A Design of a Next Generation IX using MPLS Technology

Ikuo Nakagawa
INTEC Web and Genome Informatics Corporation

Hiroshi Esaki
University of Tokyo

Kenichi Nagami
Toshiba

Abstract

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 operate[3] 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, LINX[6], NYIX[7], AMIX[8], NSPIX[9], 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)[11] technology. One of the most important features of MPLS is the independency on data-link mediums. MPLS networks can consist of any data-link medium, for example, POS (Packet Over Sonet), ATM (Asynchronous Transfer Mode), or 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 applying MPLS technology, such as scalability, simple backbone operation, and the possibility of QoS mechanism.

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 propose a next generation IX architecture using the MPLS (Multi-Protocol Label Switching) technology. We introduce MPLS technology briefly at first, and 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, unlimited bandwidth, transmission speed, widely distributable feature, and scalability.

In section 4, 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 recalculations in participating ISPs.
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 $\frac{N(N-1)}{2}$, 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’ 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 IXes 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. 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 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 be-
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, NSPIXP2 and many other major IXes are based on the LAN-IX model.

![Figure 4. IX based on LAN technology](image)

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 IX 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 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.

![Figure 5. ATM based IX](image)

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 datalink 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. New IX architecture using MPLS

In this section, we propose a new architecture to implement an IX with the MPLS (Multi-Protocol Label Switching) technology. After we introduce MPLS technology briefly, 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.

3.1. MPLS

The basic concept of MPLS technology is transmitting IP datagrams by label information instead of destination address information. We call a router that transmits an IP datagram by label information a LSR(Label Switching Router). A MPLS consists of two kinds of LSRs. Edge LSRs are border routers between a MPLS network and non-MPLS networks. Core LSRs are routers inside a MPLS network and Core LSRs transmit labeled packets. We briefly describe the behavior of packet transmission by LSRs in MPLS network.

1. LSRs in MPLS network exchange label information between each other by a signaling protocol.
2. LSRs establish virtual paths, called LSP (Label Switched Path), based on label information.

3. When an Edge LSR (called an Ingress Edge LSR) receives an IP packet which should be transmitted through a LSP, the LSR adds (pushes) label information into the packet, and transmits the packet to the next LSR defined in a LSP.

4. Core LSRs transmit labeled packets along a LSP associated with labels.

5. When an Edge LSR (Egress Edge LSR) at the end of a LSP receives the packet through a LSP, the LSR removes (pops) label information and transmits the packet by a destination address stored in the IP header.

In a MPLS network, all the Core LSRs transmit packets by label information instead of a destination address stored in the IP header. This mechanism provides a flexible packet forwarding feature and is the reason why many large ISP operators hope that MPLS has a useful routing feature known as ‘traffic engineering’.

We note that there are two major signaling protocols for MPLS. One is RSVP (Resource ReSerVation Protocol) extension[12]. When LSRs communicate with RSVP as a signaling protocol, LSRs have to control and manage topology information of the MPLS network by OSPF or IS-IS. The other signaling protocol is LDP (Label Distribution Protocol)[13]. LDP allows LSRs to control routing information ‘hop-by-hop’, that is, LSRs do not need to handle the whole topology of the MPLS network.

One of the most critical features of MPLS is that MPLS enables LSRs to transmit packets in a multi-protocol environment. That is, LSRs can carry any network protocols, for example IP, IPv6, IPX and AppleTalk through LSPs. LSRs can also establish LSPs over any data-link medium, for example POS (Packet Over Sonet), ATM (Asynchronous Transfer Mode), and a Gigabit Ethernet.

3.3. Architecture of MPLS-IX

In this section, we describe the architecture of MPLS-IX. As mentioned in section 3.2, the MPLS-IX backbone consists of Core LSRs, and participating ISPs connect their Edge LSRs to one of Core LSRs.

Fig. 7 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
First, Core LSRs need physical and data-link connections between each other. The MPLS-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 Gigabit 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)[12] in their products in the early stage of MPLS. Recently, LDP (Label Distribution Protocol)[13] 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. Currently, two major signaling protocols for the MPLS exist. Some major router vendors support RSVP in their products in the early stage of MPLS. Recently, LDP (Label Distribution Protocol) 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.

After the establishment of LSPs between Edge LSRs, ISP routers communicate with BGP4 and exchange routing information between each other. Fig. 7 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. Fig. 7 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.

Participating ISPs transmit actual data traffic through LSPs after exchanging routing information by BGP4. Fig. 8 illustrates the packet transmission mechanism in a 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 \( \langle a_A, R_A \rangle \), 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) \( LSP_B \) exists with a destination of \( R_A \), then the EDGE LSR must forward datagrams to \( a_A \) through \( LSP_B \). 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.

3.4. 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 IETF[2], 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 extensible than other IX architectures based on layer 2 technologies.

4. Result of experiment of MPLS-IX

In our research, we built a testbed to experimentally test the interconnection between ISPs over MPLS-IX. Fig. 9 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.

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 MPLS-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 MPLS 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. Conclusion

In this paper, we proposed a next generation IX architecture by applying MPLS 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 implementation, and we will also consider both the stability and the reliability of the implementation.

Acknowledgement

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

References


