Conceptual Design for a Terabit-class Super-networking Architecture

Junichi Murayama[†]*, Kenichi Matsui, Kazuhiro Matsuda, and Masaya Makino*

Abstract

Our terabit-class super-networking architecture is designed to improve both the performance and scalability of a network used by a service provider to offer Internet protocol virtual private network services. Specifically, we aim to achieve terabit-class performance among a thousand provider edge routers. To do this, we deploy IP-in-IPv6 overlay networking, cut-through optical path control, and cut-through IP forwarding, working together in cooperation. High scalability is achieved by retaining reachability in a connectionless manner and high performance is achieved by assigning cut-through forwarding resources according to traffic demands. As a result, this architecture achieves our objectives economically.

1. Introduction

An Internet protocol virtual private network (IP-VPN) service provides a closed user group with a dedicated virtual wide-area network. It achieves a low price with high security and flexibility. Low price is achieved by laying multiple IP-VPNs over a single physical provider network, thus allowing physical forwarding resources to be shared between IP-VPNs. High security is achieved by separating IP routes on IP-VPNs so that IP packets cannot be forwarded from one IP-VPN to another. This route separation also provides high flexibility: customers can design IP addressing for their IP-VPN without being constrained by potential conflicts among IP-VPNs.

Low price and high security are playing an important role in the spread of IP-VPN services to enterprise networks. In the near future, these features will also help to extend such services to home networks because remote access to a home network requires security. On the other hand, flexibility is an important feature in spreading these services to Internet service providers (ISPs).

This background requires a provider's network for both IP-VPN and Internet service to achieve terabit-class performance among a thousand provider edge routers. To simply improve performance, generalized multiprotocol label switching (GMPLS) [1] using optical paths is effective. However, if only optical paths are used to retain reachability among provider edge routers, full-mesh optical paths are required. This means that the number of optical paths increases very rapidly as the number of provider edge routers increases. Consequently, scalability is not achieved, where scalability means the ability of a provider's network to accommodate a large number of provider edge routers when the number of customers increases. A gradual increase in scale is especially important for commercial service providers to reduce service costs.

To achieve both good performance and scalability, we propose a terabit-class super-networking architecture. In this architecture, scalability is improved by using IP-in-IPv6 overlay networking where reachability is retained in a connectionless manner (IPv6: Internet protocol version 6). Performance is improved through the use of both cut-through optical path control and cut-through IP forwarding. The cut-through optical paths and IP routes are assigned according to optical and IP traffic demands, respectively. In addition, the different cut-through procedures work in cooperation. Consequently, our architecture improves both scalability and performance

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

economically.

Section 2 describes the design objectives for our architecture and section 3 shows the reference model, which forms the basis for the design. Section 4 explains the design issues and section 5 describes the architecture design. Section 6 gives a summary and conclusions.

2. Design objectives

A terabit-class super-network is designed as a single provider network for an IP-VPN service. The network architecture is designed so that small-scale routers can be added as the number of customers increases, so a large-scale provider network can be composed of a large number of small-scale routers. This is because gradual scale migration is an important issue for commercial service providers in reducing service costs. However, neither the conventional IP-VPN architecture nor the GMPLS-based Internet architecture can achieve this. An overview of the design objectives is given in **Fig. 1**. They may be summarized as follows:

(1) Target services:
- IP-VPN service
- extranet service
- multi-VPN access service
- multi-grade-security VPN service

(2) Target customers:
- enterprise network customers
- home network customers
- Internet service providers

(3) Maximum network performance:
- 100 Tbit/s (total)
- 10 Gbit/s (end-to-end)

(4) Maximum network scale:
- 10 million customer sites
- 1000 provider edge routers
- 100,000 VPNs

3. Reference model

An IP-VPN architecture should conform to the provider-provisioned virtual private network (PPVPN) framework [2] shown in **Fig. 2**. In this framework, customer edge (CE) devices accommodating local area networks (LANs) are connected to provider edge (PE) routers via access connections. PE routers are connected to provider (P) routers. Each PE router incorporates multiple forwarding instances

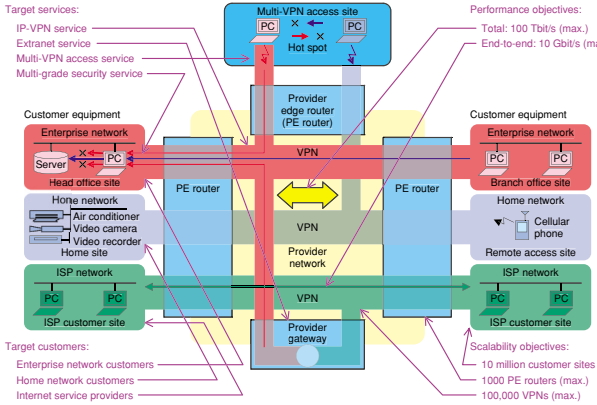
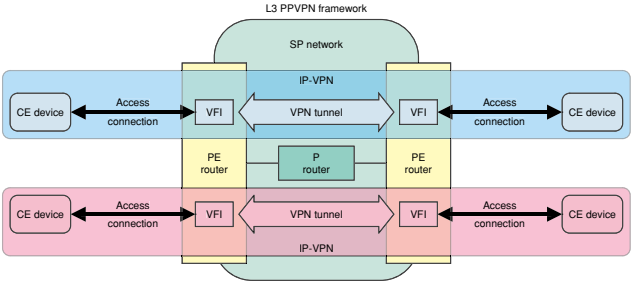


Fig. 1. Design objectives.



L3: layer 3, PPVPN: provider provisioned virtual private network, IP: Internet protocol, VPN: virtual private network, VFI: VPN forwarding instance, P router: provider router, PE router: provider edge router, CE device: customer edge device, SP network: service provider network.

Fig. 2. Reference model.

called VPN forwarding instances (VFIs). Each VFI is dedicated to a single IP-VPN. VFIs belonging to the same IP-VPN are connected to each other via virtual tunnels called VPN tunnels. P routers only relay VPN tunnels and do not perform IP-VPN routing. A PPVPN framework conforms to the overlay model. Here, the underlying network is a physical network called a service provider (SP) network, which is composed of P and PE routers connected via physical links. In contrast, the overlaying networks are IP-VPNs composed of VFIs connected via VPN tunnels.

4. Design issues

4.1 Optical layer issue

To improve the performance of an SP network in a PPVPN framework in a simple manner, optical transmission technologies are promising. Optical cross-connects (OXCs) are deployed as P routers and are connected to each other via dense wavelength division multiplexing (DWDM) links. In addition, wavelength-dedicated optical paths are established between PE routers, with the VPN tunnels being multiplexed into the optical paths.

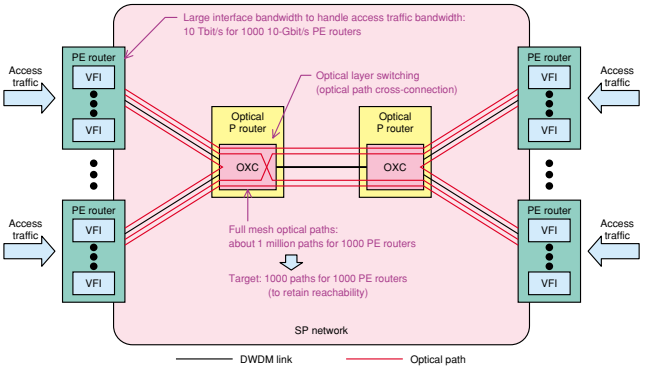
However, if reachability is retained only by optical paths, the optical paths must be arranged in a mesh topology between the PE routers, as shown in **Fig. 3**. Thus, the number of optical paths increases very rapidly as the number of PE routers increases. Approximately one million optical paths would be required to interconnect a thousand PE routers. Since a practical OXC network can handle only ten thousand optical paths, the number of PE routers that may be managed in such a network is limited to a hundred. Even if the number of PE routers is small, it is not cost-effective to assign an optical path between PE routers where the demand for traffic is small. When the number of such paths increases, the effective performance of the whole network decreases.

Consequently, reachability should be retained by electrical forwarding technologies, and optical paths between PE routers should be assigned to those routes where traffic demand is high.

4.2 Sub-IP layer issue

In a typical PPVPN implementation, electrical MPLS technologies [3] are applied to the SP network and reachability between PE routers is retained by electrical paths. In this approach, as shown in **Fig. 4**, electrical paths are arranged in a mesh topology and some electrical paths can be replaced by cut-through*1 optical paths where traffic demand is high.

*1 Cut-through is a function that replaces the IP layer packet forwarding procedure with a sub-IP layer data transmission procedure. Although a routing problem still needs to be solved because the IP address is not looked up in the replacement procedure, this function is promising for improving performance.



DWDM link: dense wavelength division multiplexing link, OXC: optical cross-connect

Fig. 3. Optical layer issue.

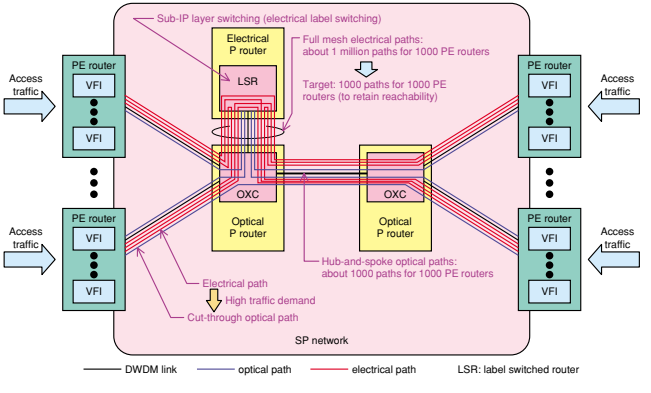


Fig. 4. Sub-IP layer issue.

However, the number of electrical paths also increases very rapidly with an increase in the number of PE routers. At a rough estimation, label switch routers, used as P routers, would need to be able to handle a million electrical paths to interconnect a thousand PE routers, but practical routers can only handle ten thousands paths.

When electrical P routers are used to merge point-to-point paths into a multipoint-to-point path, the number of electrical paths can be reduced to the same level as the number of PE routers. However, an electrical multipoint-to-point path cannot be replaced by optical point-to-point paths because the path types do not match. This means that the effective performance of the whole network is not improved.

Consequently, an electrical label should be designed such that electrical paths can be merged in P routers and an electrical multipoint-to-point path can be partially replaced by optical point-to-point paths.

4.3 IP layer issue

BGP/MPLS IP-VPNs [4] shown in **Fig. 5** are a familiar PPVPN implementation (BGP: border gateway protocol). In this implementation, VPN tunnels are established in a mesh topology between PE routers belonging to the same IP-VPN. Every VPN tunnel is mapped to IP routes at an ingress PE router. Although PE routers in the conventional hop-by-hop routing network can define a default forwarding IP route, these MPLS PE routers cannot define such a route. Thus, the number of IP routes managed in a PE router becomes the sum of the number of IP subnets in each VPN. At a rough estimation, a PE router would need to be able to manage a million IP routes to accommodate a thousand IP-VPNs, each having a thousand IP subnets, but a practical PE router can manage only ten thousand IP routes. When IP routes are established in a hub-and-spoke topology between PE routers, spoke PE routers can establish a default IP route to reduce the number of managed IP routes. However, a hub PE router may act as a performance bottleneck due to traffic concentration.

Consequently, it is important to retain reachability of the whole IP destination by using hub-and-spoke IP routes and to distribute traffic by using cut-through IP routes.

5. Architecture design

5.1 Design concept

An overview of our terabit-class super-networking architecture is illustrated in **Fig. 6**. In this architecture, the SP network is layered into an optical GMPLS network and an electrical IPv6 network. IP-

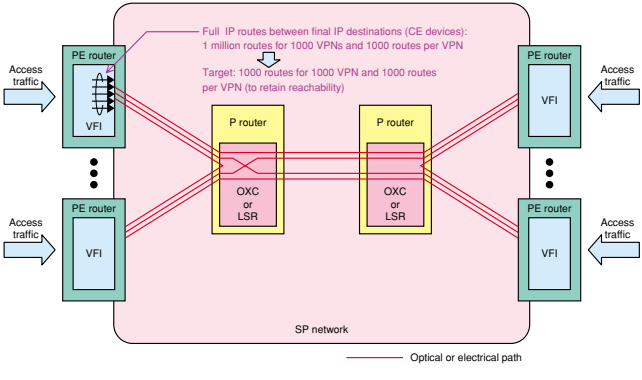


Fig. 5. IP layer issue.

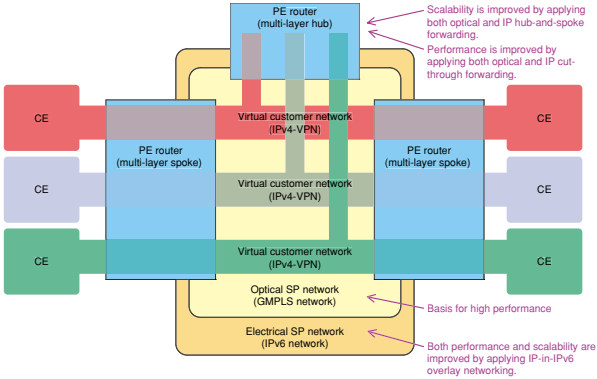


Fig. 6. Overview of the proposed architecture.

VPNs are laid over the SP network as customer IP networks. Thus, two kinds of P router are deployed: an OXC acting as an optical P router and an IPv6 router acting as an electrical P router. The PE routers deploy multi-layer functions and select the appropriate forwarding layer from the choice of optical layer, IPv6 layer, and customer IP layer.

For greater scalability, higher-layer forwarding is preferable, while lower-layer forwarding is preferable for better performance. Thus, reachability should be retained using higher-layer forwarding and a large amount of traffic should be carried using lower-layer forwarding. In our architecture, IP layer forwarding by IP-in-IPv6 forwarding is selected first. Then, IPv6 layer forwarding by cut-through IP forwarding is selected even when the demand for traffic is small. Finally, optical layer forwarding by cut-through optical path control is selected, but only when traffic demand is high. Consequently, these three technologies work in cooperation. They are overviewed in the following subsections.

5.2 IP-in-IPv6 overlay networking

As the first part of the solution, we apply IP-in-IPv6 overlay networking to the SP network of a PPVPN framework. In this technique, a customer IP packet is encapsulated into a provider IPv6 packet at the ingress PE router. The IPv6 packet is forwarded from the ingress PE router to the egress PE router via IPv6 routers acting as electrical P routers. The IP packet is then decapsulated from the IPv6 packet at the egress PE router. Since IPv6 routers forward IPv6 packets in a connectionless manner, the number of IPv6 routes to be managed in each IPv6 router can be of the same order as the number of PE routers shown in **Fig. 7**. A practical IPv6 router can manage those IP routes easily even when the number of PE routers is a thousand.

Then, to avoid wasting optical paths in each optical P router, they are arranged in a tree topology with electrical P routers forming the central nodes. Since reachability is retained in a connectionless manner by electrical P routers, the number of optical paths to be handled in each optical P router can be of the same order as the number of PE routers.

Furthermore, to avoid wasting IP routes in each PE router, they are arranged in a hub-and-spoke topology with one of the PE routers being the hub. The hub is called the default forwarder (DF). In each PE router, IP routes towards other PE routers are first aggregated into a single IP route towards the DF.

Consequently, IP-in-IPv6 overlay networking is the key to solving the sub-IP layer issue and some of the optical and IP layer issues. This is discussed in more detail in the next paper in this issue [6].

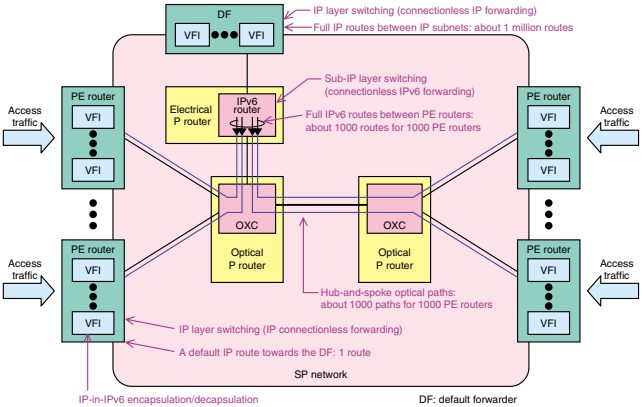


Fig. 7. IP-in-IPv6 overlay networking technology.

5.3 Cut-through optical path control

As a second part of the solution, we apply cut-through optical path control to the IPv6-based SP network. An IPv6-based P router may be a performance bottleneck in the sub-IP layer, depending on the traffic concentration. To solve this problem, we assign a cut-through optical path between PE routers where the traffic demand is high. The key to optimizing the optical path arrangement is to use a network control server (NCS) and continuously calculate the optimal arrangement. As shown in **Fig. 8**, the IPv6-based P router classifies transit packets according to a pair of source-destination IPv6 addresses and counts the number of occurrences of each pair. The NCS collects these numbers from the IPv6-based P routers and calculates the optical path configuration so that the total load on the IPv6-based P routers is minimized within the limited number of optical path interfaces of each PE router. Then the NCS triggers optical path signaling and also establishes an IPv6 route along the path. Optical paths are established between PE routers by optical GMPLS technologies. The whole network performance can be improved by determining the optimal path arrangement according to traffic demands. As a result, the number of optical paths required for the SP network can be less than ten thousand when the number of PE routers is a thousand and each PE router deploys ten optical interfaces to accommodate access traffic. This requirement can be met by a practical OXC network. However, the whole network performance depends on the path arrangement algorithm. This is discussed in more detail in the third paper in this issue [7].

5.4 Cut-through IP forwarding

As the third part of the solution, we apply cut-through IP forwarding to PE routers in the SP network. The hub DF may be a performance bottleneck in the IP layer when a large amount of traffic is concentrated there. To solve this problem, we assign a cut-through IP route between PE routers according to traffic demand, as shown in **Fig. 9**. When the DF forwards an IP-in-IPv6 packet, it also extracts IP forwarding information composed of the IP address of the destination IP subnet and the IPv6 address of the egress PE routers. The DF sends this information to the ingress PE routers identified by the source IPv6 address as a redirection message. The ingress PE router caches it as a cut-through IP route towards the egress PE router. As a result, subsequent IP packets to

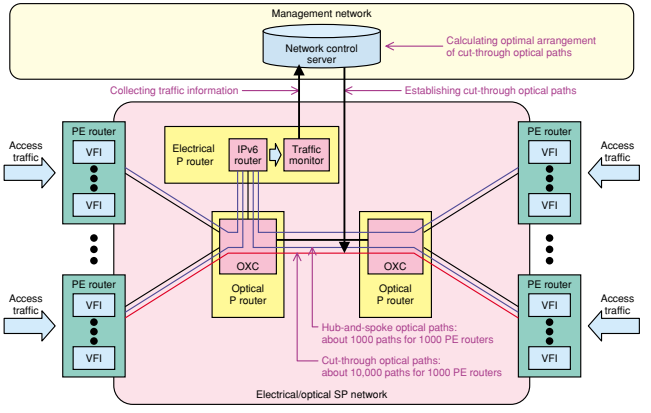


Fig. 8. Cut-through optical path control technology.

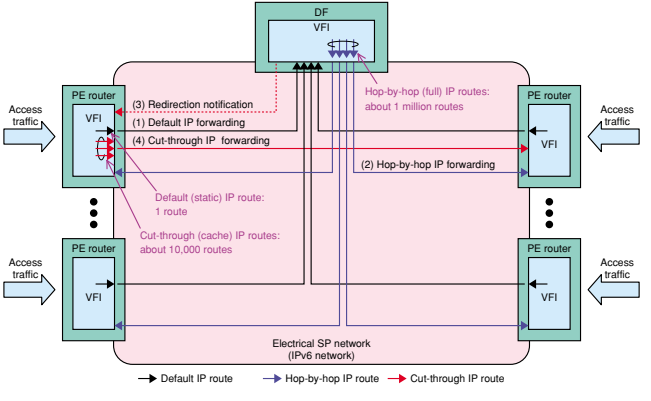


Fig. 9. Cut-through IP forwarding technology.

the same destination IP subnet are forwarded via the cut-through IP route. These routes can be removed whenever necessary because they are created automatically according to traffic demand. Thus, the number of IP routes managed in each PE router can be less than ten thousand, which can be handled by a practical PE router. After cut-through IP routes have been assigned, cut-through optical paths are also assigned when demand for cut-through IP traffic is high. This is discussed in more detail in the fourth paper in this issue [8].

6. Conclusion

This paper introduced a terabit-class super-networking architecture to provide an IP-VPN service. This architecture is designed to improve both scalability and performance. While the conventional architecture can handle only one hundred PE routers because reachability is retained by full-mesh optical paths, our architecture can handle a thousand PE routers because reachability is retained by hub-and-spoke optical paths. Moreover, our architecture enables cut-through optical path control and cut-through IP forwarding to work in cooperation by using IP-in-IPv6 overlay networking. This improves the total network performance because cut-through forwarding resources are assigned according to traffic demands. These features enable a provider network to accommodate many small-scale routers. Therefore, our architecture is cost-effective for commercial service providers for whom a small-scale network should grow gradually in proportion to the increase in the number of customers.

7. Acknowledgment

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

References

[1] E. Mannie, “Generalized multi-protocol label switching architecture,” IETF Internet-draft, draft-ietf-ccamp-gmpls-architecture-07.txt, May 2003.
[2] R. Callon and M. Suzuki, “A framework for layer 3 provider provisioned virtual private networks,” IETF Internet-draft <draft-ietf-l3vpn-framework-00.txt>, Mar. 2003.
[3] E. Rosen, A. Viswanathan, and R. Callon, “Multiprotocol label switching architecture,” IETF RFC3031, Jan. 2001.
[4] E. Rosen and Y. Rekhter, “BGP/MPLS IP VPNs,” IETF Internet-draft <draft-ietf-l3vpn-rfc2547bis-01.txt>, May 2003.
[5] A. Conta and S. Deering, “Generic packet tunneling in IPv6 specification,” IETF RFC 2473, Dec. 1998.
[6] 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.
[7] 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.
[8] T. Yagi, K. Matsui, Y. Naruse, and J. Murayama, “Cut-through IP Forwarding Technology for a Terabit-class Super-network,” NTT Technical Review, Vol. 2, No. 3, pp. 41-49, 2004.

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 the Institute of Electronics, Information and Communication Engineers (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.

Kazuhiro Matsuda
Senior Research Engineer, Supervisor, Secure Communication Project, NTT Information Sharing Platform Laboratories.
He received the B.E. and M.E. degrees in electronic engineering from Hokkaido University, Hokkaido in 1983 and 1985, respectively. Since joining NTT in 1985, he has worked on LSI CAD systems, the design of high-speed protocol processing LSIs, and managed L2/L3 VPNs. He is a member of IEICE and the IEEE Computer Society.

Masaya Makino
Senior Research Engineer, Supervisor, Secure Communication project, NTT Information Sharing Platform Laboratories.
He received the B.E. and M.E. degrees in applied mathematics and physics from Kyoto University, Kyoto in 1983 and 1985, respectively. In 1985, he joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Public Corporation (now NTT). He has been engaged in R&D in the business networking area. He is a member of IEICE.