Recommended router configuration for connecting to an Internet Exchange

1. Abstract

An Internet Exchange (IX) is an important infrastructure for exchanging Internet traffic between network service providers. For this reason, AS operators should have a common understanding and be aware of the technical requirements of connecting to an IX.

This document aims to improve the operation and stability of an IX. It describes the recommended configuration of routers connected to an IX by an AS operator, including Peering LAN, a LAN on which many individual ASes connect to.

This document does not describe information about routing outside of connecting to an IX. For example, the contents of routing information which a network service provider advertises to their peers and the contents of traffic which is exchanged between network service providers are excluded from this document.

2. Terminology

IX

A point for exchanging Internet traffic smoothly between network service providers. In this document, an IX composes of an Ethernet network and means a broadcast domain in which many specific members are connecting to.

IX Member

An organization connecting to an IX and exchanging routing information by BGP4(+). An IX member is an AS operator.

Connecting router

A router that an IX Member uses to connect to an IX.

AS operator

An organization operating an AS.

Bilateral peering

The exchange of routing information between connecting routers of 2 AS operators.

Multilateral peering

The exchange of routing information between multiple IX members within one BGP peer. This is achieved by establishing BGP peering between a connecting router of an IX member and a route server provided by the IX provider.

3. Configuration of connection

3.1. Device

A device connected directly to an IX port is a Layer 3 device (BGP router) by default.

Consult your IX provider if you are unable to connect a Layer 3 device (BGP router) through a Layer 2 device.

3.2. Access Line

It is recommended that you use a circuit enabled Link Fault Pass Through for connecting to an IX. You need to be careful when connecting a Layer 3 device (BGP router) through a long-distance access line.

Troubleshooting becomes easier if a specification of the access line is provided to the IX provider in advance.

4. Configuring a connecting router

4.1. MAC configuration

4.1.1. Source MAC address

Source MAC assesses MUST be:

- a unicast address.
- unique to one port and VLAN.

Consult your IX provider if you need to use a different MAC address for each address family (that is, you need to use different chassis or different interface on the same chassis).

*) It is recommended that each port has a unique MAC address. Having multiple source MAC addresses on one port makes it difficult for an IX to troubleshoot any issues.

*) Some IX providers implement traffic filtering based on the port MAC address. If you (IX Member) change your MAC address, you need to inform the IX of this change in advance.

4.1.2. Destination MAC address

Destination MAC addresses MUST be unicast addresses, except for the following exceptions:

- ARP (broadcast address)
- IPv6 Neighbor Discovery (multicast address)

*) Frames with broadcast, multicast and unknown-unicast address can cause Layer 2 storms and may hinder other IX members from exchanging traffic once a loop occurs.

4.1.3. EtherType

The acceptable EtherTypes for connecting to an IX are as follows.

- 0x0806（ARP
- 0x0800（IPv4
- 0x86DD（IPv6

Follow the specification of your IX service for Tag VLAN (0x8100, 0x88A8, 0x9100, etc.) and LACP(0x8809).

4.2. IP configuration

4.2.1. IP address

Configure the IP address assigned by an IX provider on a single logical port for each address family.

4.2.2. IP MTU

Follow the specification of the IX provider (it is 1500 bytes by default).

4.2.3. IPv6 link-local address

Follow the specification of the IX provider if you want to use a fixed link-local address.

4.3. Interface configuration

4.3.1. Stop sending link-local traffic

Link-local traffic MUST NOT be sent because it may harm other connecting routers and IX devices. It is also unnecessary traffic to exchange traffic of this kind on an IX. Examples of such traffic include:

- Vendor proprietary protocols, such as CDP, FDP, EDP, VTP/DTP, etc.
- Interior Gateway Protocols (IGPs), such as OSPF, IS-IS, EIGRP, etc.
- IEEE 802.1D Spanning Tree Protocol BPDU
- Multicast routing protocols, such as PIM-SM, PIM-DM, DVMRP, etc.
- BOOTP/DHCP
- DEC protocol (MOP), NetBIOS, etc.
- Protocols for gateway redundancy, such as HSRP, VRRP, ESRP, etc.
- IPv6 Router Advertisement
- Link Layer Discovery Protocol (LLDP)

4.3.2. Stop sending ICMP redirects

The Internet Control Message Protocol (ICMP) redirect function MUST be disabled because it may harm routing tables of other connecting routers.

4.3.3. Prohibit forwarding packets where the destination address is the directed broadcast address of an IX segment

This function MUST be disabled because it enables Smurf attacks. When a connecting router forwards an ICMP echo request packet with a directed broadcast address of an IX segment, other connecting routers will reply to the packet, initiating a Smurf attack.

4.3.4. Disable proxy ARP

Proxy Address Resolution Protocol (ARP) MUST be disabled to stop connecting routers replying to ARP requests for other IX members, which draws BGP packets and traffic.

4.3.5. Auto-negotiation, link fault signalling

Follow the specification of each IX service for pros and cons of configuration.

*) Auto-negotiation needs to be configured in order to enable link fault signalling for preventing a unidirectional failure in a single fiber of paired optical fiber of leased lines where Link Fault Pass Through is not enabled.

4.3.6. LACP

Follow the specification of each IX provider for the pros and cons of the configuration of the Link Aggregation Control Protocol (LACP) and recommended parameters.

*) LACP is useful for detecting unidirectional link failures between a connecting router and an IX device (such as traffic black holes caused when an interface of only one device of the link is ‘up’).

4.3.7. Minimum link in link aggregation

Configure based on the operational policy of each IX Member being sure to consider the design of traffic engineering. Troubleshooting is made easier if you (IX Member) provide the configuration to an IX provider in advancer.

4.3.8. Avoiding unnecessary link down detection (carrier-delay, hold-time, link debounce time)

Follow the recommended value of your IX provider.

*) Especially in an IX service providing physical redundancy by an optical switch. BGP session flaps and recalculation of the route is performed when a connecting router detects a momentary ‘down’ link on a physical interface on the switchover of the optical switch. To avoid this situation, make the connecting router ignore or delay the detection of the momentary link down or notification to upper protocols.

4.4. BGP configuration

4.4.1. Peer configuration

Do NOT configure BGP peering for IX members which you do not have a peering agreement with.

Immediately remove BGP peering for disconnected IX members.

*) Unnecessary ARP request (broadcast) and IPv6 Neighbor Solicitation (multicast) consume CPU resources of connecting routers of other IX members. When an IX provider re-assign the IP address previously assigned to a disconnected IX member, connecting routers send unintentional BGP packets to other IX members, to which the IP address is newly assigned to.

4.4.2. Authentication password

Consult your peer for use of authentication passwords (MD5, TCP-AO(TCP Authentication Option), etc.)

4.4.3. BFD

Consult your peer for Bidirectional Forwarding Detection (BFD) with a bilateral peer.

At the time of this writing, it is not possible to configure BFD with a BGP router which is a next-hop of routes received from a multilateral peer with a route server.

4.4.4. Maximum prefix limit

It’s recommended that you (IX Member) configure maximum prefix limit because it is effective as a defensive measure against traffic loss events caused by congestion of ports of a connecting router when the router has received a full-route from a misconfigured peer.

It’s recommended that you (IX Member) configure an alarm system to notify you when received routes are over a specific value for multilateral peers; the number of received routes from multilateral peers may increase unintentionally.

The parameter of the maximum prefix limit CAN be determined based on the operational policy of each IX Member.

4.4.5. Configuration for multilateral peer (route server)

It’s required that you (IX Member) configure the connecting router to allow it to receive routes which the first AS number of the AS Path is different from the AS number of the peer. This is because these kinds of routes are regarded as malformed routes by some router implementations but are not malformed in the context of multilateral peering.

4.5. Routing configuration

4.5.1. Do NOT advertise the prefix of IX segments

Do NOT advertise the prefix of IX segments. If possible, implement a route filter which denies the routes of IX segments from neighbor ASes.

Reachability to an IX segment from global space enables attacks such as malicious misrepresentation of routes attacks targeting TCP vulnerabilities.

When a connecting router receives the route of the connecting IX segment from other ASes, a router refers the route to the next-hop and causes traffic loss.

If the route is required to be advertised in an IGP in order to achieve traffic engineering or secure reachability internally of an AS, the advertisement should be only internal to the AS.

4.5.2. Do NOT send traffic to non-engaged peers

This situation often happens when an IX member connects multiple routers to an IX and establishes a BGP peer on only one router.

This causes unidirectional packet flow and prevents correct MAC learning on Layer 2 devices. This causes unknown unicast flooding which affects the traffic exchange of other IX members.

This can result in an incident similar to bandwidth theft attack, which makes other IX members carry traffic by configuring a static route.

4.5.3. Configure Next-Hop-Self

It's recommended that you (IX Member) configure Next-Hop-Self when you are going to advertise routing information received from a neighbor AS on an IX to iBGP in your AS. Inform your IX provider if Next-Hop-Self cannot be configured due to a routing design issue.

- To avoid traffic loss caused by the advertisement of the prefix of IX segments from neighbor ASes (as described above).
- To prevent harmful traffic exchange (as described above).
- Next-Hop-Self cannot be configured if ECMP of IGP is enabled for load balancing when multiple routers are connected to an IX.
- iBGP Multipath can be a solution for load balancing when configuring Next-Hop-Self.

5. Acknowledgements

First edition

The authors would like to thank Akira Kato of the WIDE Project and Toshinori Ishii of Internet Multifeed for editing the document, as well as IRS Meeting attendees and engineers of Japan Internet Exchange for their comments.

DIX-IE, AMS-IX, and LINX have published similar documents for their IX members, which were very helpful in drafting this document.

We would also like to thank Kuniaki Kondo, Tomoya Yoshida, Ryoko Nakanishi and the IRS Meeting secretariat for their support.

Revised edition

The authors would like to thank Shishio Tsuchiya, Shin Shirahata, Toshinori Ishii, and Peering BoF attendees for their comments.

6. References

- "Allowed Traffic Types on Unicast Peering LANs", AMS-IX(on-line). Available at https://ams-ix.net/technical/specifications-descriptions/allowed-traffic
- "LINX Memorandum of Understanding", Appendix 1, LINX(online). Available at https://www.linx.net/governance/mou
- Tomoya Yoshida, "Routing design on ISP backbone network - best practice”, Japan Network Information Center(online). Available at http://www.nic.ad.jp/ja/materials/iw/2004/proceedings/T7.pdf

7. Authors

Name: Toshitaka Hirao
Affiliation: KDDI Corporation
Contact: to-hirao@kddi.com

Name: Masataka Mawatari
Affiliation: Japan Internet Exchange Co., Ltd.
Contact: mawatari@jpix.ad.jp

Name: Yuya Kawakami
Affiliation: Internet Multifeed Co.
Contact: kawakami@mfeed.ad.jp

Name: Hiroyuki Ashida
Affiliation: BBIX Inc.
Contact: ashida@bbix.net

8. Indemnity

9. Distribution

Appendix A: Use of database services useful for peering

A-1. IRR（Internet Routing Registry)

IRR is a database of global routing information of the Internet and AS number. It’s useful for confirming contact information of routes and AS numbers and configuring routing control and route filtering as a reference. A representative IRR in Japan is JPIRR operated by JPNIC.
https://www.nic.ad.jp/ja/irr/index.html

A-2. PeeringDB

PeeringDB is a database of networks operated by AS operators, IXes, and data centers. It’s useful when for connecting to an IX, or checking if there is a network you want to peer with on an IX.
https://www.peeringdb.com/