Cut-through IP Forwarding Technology for a Terabit-class Super-network

Takeshi Yagi†, Kenichi Matsui, Yuuichi Naruse, and Junichi Murayama

Abstract

Our cut-through IP forwarding technology for a terabit-class super-network can reduce the processing load on provider edge routers. It is designed to improve the scalability of Internet protocol virtual private networks (IP-VPNs) while maintaining terabit-class forwarding performance achieved by cut-through optical path control. It lets provider edge routers accommodate a thousand IP-VPNs. Specifically, it deploys a hub-and-spoke IP routing scheme, redirection control scheme, and purge control scheme. The hub-and-spoke IP routing scheme improves the number of IP-VPNs accommodated in spoke provider edge routers by aggregating IP routes in each VPN. The redirection and purge control schemes maintain terabit-class forwarding performance by controlling IP routes according to traffic demand to avoid traffic concentration at the hub provider edge router. These technologies can efficiently improve scalability because highly reliable control is performed when a cut-through IP route is assigned unless state management is unnecessary.

1. Introduction

To achieve economical Internet protocol virtual private network (IP-VPN) service, it is important to improve network scalability and accommodate a large number of IP-VPNs in a router network having terabit-class forwarding performance. IP-in-IPv6 overlay networking technology [1] and cut-through optical path control technology [2] upgrade the forwarding performance of the backbone network to the terabit class by reducing the processing load on provider routers (P routers). As a complementary approach, this paper describes how we can increase the number of IP-VPNs accommodated in the network by reducing the processing load on provider edge routers (PE routers).

In the conventional router network used to provide Internet service, the network topology is like a tree, and a large number of small PE routers may be accommodated by using a small number of large P routers. In the network, it is easy to increase the number of small PE routers according to traffic demand and expand the network scale gradually. Thus, the network has excellent economical efficiency. In the network, the routing processing and traffic loads of PE routers are light, while those of P routers are heavy. Therefore, P routers are likely to become a performance bottleneck. Moreover, when the network offers IP-VPN service, these loads increase to several times the number of IP-VPNs. This further raises the likelihood of P routers becoming a bottleneck.

To solve this problem, it is usual to deploy an overlay network and establish forwarding tunnels between PE routers when the network accommodates a number of IP-VPNs. One example of such a network is BGP/MPLS IP-VPNs [3]. In this network, the routing processing load of P routers can be reduced because P routers need manage only forwarding tunnels and need not manage the route information in each IP-VPN. In addition, routing is simplified, so optical cross-connects can be deployed as P routers. In this case, it is easy to achieve terabit-class forwarding performance by assigning cut-through opti-

† NTT Information Sharing Platform Laboratories
Musashino-shi, 180-8585 Japan
E-mail: yagi.takeshi@lab.ntt.co.jp

cal paths for forwarding tunnels using a cut-through optical path control scheme.

However, the numbers of routing peers and IP routes managed by PE routers increase in the overlay network [4]. In the basic overlay network, each PE router establishes every other PE router as a routing peer and establishes IP routes by routing protocol. But the number of routing peers is limited to about a hundred [5], so the maximum number of PE routers that can be accommodated in the network is only a hundred. To solve this problem, it has been proposed that BGP [6] should be deployed as a routing protocol and a BGP route reflector [7] deployed as a device to act as a routing peer. The reflector reduces the number of routing peers to one per PE router by changing the routing peer topology from a mesh topology to a hub-and-spoke topology. In addition, the number of routing peers per reflector can be reduced by setting up two or more reflectors in a tree topology, so the number of PE routers in the network can be increased to about a thousand.

However, IP routes are established in a full-mesh topology between all PE routers in each VPN. Therefore, the number of IP routes to be managed in a PE router becomes the sum of the number of IP routes in each VPN. With IP-VPNs spreading to enterprise networks, each IP subnet of an enterprise network depends on the area, branch, and section. If, for example, an enterprise network is composed of ten areas, ten branches, and ten sections, then the number of IP subnets is about a thousand. To accommodate a thousand IP-VPNs, with a thousand IP subnets defined in each VPN, a PE router would need to manage approximately a million IP routes. However, a practical PE router can manage only ten thousand IP routes, so the number of VPNs accommodated in the network is limited to only ten. Furthermore, when a route change occurs among only some of the PE routers, these PE routers distribute route information to all PE routers. So, when there is a locally unstable part of the network, all PE routers must re-calculate route information even if only a few PE routers were concerned with the route change. Thus, the processing load of all PE routers increases and the forwarding performance of the whole network may fall.

The technology in which PE routers request IP routes between PE routers from a server using next-hop resolution protocol (NHRP) [8] according to traffic demand can reduce the number of routing peers and IP routes managed by each PE router to the same number as the number of servers, i.e., one or a few. But in this technology, when a PE router requests IP routes for a large number of destination IP addresses, it performs state management for each destination IP address to avoid duplicated requests and to control retry requests. Furthermore, the number of packets accumulated in a PE router increases because of the delay for resolving IP routes, so the delay and fluctuation in user communication may increase.

Consequently, when the network accommodates a number of IP-VPNs, it is important to reduce the number of routing peers and IP routes managed by each PE router while avoiding state management and packet accumulation in each PE router in the overlay network.

To solve this problem, we propose cut-through IP forwarding technology for a terabit-class super-network (TSN). By deploying IP-in-IPv6 overlay networking technology and cut-through optical path control technology, it could reduce the routing processing load and traffic load of P routers and upgrade the forwarding performance of a service provider network (SP network) to the terabit class. The cut-through IP forwarding technology is composed of a hub-and-spoke IP routing scheme, a redirection control scheme, and a purge control scheme. The hub-and-spoke IP routing scheme reduces the processing load of PE routers while retaining reachability. The redirection control scheme achieves cut-through IP forwarding according to traffic demand using a control procedure that is stateless but reliable. The purge control scheme removes cut-through IP routes when they are no longer justified, which helps stateless but reliable redirection control. In our technology, a hub PE router retains reachability by maintaining all IP routes in the network. So a PE router needs to manage only one IP route towards a hub PE router for each IP-VPN to retain reachability. Thus, such a router can accommodate more than a thousand IP-VPNs. Even when cut-through IP forwarding is performed, the number of IP routes managed in the PE router can be kept to less than ten thousand.

Section 2 describes the cut-through IP forwarding technology that forms the proposed network architecture. Section 3 describes the node architecture as implemented in our prototype. Section 4 evaluates the scalability of our technology. Finally, section 5 provides a conclusion and brief summary.

2. Cut-through IP forwarding technology

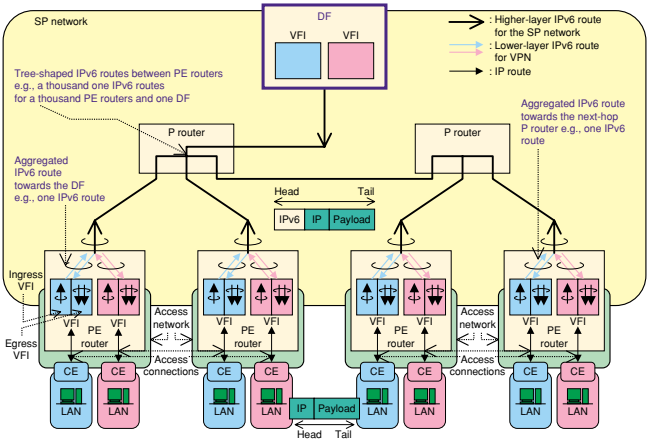
2.1 Hub-and-spoke IP routing scheme

To reduce the processing load of PE routers, it is important to reduce both the number of routing peers

that perform the routing protocol and the number of IP routes established in the IP-in-IPv6 forwarding table. Therefore, we establish the IP routes in a hub-and-spoke topology for each IP-VPN, as shown in **Fig. 1**. Here, the PE router composed of hub VFIs is called a default forwarder (DF).

Any routing protocol such as BGP and OSPF [9] can be deployed in this network and the routing topology is also established as a hub-and-spoke topology, so each spoke VFI only has to establish one routing peer towards the DF's VFI. When the DF's VFIs establish a large number of routing peers, two or more DFs are set up in a tree topology to avoid them becoming a bottleneck for IP routing. In addition, spoke VFIs can aggregate IP routes towards the SP network into static IP routes towards the DF's VFI. Thus, in the spoke VFIs, the number of IP routes to be managed in its IP-in-IPv6 encapsulation table could be reduced to only one. This route towards the DF's VFI is called the default route. A default route is not affected by the routing protocol so this route is not affected by route changes in the IP-VPN. Thus, this technology should not only reduce the number of IP routes but also avoid performance degradation caused by route changes.

On the other hand, the DF must manage all IP routes for all destination IP addresses using IP routing protocol. To avoid a routing loop, in the TSN, a PE router's VFIs are not prohibited from sending an IP packet received from the access network back to the access network but they are prohibited from sending one from the SP network back to the SP network, as shown in **Fig. 2**. When a PE router's VFI receives an IP packet from an access connection or from the SP network, it behaves as an ingress or egress VFI, respectively. Thus, an IP packet received from the SP network is never sent back to the SP network. But only DF's VFIs are permitted to send an IP packet received from the SP network back to the SP network. Instead of this, the DF's VFIs do not accommodate users by means of an access interface. When the DF's VFIs receive IPv6 packets from the SP network, they



SP network: service provider network, PE router: provider edge router, P router: provider router, CE: customer equipment, DF: default forwarder, VPN : virtual private network, VFI : VPN forwarding instance, LSP: label switched path

Fig. 1. Hub-and-spoke IP routing scheme.

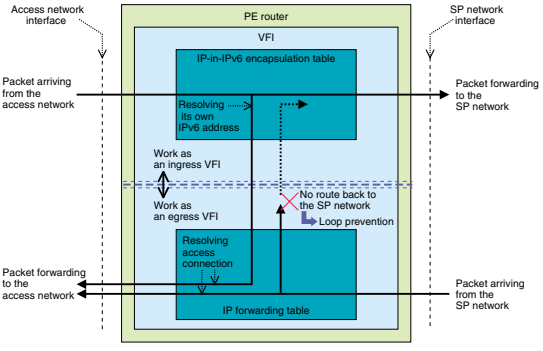


Fig. 2. Packet forwarding scheme in a VFI.

remove the IPv6 header from each user IP packet. Then, the destination IPv6 address is resolved again from the destination IP address using the DF's IP-in-IPv6 encapsulation table and a new IPv6 header is attached to each user IP packet. This newly encapsulated packet is sent back to the SP network.

2.2 Redirection control scheme

Although the hub-and-spoke IP routing is effective in reducing the processing load of the spoke VFIs, the forwarding performance in the network might be degraded because of traffic concentration at the hub. To address this problem, the hub-and-spoke IP routing should be retained to provide reachability, but cut-thorough IP forwarding should be implemented to improve performance. However, if PE routers try to resolve a destination IPv6 address by requesting the server to inform them of IP routes, then the forwarding performance might degrade because of state management. To solve this problem, we apply a redirection control scheme to resolve the destination IPv6 address in a PE router, as shown in **Fig. 3**.

In this scheme, when a VFI in the DF forwards an IP packet from an ingress VFI to an egress VFI, it sends a redirection message to the ingress VFI. This message contains IP-in-IPv6 encapsulation information used for packet forwarding. The ingress VFI is identified from the source IPv6 address of the received IPv6 packet.

When the originating ingress VFI receives the redirection message, it adds the notified information to the IP-in-IPv6 encapsulation table as a cut-through IP route. This route is treated as a cached route that can be removed at any time, while the default IP route is treated as a static route that cannot be removed. The ingress VFI first searches the IP-in-IPv6 encapsulation table for entries other than the default IP route, and, if it finds one, forwards the received packet to the SP network using the cut-through route. Only if no cut-through IP route entry is found does the VFI look up the default IP route and forward the packet to that route in the SP network. In this way, once a cut-through route has been set up, subsequent IP packets to the same destination are forwarded through the cut-through IP route.

This form of control is stateless but reliable. If a redirection message is dropped in the SP network, IP packets are still sent to the DF's VFI, and this results in redirection messages being sent again to the ingress VFI. Even if the ingress VFI were to remove the cut-through IP routes without any reason, IP packets would still be sent to the DF's VFI and redirection messages would still be returned. On the other hand, to suppress the generation of superfluous redirection messages in a short period, the DF's VFIs may optionally perform a state-dependent procedure in which the transit IP packets are classified into source-destination IP address flows and the number

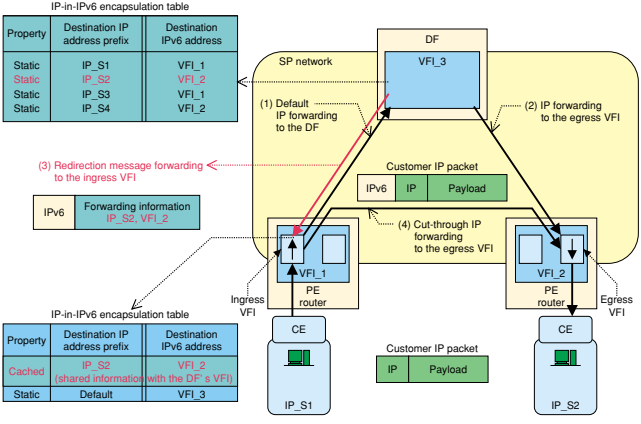


Fig. 3. Redirection control scheme.

of packets per flow is counted periodically. With this procedure, only a single redirection message is sent in a short period even if IP packets arrive at the DF in a bursty manner.

When two or more DFs are set up in a tree topology, a DF may receive redirection messages from other DFs. In this case, the DF does not add the notified information to the IP-in-IPv6 encapsulation table as a cut-through IP route. This mechanism can avoid loop forwarding caused by cut-through IP forwarding when DFs are permitted to send an IP packet received from the SP network back to the SP network. An ingress VFI receives a redirection message from the DF located one hop further from the DF to which the ingress VFI sent an IP-in-IPv6 packet. Consequently, the ingress VFI can receive the redirection message containing IP-in-IPv6 encapsulation information used for forwarding packets to the egress VFI and performs cut-through IP forwarding between ingress-egress VFIs.

In addition, when the traffic between ingress and egress VFIs increases, a cut-through optical path is automatically established between them, which provides a large bandwidth.

2.3 Purge control scheme

The PE router's VFIs can manage notified redirection information freely. Therefore, they can remove that information immediately or hold it over a long period. Considering the DF's traffic load and the state management load of PE routers, it is desirable to hold notified redirection information semi-permanently in PE routers. However, some redirection information becomes obsolete because of updates to the route information when the destination IP subnet moves between more than two PE routers in a network such as a mobile telecommunication network. It is desirable to correct this sort of redirection information so that normal packet forwarding is not disturbed by obsolete redirection information. However, correcting redirection information regardless of the existence of packet forwarding is undesirable because the IP-in-IPv6 encapsulation table management load is increased in the same way that the IP route management load in a mesh topology using routing protocol becomes heavy. To solve this problem, we deploy a purge control scheme that corrects redirection information according to packet forwarding, as shown in **Fig. 4**.

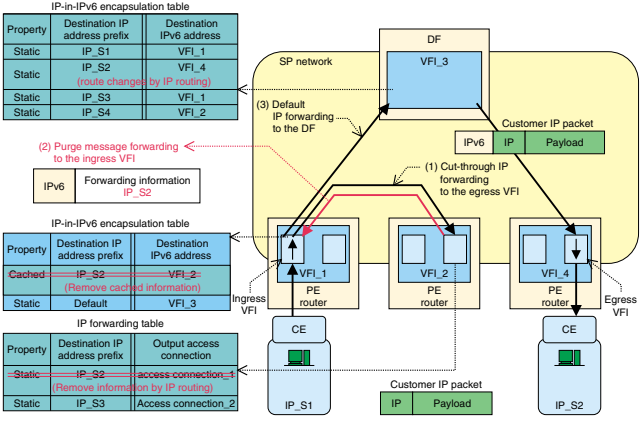


Fig. 4. Purge control scheme.

If IP packets are sent using such outdated routes, they are discarded at the egress VFIs because these VFIs are prevented from forwarding IP packets from the SP network back to the same network. Thus, such obsolete routes should be removed. When an incorrect egress VFI discards an IP packet because there is no route towards the correct access network, it sends a purge message to the ingress VFI. This message contains the destination IP address of the discarded IP packet. The ingress VFI is identified from the source IPv6 address of the attached IPv6 header of the discarded IP packet.

The ingress VFI that receives the purge message removes the cut-through IP route corresponding to the notified IP address from the IP-in-IPv6 encapsulation table. Here, the default IP route is never removed by purge control because it is needed to retain reachability. In this case, subsequent IP packets to the same destination are forwarded through the default IP route.

In the DF's VFI, the IP-in-IPv6 encapsulation table is managed using an IP routing protocol, so IP packets are forwarded towards the appropriate egress VFIs. In addition, the DF's VFI performs redirection control again, so when this has been completed, subsequent IP packets to the same destination are forwarded through the cut-through IP route from the ingress VFI to the appropriate egress VFI. Even if cut-through IP routes are established according to transient IP routing information, forwarding loops are removed and corrected by this purge control scheme.

This control is also stateless but reliable, like the redirection control. If a purge message becomes extinct in the SP network, purge control is applied again after user IP packets arrive at the incorrect egress VFI. In addition, optional state-dependent flow control to suppress duplicated controls is possible. With this procedure, only a single purge message is sent in a short period even if IP packets arrive in a bursty manner at an incorrect egress VFI.

3. Node design

We implemented a DF, as shown in **Fig. 5**. The DF is similar to PE routers deployed in IP-in-IPv6 networking in that they are composed of a core interface package (CIP) and a switch package (SWP). Howev-

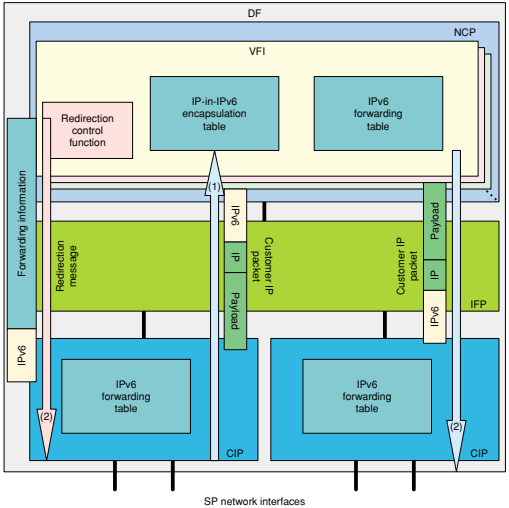


Fig. 5. Node design.

er, the DF does not have an access interface package (AIP) because it is exclusively a relaying device on the user IP layer, so the DF performs the forwarding process on the user IP layer in a network control package (NCP). Therefore, a VFI is allocated in an NCP. An IP-in-IPv6 packet received by a CIP from the SP network is forwarded to a VFI in an NCP via SWP. The VFI removes the IPv6 header and applies the forwarding process to a decapsulated user IP packet on the user IP layer. The VFI attaches a new IPv6 header to the user IP packet and forwards the newly encapsulated user IP packet to a CIP. The CIP sends this IP-in-IPv6 packet back to the SP network. This processing sequence results in a hub-and-spoke IP routing scheme. In addition, the NCP generates redirection messages at the time of packet forwarding and sends them to ingress PE routers. This process achieves a redirection control scheme. Purge control is performed between PE routers, so a DF need not be concerned with purge control.

4. Scalability evaluation

To compare our scheme with the conventional technology, we calculated the numbers of IP-VPNs that can be accommodated by PE routers to determine how many IP-VPNs can be accommodated by the SP network. With the conventional technology, the number of IP routes to be managed in the IP-in-IPv6 forwarding table for each PE router (Pc) is expressed as the product of the number of IP-VPNs accommodated in each PE router (Nv) and the number of IP subnets defined in each IP-VPN (Ns).

$$Pc = Nv\ Ns \quad (1)$$

On the other hand, in a TSN, the maximum number of IP routes to be managed in the IP-in-IPv6 forwarding table for each PE router (Pp) is expressed as

$$Pp = (1 + Nc)\,Nv \quad \text{for } 0 < Nc < Ns, \tag{2}$$

where Nc is the number of cut-through IP routes for each VPN.

Figure 6 shows how the numbers of IP routes vary with the number of VPNs (Nv) for two cases, BGP/MPLS IP-VPNs and the TSN, when the number of IP subnets in each VPN was fixed at a thousand.

Practical PE routers can manage about ten thousand IP routes towards the SP network in the IP forwarding table. Thus, in BGP/MPLS IP-VPNs, only ten IP-VPNs can be supported in each PE router if each VPN is composed of a thousand IP subnets. On the other hand, in the TSN, a maximum of ten thousand IP-VPNs can be accommodated, although cut-through IP routes cannot be established in this condition. For example, let us consider a network composed of ten star topologies with all PE routers belonging to all of the star topologies. The central PE routers of the stars are located in major cities, company head offices, or data centers etc., and traffic is concentrated there. In this case, even when a PE router other than the central one establishes ten cut-though IP routes for each IP-VPN, the PE router can still accommodate thousands of IP-VPNs.

When a thousand PE routers are accommodated in the SP network and each IP-VPN is composed of ten VFIs, the number of IP-VPNs accommodated in the SP network is a hundred times that accommodated in each PE router. In this case, in BGP/MPLS IP-VPNs, only a thousand IP-VPNs can be accommodated in the SP network because a PE router can accommodate only ten IP-VPNs. On the other hand, in a TSN, a hundred thousand IP-VPNs can be accommodated in the SP network because a PE router can accommodate a thousand IP-VPNs. Thus, a TSN can support one hundred times as many IP-VPNs as a corresponding BGP/MPLS IP-VPN for this evaluation condition.

In this way, while maintaining the forwarding performance, a TSN can reduce the number of forwarding table entries of PE routers and can accommodate a lot of IP-VPNs compared with BGP/MPLS IP-VPNs.

5. Conclusion

Our cut-through IP forwarding scheme is designed

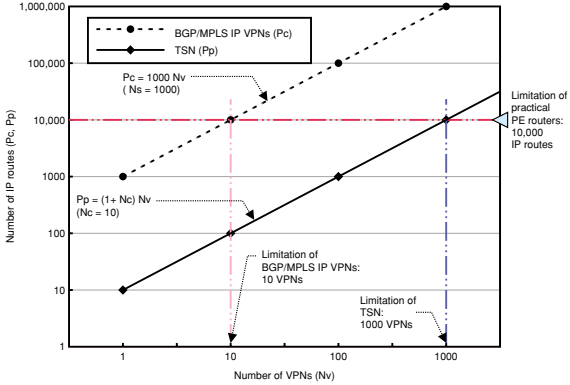


Fig. 6. Scalability evaluation.

for a terabit-class super-network used as a provider network to accommodate a large number of IP-VPNs in which a large number of IP subnets are defined. In the conventional architecture, only ten IP-VPNs with a thousand IP subnets defined in each VPN can be accommodated by full-mesh IP routes. In our architecture, on the other hand, a thousand IP-VPNs can be accommodated under the same conditions because reachability is retained by the hub-and-spoke IP routing scheme. In addition, performance degradation is suppressed by the redirection control scheme. Consistency between IP routing information and IP forwarding information is ensured by the purge control scheme. Since a terabit-class super-network enables a service provider network to accommodate a lot of small-scale provider edge routers, a larger number of IP-VPNs can be accommodated in it.

6. Acknowledgment

This research was supported by a grant from the Telecommunications Advancement Organization of Japan (TAO).

References

[1] Y. Naruse, T. Yagi, K. Matsui, and J. Murayama, "IP-in-IPv6 Overlay Networking Technology for a Terabit-class Super-network," NTT Technical Review, Vol. 2, No. 3, pp. 21-31, 2004.
[2] K. Matsui, T. Yagi, Y. Naruse, and J. Murayama, "Cut-through Optical Path Control Technology for a Terabit-class Super-network," NTT Technical Review, Vol. 2, No. 3, pp. 32-40, 2004.
[3] J. Murayama, K. Matsui, K. Matsuda, and M. Makino, "Conceptual Design for a Terabit-class Super-networking Architecture," NTT Technical Review, Vol. 2, No. 3, pp. 12-20, 2004.
[4] E. Rosen and Y. Rekhter, "BGP/MPLS IP VPNs," IETF Internet-draft<draft-ietf-l3vpn-rfc2547bis-01.txt>, May 2003.
[5] http://www.cisco.com/japanese/warp/public/3/jp/service/tac/459/highcpu-bgp-j.html#subtopic4b
[6] Y. Rekhter and T. Li, "A Border Gateway Protocol 4," RFC1771, Mar. 1995.
[7] T. Bates, R. Chandra, and E. Chen, "BGP Route Reflection – An Alternative to Full Mesh IBGP," RFC2796, Apr. 2000.
[8] B. Fox and B. Petri, "NHRP Support for Virtual Private Networks," RFC2735, Dec. 1999.
[9] J. Moy, "OSPF Version 2," RFC2328, Apr. 1998.

Takeshi Yagi
Secure Communication Project, NTT Information Sharing Platform Laboratories.
He received the B.E. degree in electrical and electronic engineering and the M.S. degree in science and technology from Chiba University, Chiba in 2000 and 2002, respectively. He joined NTT in 2002. His studies focus on IP-VPN architecture and his current research interests include IP routing and forwarding technology, traffic monitoring technology, and layer cooperation technology. He is a member of the Institute of Electrical Engineers of Japan (IEEJ) and the Institute of Electronics, Information and Communication Engineers of Japan (IEICE).

Kenichi Matsui
Secure Communication Project, NTT Information Sharing Platform Laboratories.
He received the B.E. degree in information engineering and the M.S. degree in information sciences from Tohoku University, Sendai, Miyagi in 1995 and 1997, respectively. He joined NTT in 1997. His work focuses on IP networking and his research interests include traffic engineering for optical IP networks and MPLS, on-demand QoS management, and managed IP multicast platforms. He is a member of IEICE, the Information Processing Society of Japan, and the IEEE Computer Society.

Yuuichi Naruse
Research Engineer, Secure Communication Project, NTT Information Sharing Platform Laboratories.
He received the B.S. and M.S. degrees in physics from Tohoku University, Sendai, Miyagi in 1990 and 1992, respectively. He joined NTT in 1992. He has been engaged in R&D of ATM-LAN systems, Fibre-Channel network systems, and IP-VPN service platforms.

Junichi Murayama
Senior Research Engineer, Secure Communication Project, NTT Information Sharing Platform Laboratories.
He received the B.E. and M. E. degrees in electronics and communication engineering from Waseda University, Tokyo in 1989 and 1991, respectively. Since joining NTT in 1991, he has been engaged in R&D of ATM networks, large-scale IP networks, and IP-VPN service platforms. He is a member of IEICE.