

Invited Paper

A Design of Next Generation IX using MPLS Technology

IKUO NAKAGAWA,^{†1} HIROSHI ESAKI,^{†2} YUTAKA KIKUCHI^{†3}
and KENICHI NAGAMI^{†4}

An IX (Internet eXchange) is a mechanism to interconnect many networks to each other. Currently, an ISP (Internet Service Provider) establishes numerous interconnections to other ISPs. Although ‘private peering’ is one way for an ISP to interconnect to other ISPs with individual links, connecting to an IX is a more efficient way to establish and maintain a large number of peerings (or ‘public peerings’) with other participating ISPs.

Currently, two major IX architectures exist. One uses LAN (Local Area Network) technologies such as FDDI, Ethernet or Gigabit Ethernet to interconnect ISPs to each other. The other IX architecture is based on ATM (Asynchronous Transfer Mode) technology, which uses PVCs (Permanent Virtual Circuit) between participating ISPs. Both LAN and ATM based IXes have several problems, for example, bandwidth limitation, operational cost, less scalability, and dependency on data-link mediums.

In this paper, we propose a next generation IX architecture based on MPLS (Multi-Protocol Label Switching) technology. MPLS provides a data-link independent virtual path, called LSR (Label Switched Path), between MPLS capable routers. MPLS technology is also useful with a traffic engineering capability. We apply this MPLS technology to an IX. A MPLS based IX has the advantages of the independency of data-link mediums, unlimited bandwidth, scalability, and widely distributed features.

1. Introduction

An IX (Internet eXchange) is a mechanism to interconnect many networks to each other. Currently, an ISP (Internet Service Provider) establishes numerous interconnections to other ISPs. Although ‘private peering’ is one way for an ISP to interconnect to other ISPs with individual links, connecting to an IX is a more efficient way to establish and maintain a large number of peerings (or ‘public peerings’) with other participating ISPs.

Recently, a large number of IXes operate³⁾ to exchange large volumes of traffic between participating ISPs. For example, PAIX (Palo Alto Internet eXchange)⁴⁾ is one of the largest IXes in the world. The MAE (Metropolitan Area Network)⁵⁾ also provides several IX points in the United States, to exchange traffic between ISPs. Similarly, LINX⁶⁾, NYIIX⁷⁾, AMX-IX⁸⁾, NSPIXP⁹⁾, and many other IXes exchange Internet traffic between participating ISPs.

In this paper, we propose a next generation IX architecture using MPLS (Multi-Protocol Label Switching)¹²⁾ technology. MPLS enables abstraction of network devices. MPLS provides

virtual path between network nodes and inherit physical and data-link layer dependency. That is, MPLS networks can consist of any data-link medium, for example, POS (Packet Over Sonet), ATM (Asynchronous Transfer Mode), or GbE (Gigabit Ethernet). As a result, an IX based on MPLS technology, called **MPLS-IX**, takes advantage of migration of data-link mediums. A **MPLS-IX** also has the advantage of scalability or simple backbone operation.

In section 2 we introduce the basic concept of an IX, and an IX policy model called a ‘bilateral’ model. We describe current IX architectures such as LAN technology based IX or an ATM technology based IX. We also discuss problems existing IXes face.

In section 3, we discuss about abstraction of network devices. MPLS provides virtual network mechanism which inherit any physical and data-link medium of network devices. A design of new IX architecture proposed in this paper stands on the abstraction of network devices.

In section 4, we propose a next generation IX architecture using the MPLS (Multi-Protocol Label Switching) technology. We describe how to apply the MPLS technology to an IX. We also discuss about key features of **MPLS-IX**, such as independency of data-link mediums, unlimitation of transmit speed, widely distributable feature, and scalability.

†1 INTEC Web and Genome Informatics Corporation

†2 University of Tokyo

†3 Kochi University of Technology

†4 Toshiba

In section 5, we report the results of experimental test of our proposed IX architecture with MPLS capable routers. We confirm normal behavior of traffic exchange in **MPLS-IX**. We also ensure that our proposed architecture provides redundancy inside the IX as well as path recalculation in participating ISPs. We also evaluate performance and scalability of **MPLS-IX**.

2. IX - Internet eXchange

First, we describe the basic IX mechanism and current IX technologies. To understand the IX mechanisms, we refer to ‘private peering’ mechanism, first. We also mention an IX policy model, called a ‘bilateral’ model, which is an important factor for IX implementations.

In section 2.3 and section 2.4, we review current IX technologies: a LAN technology based IX, and an ATM technology based IX. We also discuss about problems that current IX technologies face.

2.1 IX model

In the Internet, two main ways to achieve interconnection between ISPs exist. Private peering is a method to establish an interconnection between two ISPs. In other words, two ISPs prepare and operate a dedicated physical point-to-point circuit between each other, and exchange traffic over the circuits. When an ISP wishes to interconnect to multiple ISPs, the ISP has to draw multiple physical circuits for each ISP to individually exchange data traffic.

Fig. 1 represents a typical case of interconnection between multiple networks with the private peering model. As shown in this figure, an ISP has to prepare and operate individual physical circuits for each ISP. To complete fully meshed interconnections, the number of individual interconnection circuits is in total $N(N-1)/2$, where N is the number of ISPs that want to interconnect. As a result, a model of private peering with N ISPs needs $O(N^2)$ interconnection circuits. Obviously the model of private peering does not provide clear scaling properties.

On the other hand, IX(Internet eXchange) reduces the total cost of dedicated lines between ISPs. An IX is a specific ‘field’ where N ISPs can make interconnections to each other. An ISP that wants to interconnect to others draws a single physical circuit into the IX. Fig. 2 illustrates the basic model of an IX.

In this model, preparing a specific ‘field’

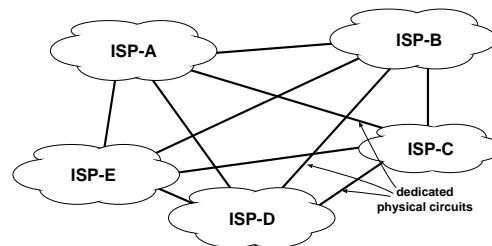


Fig. 1 Private peering

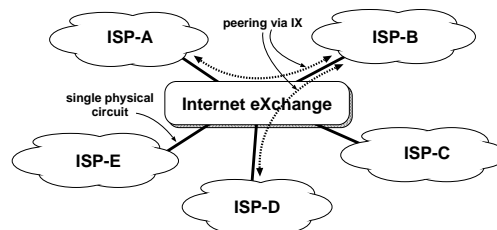


Fig. 2 Internet eXchange

where ISPs can exchange data traffic achieves the same functionality of complete private interconnections between these N ISPs. Also shown in Fig. 2, the total number of physical circuits is only N , e.g., $O(N)$. A participating ISP needs no additional individual circuits, which is why we consider the IX model an efficient way to achieve numerous interconnections between ISPs.

2.2 IX policy model

In an environment of interconnections, the total volume of traffic between two ISPs is decided by routing information exchanged by each of the ISP routers. For an ISP, incoming traffic depends on the outgoing routing information, and outgoing traffic is the outcome of accepted routing information. In this way, routing policy is important for all the ISPs in controlling their incoming or outgoing traffic. This situation is also true in the IX environment. As a result, IXes are now active policy elements in the Internet. Likewise, IX policy model is an important factor in implementing IX technologies.

In current IX environments, participating ISPs have a higher expectation of flexibility in policy control from an exchange structure. These ISPs themselves determine the routing policy in controlling both incoming and outgoing traffic; that is, each ISP wants to control incoming and outgoing routing information individually exchanged with other ISPs. Partici-

participating ISPs disregard a situation where IX operators decide or affect ISP routing policy.

To make participating ISPs individually control routing information, a policy model of the IX is based on the ‘bilateral’ model; any two participating ISPs can themselves decide their routing policy without the control of IX operators. In this model, an IX provides only a basic functionality which allows any two ISPs to interconnect to each other. The IX operators do not care about routing information exchanged between participating ISPs.

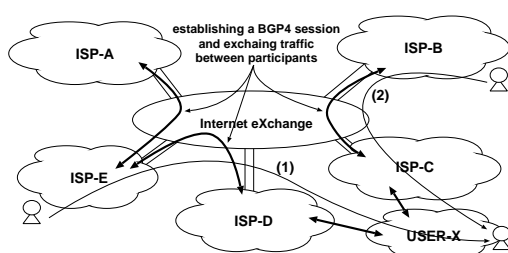


Fig. 3 Policy Model

Fig. 3 is an example of the ‘bilateral’ policy model in an IX. In this figure, three interconnections exist in the IX. In one interconnection, for example, ISP-B and ISP-C interconnect to each other and exchange routing information between their routers. Note that USER-X buys transit connectivity from both ISP-C and ISP-D, and these ISPs announce the route for USER-X via the IX. From the IX’s point of view, there are two different routing entries for the specific user USER-X on the IX. If the IX is a single router or a set of routers, routing policy is decided by the IX itself because the forwarding table for a routing prefix normally has only one next-hop entry in a router. Instead, as shown in this figure, the bilateral policy model allows participating ISPs to decide the forwarding path themselves, such that a user of ISP-E transmits datagrams through ISP-D, and a user of ISP-B chooses paths through ISP-C.

2.3 LAN based IXes

One of the most well known implementations of the IX model is the use of LAN (Local Area Network) technologies, such as FDDI or the Ethernet. An implementation of the LAN based IX is simple because an IX provider only needs to prepare a LAN switch and participating ISPs connect their routers into the switch. Hereafter, we refer to these kinds of IXes as ‘LAN-IX’. Currently, PAIX, LINX,

NYIIX, NSPIX2 and many other major IXes are based on the LAN-IX model.

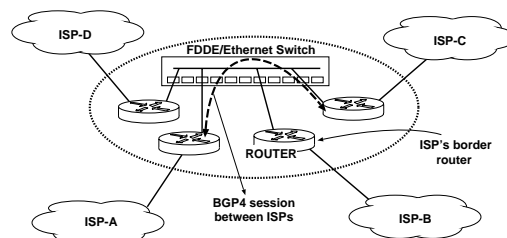


Fig. 4 IX based on LAN technology

Fig. 4 illustrates the basic architecture of the LAN-IX. In the LAN-IX, the IX itself consists of a set of LAN switches, for example, FDDI switches or Ethernet Switches. In general, when a participating ISP wants to connect its router into the IX, the ISP has to prepare its border router to be located near the LAN switches, because there is a fiber or cable length restriction in most LAN mediums. The LAN-IX is sometimes referred to as the, ‘concentrated model’.

Another important characteristic in the LAN-IX architecture is that a LAN-IX uses a shared subnet for exchanging actual traffic between participating ISPs. As shown in Fig. 4, LAN switches provide a shared subnet, called an ‘exchange subnet’. For the participating ISP routers, an IX operator assigns an IP address in the exchange subnet, and the ISP connects its router into the exchange subnet with the assigned IP address. Since the functionality of the IX only provides LAN communication between ISPs, ISP routers can communicate by LAN protocols, such as FDDI or Ethernet. As described in Section 2, this architecture achieves the bilateral policy model of the LAN-IX and allows participating ISPs to establish BGP4 sessions directly over LAN switches.

Problems of LAN-IXes

Although a shared exchange subnet makes it easy for participating ISPs to configure data-link layer (LAN) interfaces and set up routers to communicate with each other in a LAN-IX, this architecture results in several restrictions and problems as follows:

(1) Switching speed

ISPs require a higher volume of traffic exchange in a LAN-IX. For example, although some of the largest ISP backbones consist of 10Gbps(OC-192)

in POS(Packet over Sonet) links, most of the major LAN-IXes provide only 100Mbps or 1Gbps throughput with Ethernet technology. An interface speed of 1Gbps is not fast enough to exchange data traffic between large ISPs in the current Internet.

(2) **Security**

In a LAN-IX, participating ISPs' routers connect to a shared subnet to exchange traffic with each other. In a LAN-IX, a third party router can send any bogus packet to another router, or inject unexpected traffic into other routers. For example, an ISP can forward all the traffic into another ISP router by manually configuring the next-hop attributes in the ISP router. This type of configuration is called a 'third party next-hop' and is still a critical problem in current LAN-IX architecture.

(3) **Additional routers**

A participating ISP has to locate its router physically near a LAN-IX, because of a physical cable or fiber length restrictions. An ISP usually brings its router into the building where the LAN-IX's switch is located, and the ISP also prepares another leased line from an ISP location into the router located near the LAN-IX.

(4) **Scalability**

A LAN-IX uses fixed size shared subnet as an 'exchange subnet'. A fixed size network address space is not scalable, because an expanding exchange subnet requires changes in the network address and the network mask of all participating routers.

2.4 ATM based IXes

Another architecture adopted by some of the major IXes is based on ATM(Asynchronous Transfer Mode) technology. In this case, an IX is ATM switched network, and participating ISPs connect their ATM routers into one of the ATM switches provided by the IX. We call this kind of IX, 'ATM-IX'.

Since ATM switches provide virtual circuits, called PVC(Permanent Virtual Circuit) between ATM routers, a participating ISP of an ATM-IX can establish interconnections to other ISPs over virtual circuits. Because ATM devices can handle many PVCs in a single physical link, participating ISPs of an ATM-IX can

interconnect to many other ISPs through a single physical link.

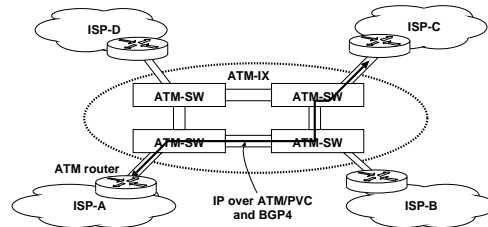


Fig. 5 ATM based IX

Fig. 5 is an example of ATM-IX implementation. In this figure, ISP-A and ISP-C interconnect to each other. Both ISP-A and ISP-C connect their ATM routers into the IX, and an IX provider configures ATM switches to establish a PVC between these two routers. Since this PVC acts as a point-to-point link between ISP routers, ISP routers can communicate directly over the PVC. In the ATM-IX architecture, the entire functionality of the IX provides only data-link connectivity as ATM PVCs. This architecture makes an ATM-IX 'bilateral', and allows participating ISPs to establish BGP4 sessions and to transmit data traffic over PVCs.

Problems of ATM-IXes

We can assume that an ATM PVC is a virtual point-to-point circuit between two participating ISPs in an ATM-IX. However, using ATM technology to transmit IP datagrams has several problems such as cell transmitting speed, and overhead. These problems are also critical in ATM-IXes. Next, we discuss several ATM-IX problems as follows:

(1) **Switching speed**

In ATM-IXes, ATM switching speed inside the IX is problematic because ATM cell switching requires high performance and an expensive forwarding table look up. Although most current ATM-IXes provide up to a 622Mbps(OC-12) ATM link for exchanging data traffic, this speed is not fast enough to exchange traffic between large ISPs in the current Internet.

(2) **Overhead**

Communicating with TCP/IP protocols over ATM switches has an overhead problem, namely the 'cell tax'. ATM-IXes also have the same problem. ATM protocol is designed to transmit a small

and fixed size packet consisting of 48 octets of data and 5 octets of header; that is, at least 9.4% of header overhead exists when communicating with an ATM. When communicating with TCP/IP protocols over ATM networks, the overhead might be more than 15% in a high speed network.

(3) Operational cost and scalability

Since an IX has to configure and manage many PVCs between ISPs' routers, operational and management costs are expensive and the scalability problem remains. When an IX is implemented with ATM PVC technology, up to $O(N \times N)$ PVCs are needed to interconnect N participating ISPs to each other, and all of these PVCs must be configured individually.

3. Abstraction of network devices

Before we propose a new IX architecture, we discuss about abstraction of network devices by MPLS technology. In this section, we introduce MPLS technology, followed by the discussion and evaluation of abstraction of network devices by MPLS.

3.1 MPLS overview

MPLS (Multi-Protocol Label Switching) is a new routing paradigm, discussed and standardized in IETF²⁾. The basic concept of MPLS technology is transmitting a data packet by label information instead of destination address stored in the original data packets.

Although MPLS stands for *multi-protocol* and allows us to transmit any network layer protocol such as IP, IPX and AppleTalk, we discuss about transmitting IP datagram in this paper.

A MPLS network consists of LSR (Label Switching Routers) which recognize label information for each data packet. 2 kinds of LSRs exist in a MPLS network. An Edge LSR is a border router between a MPLS network and non-MPLS networks. A Core LSR is router inside a MPLS network and Core LSRs transmit label encapsulated packets.

A LSR establish a virtual path, called LSP (Label Switched Path), by a signaling protocol, such as RSVP-TE or LDP. LSP is a sequence of LSRs in which a label encapsulated packet should traverse in that order.

Fig. 6 shows basic concept of a MPLS. We briefly introduce the packet forwarding behavior in a MPLS network with this figure.

(1) A LSR establish a LSP (Label Switched

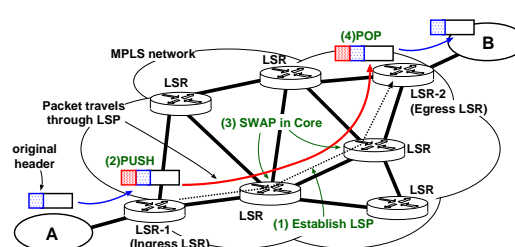


Fig. 6 Concept of MPLS

- (1) Path) by a signaling protocol.
- (2) When an Edge LSR (called an Ingress Edge LSR) receives an IP packet which should be transmitted through a LSP, the LSR adds (**PUSH**es) label information into the packet, and transmits the packet to the next LSR defined in the LSP.
- (3) Core LSRs replace (**SWAP**) label information of data packets and transmit them to the next LSR in the LSP.
- (4) When an Edge LSR (Egress Edge LSR) at the end of the LSP receives the packet, the LSR removes (**POP**s) label information and transmits the packet to the destination stored in the original IP header.

MPLS has a benefit of flexibility in forwarding data packets. LSRs only look up label information when they forward packets. IP header information has no affect in routing decision in LSRs. A typical application of MPLS is 'traffic engineering'¹³⁾, by which ISP operators can design and control backbone traffic efficiently.

MPLS also provides data-link medium independency in consisting MPLS network. Any physical and data-link medium is available for Edge-Core or Core-Core interconnection. Currently we are using POS (Packet Over Sonet), ATM (Asynchronous Transfer Mode) and GbE (Gigabit Ethernet) for our MPLS network. Even POS OC-768 which is the 40Gbps circuit and the fastest interface in the current technology, is available for a MPLS backbone.

3.2 Abstraction of network devices

Using MPLS technology enables abstraction of network devices. In a MPLS network, a LSR has a virtual network device which is connected to other LSRs by some LSPs. A LSR also transmits data packets through LSPs. A LSR logically separates LSPs from physical devices, so that the LSR could manage redundancy or load balancing.

In a MPLS network, a LSR has two kinds of connections. One is real connections to neigh-

bor LSRs, where ‘real’ means the physical (layer 1) devices / circuits and data-link medium connections. LSRs operate and manage real connections for a ‘control plane’ which is a network to realize virtual connections, e.g. LSPs, described below.

A LSR also has virtual connections, e.g. LSPs to other LSRs. An Ingress Edge LSR handles routing information for a specific destination of data packets and assigns LSP to those destination. In other words, an Ingress Edge LSR assigns virtual connection for data packets, instead of assigning physical interface nor physically neighboring routers. We call this virtual network as ‘data plane’.

Fig. 7 shows the usage of virtual network devices in a MPLS network. LSR-1 and LSR-2 are Edge LSRs in the MPLS network. LSR-1, LSR-2 and two Core LSRs have real network devices and real circuits between each other in this figure. For example, LSR-1 has a GbE interface and GbE connection to neighboring Core LSR. All physical and data-link connections consists of control plane of the MPLS network.

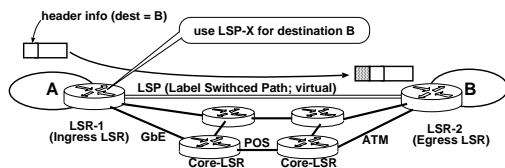


Fig. 7 Abstraction of network devices

LSR-1 also have virtual network devices, that is, LSP-X which is terminated at LSR-1 and LSR-2. The LSP-X act as virtual connection between LSR-1 and LSR-2. MPLS allows LSR-1 to assign LSP-X for the destination of network B, instead of assigning physical interface. LSR-1 also transmit data packet for the network B through the virtual connection, e.g., LSP-X.

Abstraction of network devices, that is, using virtual network devices and virtual connections, provides numerous benefits in consisting high speed network.

- Scalability.

Since the control plane is an IP network of LSRs, MPLS network could be hierachical and easy to extend.

- Data-link medium independency.

MPLS is multi-protocol in data-link medium. We can use any physical and atalink medium such as POS, ATM or GbE.

- Redundancy.

LSRs separate virtual connections, e.g. LSPs from physical interface. A LSR can have an alternate path for a LSP. A LSR also changes the router for a LSP when any trouble exists in the current path.

- Load balancing.

A LSR can establish multiple LSPs for a single destination so that LSR can transmit traffic through physically separated paths.

4. MPLS-IX Architecture

In this section, we propose a new IX architecture **MPLS-IX** which is based on the MPLS (Multi-Protocol Label Switching) technology. As denoted in Sec. 3, MPLS provides abstraction of network devices. **MPLS-IX** is an implementation of virtual network mechanism of MPLS.

In this section, we describe the new IX model and MPLS based IX architecture. In the latter part of this section, we discuss the benefits of MPLS based IXes.

4.1 Model of MPLS-IX

In **MPLS-IX**, we use MPLS mechanism between participating ISPs. As usual, an ISP uses MPLS in its closed network, and does NOT use any MPLS mechanism in inter-domain environment. Instead, in our proposing architecture, we use inter-domain MPLS mechanism between participating ISPs.

The basic model of **MPLS-IX** consists of two parts, that is, (1)establishing LSPs (Label Switching Paths) between participating ISPs and (2)transmitting actual data traffic through LSPs between those ISPs. As denoted in Sec. 3, we can assume that a LSP is a virtual connection between LSRs. LSRs, that is participating ISPs routers, transmit any actual data packet through LSPs.

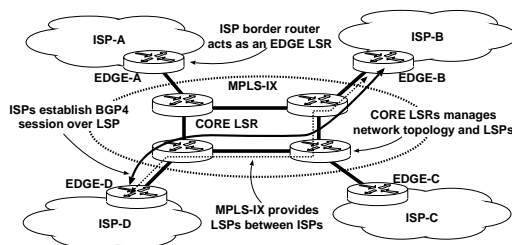


Fig. 8 MPLS-IX

In the **MPLS-IX** model, the main part of the **MPLS-IX** is a network of Core LSRs, called a ‘IX backbone’. Because **MPLS-IX** is a net-

work of MPLS capable IP routers, we can apply normal IP operation and management technologies to **MPLS-IX**, thereby controlling topology information, and obtaining redundancy, for example.

When an ISP participates with a **MPLS-IX**, the ISP connects a MPLS capable router to the nearest Core LSR. A participating ISP router acts as an Edge LSR in the MPLS network. To exchange traffic over a **MPLS-IX**, an ISP has to establish LSPs to other ISP routers, called peering routers, and exchange routing information over the LSP.

4.2 Architecture of MPLS-IX

In this section, we describe the architecture of **MPLS-IX**. As mentioned in section 4.1, the IX backbone consists of Core LSRs, and participating ISPs connect their Edge LSRs to one of Core LSRs.

Fig. 9 illustrates an example of establishing LSPs and exchanging routing information between participating ISPs in a **MPLS-IX**. In **MPLS-IX**, the following steps are necessary to achieve actual data traffic exchange:

- (1) Preparing physical and data-link connections between routers
- (2) Enabling MPLS and Running a LDP between MPLS routers.
- (3) Establishing LSPs between Edge routers that desire to communicate with each other
- (4) Exchanging routing information by BGP4 between Edge routers

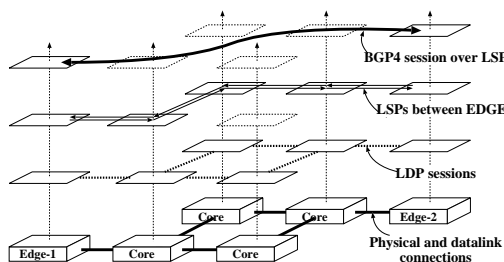


Fig. 9 MPLS-IX architecture

First, Core LSRs need physical and data-link connections between each other. The IX backbone consists of connections between Core LSRs. Edge LSRs also need to connect to one of the Core LSRs. As noted several times, one of the key features of the **MPLS-IX** is the independency of data-link mediums. In other words, both Core-Core and Core-Edge connections can consist of ATM, POS, FDDI or Giga-

bit Ethernet as data-link mediums.

To apply MPLS technology to an IX, we need to enable MPLS features and to run a signaling protocol between MPLS routers. Currently, two major signaling protocols for the MPLS exist. Some major router vendors support RSVP (Resource reSerVation Protocol)¹⁴ in their products in the early stage of MPLS. Recently, LDP (Label Distribution Protocol)¹⁵ is also available in major router vendors' products as another solution. In this paper, we use LDP as the signaling protocol because LDP has flexibility in managing LSPs in a **MPLS-IX**.

Edge LSRs, which are participating ISP border routers have to establish LSPs to exchange routing information and actual data traffic over **MPLS-IX**. Fig. 9 illustrates Edge-1 and Edge-2 establishing LSPs between each other. Since MPLS defines a LSP to be unidirectional, both Edge-1 and Edge-2 have to set up LSPs to establish bi-directional virtual paths.

After the establishment of LSPs between Edge LSRs, ISP routers communicate with BGP4 and exchange routing information between each other. In Fig. 9, Edge-1 and Edge-2 communicate with BGP4, to exchange routing information.

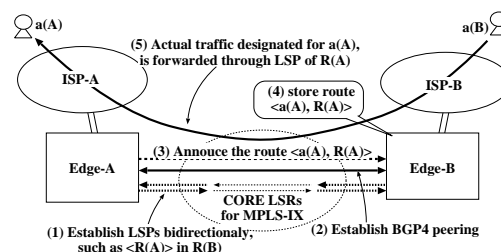


Fig. 10 actual transfer through LSP

Participating ISPs transmit actual data traffic through LSPs after exchanging routing information by BGP4. Fig. 10 illustrates the packet transmission mechanism in the **MPLS-IX**. Suppose that ISP-A and ISP-B connect to **MPLS-IX** and they establish both LSPs and a BGP4 session between their routers. If ISP-A announces a route for an address space a_A with the next-hop attribute R_A , then R_B obtains routing information such as (a_A, R_A) , and installs this route into its forwarding table. MPLS label encapsulation specification defines the behavior of Edge LSRs so that, if (1) Edge LSR has a route to a_A with next-hop R_A , (2) no LSP exists for the destination a_A , and (3) LSP_x

exists with a destination of R_A , then the Edge LSR must forward datagrams to a_A through LSP_x . This mechanism allows Edge LSRs to establish LSPs on a peer basis, instead of on a route basis so that **MPLS-IX** can reduce the total number of LSPs in its backbone.

4.3 Benefits of MPLS-IX

MPLS-IX architecture has the benefit of applying MPLS technology to the IX architecture proposed in this paper. The most important feature in applying MPLS technology is the independency of data-link mediums. As a result, our architecture contains the following features:

Migration of data-link mediums

A participating ISP can connect its router with any data-link medium. MPLS works fine over any of POS, ATM, or Gigabit Ethernet. An ISP can choose any medium that MPLS supports. The Independency of data-link mediums provides flexibility in implementing an IX, especially when installing and operating participating ISP routers. One can choose either the cheapest medium or the best performance medium.

Highest speed capability

Since **MPLS-IX** works with not only ATM or Gigabit Ethernet but also with POS links, the IX provides the highest speed connectivity between participating ISPs, such as 10Gbps(OC-192) or more. Furthermore, as discussed in IETF²⁾, MPLS will support WDM or DWDM technologies, and higher speed data-links will be available in the near future.

Widely distributed IX

By using WAN (Wide Area Network) interfaces such as ATM or POS, a **MPLS-IX** provider can expand Core LSRs to widely distributed areas. On the other hand, an ISP can also connect its Edge LSR with a WAN interface. An ISP does not need to put an ISP router into the IX's co-locating spaces.

Scalability

MPLS-IX has a scalability feature since Core LSRs hold only topological information for a MPLS network and LSP information. Core LSRs do not hold any routing information exchanged between participating ISPs. Additionally, since **MPLS-IX** is an IP network, the IX is more extendable than other IX architectures based on layer 2 technologies.

5. Evaluation

In our research, we tested basic feature of **MPLS-IX**, that is, we built a testbed and ex-

change traffic over the testbed. We also evaluated the performance and scalability of a typical implementation of **MPLS-IX**. In this section, We report the outcome of these evaluation.

5.1 Behavior of basic features

In our research, we built a testbed to experimentally test the interconnection between ISPs over **MPLS-IX**. Fig. 11 briefly illustrates the structure of our testbed. In this figure, Core-1~5 and Edge-1~3 represent Core LSRs and Edge LSRs, respectively. In a **MPLS-IX**, the IX backbone consists of Core LSRs. We note that the IX provider prepares and operates all the Core LSRs, Core-1~5. Edge LSRs are participating ISP border routers, and are operated by each ISP. We also note that we used Juniper routers for all the MPLS routers in this testbed.

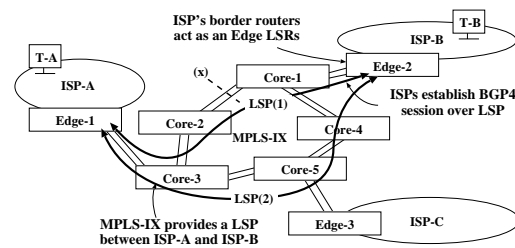


Fig. 11 MPLS-IX testbed

In our testbed, we configured Core and Edge LSRs as follows:

- (1) Enabling MPLS and a LDP on both Core and Edge LSRs. As noted before, we use LDP as a signaling protocol.
- (2) Configuring an OSPF protocol between Core LSRs. An IX provider runs the OSPF only in the IX backbone and does not allow participating ISPs to run the OSPF in their Edge LSRs.
- (3) Configuring static routes in Edge LSRs. By configuring both LDP and static routes in Edge LSRs, Edge LSRs establish LSPs to peering routers.
- (4) Configuring BGP4 in Edge LSRs. In **MPLS-IX**, a participating Edge LSR needs to establish BGP4 sessions with peering routers. In our testbed, we established three BGP4 sessions between Edge-1 and Edge-2, Edge-2 and Edge-3, and Edge-1 and Edge-3.

After we configured all the routers as previously described, we made three tests to ensure the behavior of traffic exchange in **MPLS-IX**. The first test examined the normal behavior

of the **MPLS-IX** interconnection model. Two other test simulate illegal cases.

Normal case:

Edge-1 and Edge-2 established a BGP4 and exchange data traffic over LSPs between these routers. In this figure, two terminals T-A and T-B communicated through the LSP (1). This test shows that the two ISPs interconnected to each other over a **MPLS-IX** can exchange data traffic over LSPs.

Case of link failure:

We disconnected a physical link at 'x' to simulate link failure. We confirmed that two terminals, T-A and T-B, could still communicate through LSP (2). **MPLS-IX** is a network that provides redundancy in the IX backbone. This test shows that **MPLS-IX** provides backup routes in its backbone.

Case of critical failure:

We shutdown router Core-5 after disconnecting the physical link at 'x' to simulate router failure. In this case, after a BGP4 Keepalive timeout, Edge-1 and Edge-2 disconnected the BGP4 session. In other words, Edge-1 and Edge-2 released routing information which had been exchanged between these routers, and both Edge-1 and Edge-2 routers selected another route instead of the withdrawn routes.

5.2 Evaluation of performance

We also evaluated performance of packet forwarding by MPLS routers (LSRs). In theory, MPLS packet forwarding requires additional 4 octets space for each packet to store label information. In this section, we discuss and evaluate how the MPLS packet forwarding mechanism reduces performance of packet forwarding in **MPLS-IX**.

At first, we calculate maximum bandwidth of a circuit. We assume that two LSRs (Label Switching Routers) connect each other by a single circuit C . We define that the maximum line speed of C is S [bps] and data-link header length is L [octets]. In this case, we can represent maximum throughput for packets whose data length is x , as $T_1 = S/((L + x) \times 8)$ [pps].

On the other hand, when we use MPLS mechanism to transmit data packet, we need 4 octets more to store label information. That is, maximum throughput of a MPLS environment is $T_2 = S/((L + 4 + x) \times 8)$ [pps].

Fig. 12 shows the logical values and actual values packet forwarding performance for the case when C is GbE (Gigabit Ethernet). We measured actual packet forwarding perfor-

mance with packet length x is 64, 256, 512, 1024, 1496. We used Hitachi GR2000 to measure the actual values, but most LSR implementations (which support hardware forwarding) achieves similar values.

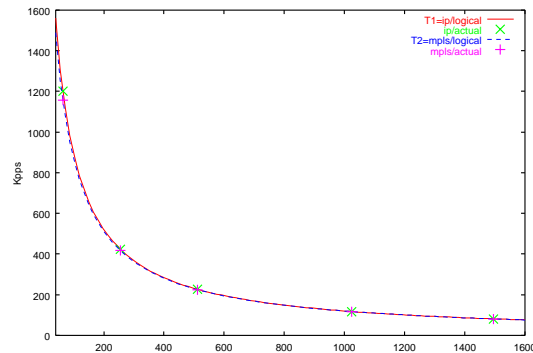


Fig. 12 Throughput of MPLS (1)

For example, when packet length x is 40 which is the minimum data length (without any data) of normal IP packet, T_1 and T_2 represent as follows:

$$\begin{aligned} T_1 &= 1,000,000,000/((38 + 40) \times 8) \\ &= 1,602,564 [pps] \\ T_2 &= 1,000,000,000/((48 + 4 + 40) \times 8) \\ &= 1,524,390 [pps] \end{aligned}$$

From these equations, we find that MPLS mechanism reduces $(T_1 - T_2)/T_1 = 0.04878$ of packet forwarding performance.

Similarily, when packet length x is 1500 which is typical data size of burstable data traffic, $T_1 = 81,274$, $T_2 = 81,063$ and and reduction of packet forwarding performance is 0.002596.

We also refer to the maximum throughput in bandwidth. It's obviously that bandwidth is represented in $x \times T_1$ for normal IP packets and $x \times T_2$ for MPLS encapsulated packets. Fig. 13 show the maximum bandwidth of normal IP packets and MPLS encapsulated packets. In this figure, lines represent logical values, and points represent actual measured values in our test.

From operational point of view, ISP engineers or architect decide to upgrade their backbone when the usage of physical circuits could keep 30~70. In this sense, we can assume that affect of MPLS encapsulation is small enough both in theory and in actual packet forwarding.

6. Conclusion

In this paper, we proposed a next generation IX architecture **MPLS-IX** by applying MPLS

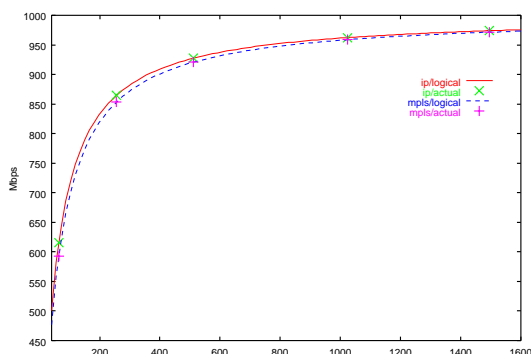


Fig. 13 Throughput of MPLS (2)

technology for interconnection between ISPs. IXes which are based on MPLS technology have the following benefits:

- (1) Migration of data-link medium. ISPs can connect into the IX and interconnect to other ISPs with data-link mediums such as POS, ATM, and the Gigabit Ethernet.
- (2) Unlimited bandwidth capability. ISPs can transmit a high volume of traffic, for example, up to 10Gbps (POS OC-192) or more.
- (3) Widely distributed IX. An IX provider can distribute the Core LSRs of **MPLS-IX** to widely distributed areas. Participating ISPs also need no additional routers in IX spaces.
- (4) **MPLS-IX** is highly scalable. Core LSRs have only topological information for the MPLS network, and hold no routing information exchanged between participating ISPs. Additionally, the **MPLS-IX** backbone is an IP network, and thus, an IX provider can easily extend the IX structure.

We also built a **MPLS-IX** testbed, and tested traffic transmission between participating ISPs. In this test, we confirmed that ISP routers transmitted data traffic over LSPs in the **MPLS-IX**. We ensured that path recalculation in the MPLS backbone also worked well after partial physical link failure.

As the Internet becomes more and more important to telecommunication infrastructure, IXes also play an important role in the Internet. ISPs need not only to exchange higher volume traffic with each other, but also need stable and reliable mechanisms to transmit commodity traffic.

We will do additional research regarding the performance evaluation of a **MPLS-IX** imple-

mentation, and we will also consider both the stability and the reliability of the implementation.

Acknowledgments

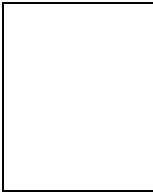
We would like to thank Dr. Hayashi of Reitaku University for his helpful advice, as well as Mr. Matsushima of Japan Telecom and Mr. Nishio of the Internet Research Institute who provided us with many useful comments.

References

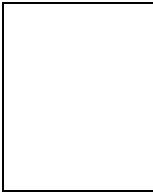
- 1) Geoff Huston: "Interconnection, Peering and Settlements", The Internet Protocol Journal, Vol 2, No 1, Mar. 1999.
- 2) IETF: "Internet Engineering Task Force", <http://www.ietf.org/>
- 3) Bill Manning: "Exchange Point Information", <http://www.ep.net/>
- 4) PAIX: "Palo Alto Internet eXchange", <http://www.paix.net/>
- 5) WCom: "MAE Information", <http://www.mae.net/>
- 6) LINX: "London InterNet eXchange", <http://www.linx.net/>
- 7) Telehouse: "New York International IX", <http://www.nyiix.net/>
- 8) AMS-IX: "Amsterdam IX", <http://www.ams-ix.net/>
- 9) WIDE Project: "NSPIXP", <http://jungle.sfc.wide.ad.jp/NSPIXP/>
- 10) Ikuo Nakagawa, Eisuke Hayashi, Toru Takahashi: "Direction of Next Generation Internet eXchanges", Transaction: Communications Special Issue on "Internet Technology", IEICE, 2001
- 11) Y. Rekhter, T. Li: "A Border Gateway Protocol 4", IETF RFC1771, Mar. 1995
- 12) E. Rosen, A. Viswanathan, R. Callon: "Multiprotocol Label Switching Architecture", RFC3031, Jan., 2001.
- 13) D. Awduche, J. Malcolm, J. Agogbua, M. O'Dell, J. McManus: "Requirements for Traffic Engineering Over MPLS", RFC2702, Sep., 1999.
- 14) D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan, G. Swallow: "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC3209, Dec., 2001.
- 15) L. Andersson, P. Doolan, N. Feldman, A. Fredette, B. Thomas: "LDP Specification", RFC3036, Jan., 2001.

(Received ??? ??, ????)

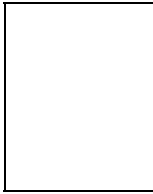
(Accepted ??? ??, ????)



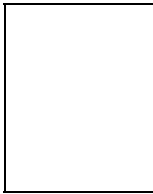
Ikuo Nakagawa was born in Toyama, Japan on August, 1968. He received the M.E. degree in system science from Tokyo Institute of Technology in 1993. He joined research division of INTEC Inc. in 1993, where he had been engaged in research on network operation and management, routing technology, and internet exchanges. He is a member of Information Processing Society of Japan and Internet Technology Research Committee. He is a board member of Next Generation IX Consortium, and a committee member of Toyama Regional IX Consortium. He also engages in IPv6 Deployment Committee of IAJapan.



Hiroshi Esaki He received the B.E. and M.E. degrees from Kyushu University, Fukuoka, Japan, in 1985 and 1987, respectively. And, he received Ph.D from University of Tokyo, Japan, in 1998. In 1987, he joined Research and Development Center, Toshiba Corporeation, where he engaged in the research of ATM systems. From 1998, he works for University of Tokyo as an associate professor, and works for WIDE project as a board member. He has been at Bellcore in New Jersey (USA) as a residential researcher from 1990 to 1991, and has engaged in the research on high speed computer communications. From 1994 to 1996, he has been at CTR (Center for Telecommunications Reserch) of Columbia University in New York (USA) as a visiting scholar. He has joined the University of Tokyo as an associate professor, and is currently interested in a high speed internet architerture, including MPLS technology, IP version 6 technology and mobile computing.



菊池 豊 1992年東京工業大学博士課程単位取得退学。同年より同大学情報工学科助手。1997年より高知工科大学情報システム工学科助教授。地域指向型のインターネットトラフィック交換の研究を行う。情報処理学会 DSM 研究会幹事。KPIX 実験研究協議会会長。博士(工学, 東京工業大学, 1994)。



永見 健一 1992年東京工業大学理工学研究科修士課程終了。同年(株)東芝入社。IETF MPLS WG で標準化活動を行い, CSR および MPLS に関する RFC を提出。現在, 東芝開発センターで MPLS および IPv6 の研究に従事。工学博士(東京工業大学, 2001)。