of routes increases in accordance with the increase in the number of users connected. In a case, for example, in which there are 200 users on a certain IP-VPN and each of them uses 1,000 routes, the number of routes handled by the IP-VPN will be 200 × 1,000, or 200,000. At present 200,000 is the upper limit for the route information that can be handled by the routers that are used as route reflectors, so that more introducing powerful route reflectors is an urgent need in IP-VPNs. The issues facing the market may therefore be stated as:

1) responding to the growing number of routes; and
2) responding to the increase in convergency time.

With respect to item 1), it would be possible to replace the routers in use with higher-cost backbone routers, but this not only would fail to address the increase in convergency time in item 2) but also would result in increased capital cost. It is therefore important to supply the market with route reflectors with the scalability and high-speed performance to provide solutions to these problems.

3. ESTABLISHING A PERFORMANCE EVALUATION ENVIRONMENT

The first problem we focused on here in developing the route reflector was the performance evaluation environment.

Since the market positioning of the route reflector differs from that of routers developed in the past, our target was route information and connection partners an order of magnitude different from before. This made it impossible to evaluate the performance of the route reflector using the conventional evaluation environment. That is to say, it would not be possible to determine whether the route reflector would meet the development targets. This made it necessary to establish a new performance evaluation environment. For example, an environment capable of evaluating the route convergency time required to re-distribute 1 million routes (500,000 × double redundancy) to 100 peers would require 100 units capable of simulating route information exchange under the following protocol:

o route information exchange by iBGP (required in the following as a function of the iBGP);
o Ability to create a minimum of 5,000 IP-VPN routes and transmit them to iBGP peers; and
o Ability to attach attribute information such as RD (route distinguisher) or SOO (site of origin) tags to IP-VPN routes.

A unit capable of satisfying the above requirements would be a router used in an IP-VPN, but the high cost of such routers would mean that purchasing 100 of them would involve enormous expense. We therefore decided on a policy of running the routing software developed for our route reflector, which satisfies the above requirements, on ordinary personal computers, to develop a system capable of simulating a large scale, yet inexpensive, network.

As the operating system running on the PCs, we adopted NetBSD, which is also used for our route reflector. Just to load the simulator functions for a single unit into NetBSD would require a total of 100 PCs. This, while cheaper than setting up a series of routers, would be extremely complex to operate. Accordingly we performed multiple IP address settings (ifconfig alias settings) on a single physical interface (Fast Ethernet) to enable loading of multiple simulator functions into NetBSD. We also upgraded the routing software to designate the IP addresses used.

By means of this upgrade it became possible to run multiple simulator functions on a single NetBSD. The hardware specifications for the PC were: CPU clock speed of 1 GHz with 1 GB of RAM. Further, we adopted a structure with adequate performance for the route reflectors to be evaluated, assuming 10 route reflectors operating on a single PC. (Our route reflectors have a CPU clock speed of 700 MHz with 768 MB of RAM.) In this way it became possible to carry out simulation of route control for 100 units using 10 PCs (hereinafter referred to as the simulator system). Figure 3 shows the

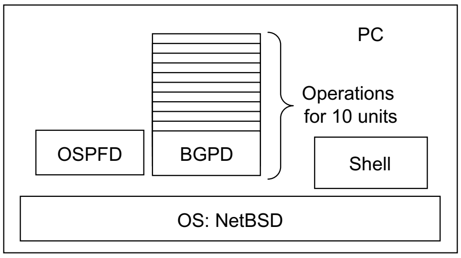

Figure 3 Structure of simulator system.

```
!
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
```

Figure 4 Example of configuration settings for BGP daemon of simulator system.

structure of the system.

In addition, the BGP daemon used by the simulator system must be capable of route creation for the simulation, as well as setting of the attribute information for those routes. The setting of routes and attribute information was made possible by developing these functions anew for the BGP daemon, and using the setting processing portion of the existing daemon.

Figure 4 shows an example of configuration settings for the BGP daemon of the simulator system. The coding following “network” sets routes and attribute information. Writing 5,000 lines of these network commands introduces route information for 5,000 routes into the BGP daemon, making it possible to send the 5,000 routes to the route reflector.

Figure 5 shows the route reflector evaluation environment that was configured in this work. The system to be evaluated is route reflector 3 (RR3). This is a redundant configuration in which RR3 provides iBGP connection with RR1 and RR2. RR1 and RR2 are set to client route reflectors RR3 and RR4, respectively, in a hierarchic relationship, so that RR1 and RR2 are known as “higher-order” route reflectors; RR3 and RR4 as “lower-order”.

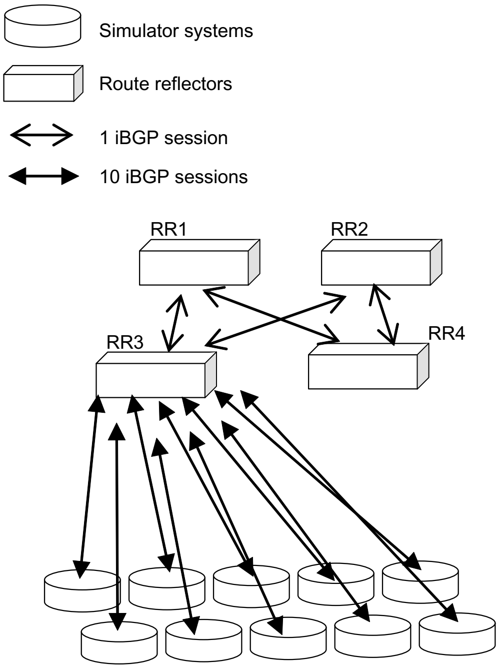

Figure 5 Route reflector evaluation environment.

Although not shown, even if the iBGP is connected between RR4 and the simulator system, the route information created by the simulator system is first transmitted to RR4. The configuration is such that route information is then transmitted from RR4 to RR1 and RR2 respectively and finally route information from both route reflectors RR1 and RR2 is handed over to RR3. The performance as RR3 can be represented as the total time for all processing: processing of route receiving from RR1 and RR2, and reflection of received routes (only those routes permitted at the filter setting) to the simulator systems (100 units).

4. PERFORMANCE ENHANCEMENT

In an ordinary installation of a border gateway protocol (BPG), when a session with a peer is interrupted, deletion of the route received from that peer (deletion from the route information table and notification of deletion to other peers) is carried out, and if during that period a repeat connection request comes from the peer it is denied, and responded to only after route deletion processing has been completed (Figure 6.1).

As long as the number of routes is small this presents no major problem, since even completing route deletion before responding to the connection request requires no great amount of time. However with increases in the number of routes and number of transmitting peers, the time required for route deletion also increases, and in a large-scale network of a million routes, becomes too long to be ignored. Also, until processing of that disconnection is completed, it is impossible to make the next connection.

In our own evaluation, it was found that in reality, when interrupting a session with a peer from a condition

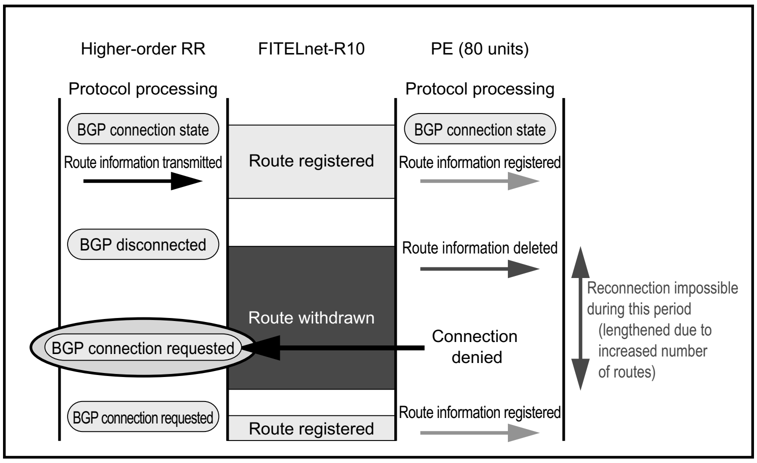

Figure 6.1 General deletion processing.

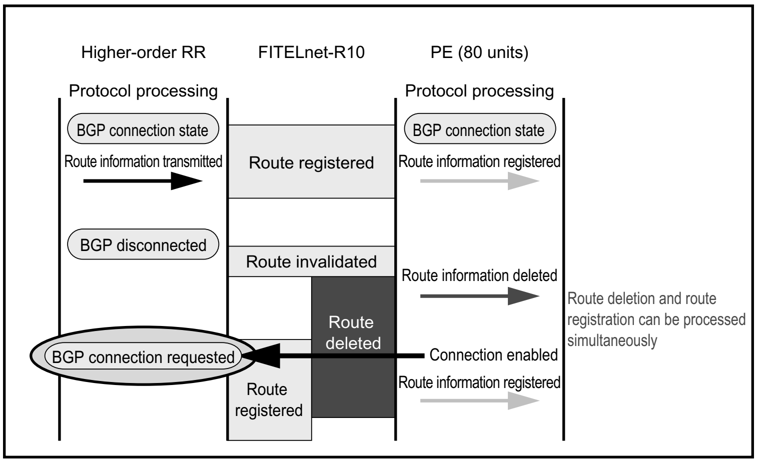

Figure 6.2 Deletion processing after change.

in which 1 million routes were received and 100,000 routes were transmitted to each of 80 peers (or a total of 8 million routes transmitted), reconnection required approximately 5 min. Accordingly we considered changing deletion processing as a means of shortening the time to reconnection.

In an analysis of deletion processing it was found that virtually all of the time was time for transmitting delete information to other peers. No protocol conflict can be allowed to occur due to changes in deletion processing. In the case under study, we realized that no protocol conflict will occur even if transmission of deletion information to other peers is not carried out immediately.

It was therefore decided to add a process to allow transmission processing not to be carried out immediately, whereby a token is set to invalidate the deleted route information. We also effected a change so that invalidation is finished before deletion information is transmitted to other peers. Since invalidation is merely a processing flag operation, it becomes possible to respond immediately to a reconnection request from a peer (Figure 6.2).

5. SUMMARY

As a result of the development work reported here, it has been possible to configure an evaluation environment for large-scale networks at a reasonable cost. And using this evaluation environment we have also developed a route reflector offering high performance and superior scalability.

In the convergency of 1 million routes (500,000 × double redundancy), the route reflector developed here offers convergency times as follows:

BGP session reconnection time:	12 min
Reset execution time:	15 min
Power supply on/off switching time:	15 min

6. CONCLUSION

The routing software for the route reflector developed in this work offers superior scalability and high performance, and can therefore be used in common with Furukawa Electric’s metro edge routers. It is anticipated that the evaluation environment will also be applicable in the next

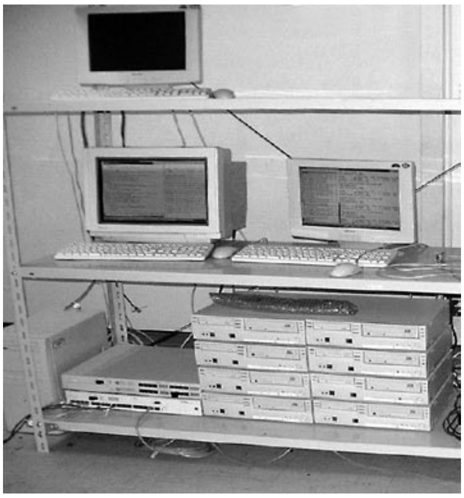

Figure 7 Evaluation environment equipment.

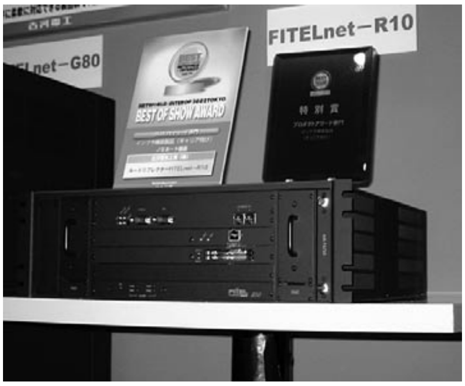

Figure 8 FITELnet-R10 route reflector.

phase of development.

We are confident that the expertise acquired in developing this route reflector will allow us to continue to bring out the most advanced products.

REFERENCE

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