BGP4+ Peering Using IPv6 Link-local Address
draft-kato-bgp-ipv6-link-local-00.txt

Status of this Memo

This document is an Internet-Draft and is in full conformance with all
provisions of Section 10 of RFC2026.

Internet-Drafts are working documents of the Internet Engineering Task
Force (IETF), its areas, and its working groups. Note that other groups
may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months
and may be updated, replaced, or obsoleted by other documents at any
time. It is inappropriate to use Internet-Drafts as reference material
or to cite them other than as ``work in progress.''

The list of current Internet-Drafts can be accessed at
http://www.ietf.org/1id-abstracts.html

The list of Internet-Draft Shadow Directories can be accessed at
http://www.ietf.org/shadow.html.

Distribution of this memo is unlimited.

The internet-draft will expire in 6 months. The date of expiration will
be April 20, 2002.

Abstract

In IPv6 it is possible to use link-local addresses to specify a E-BGP
connection. In this case, no global address space is necessary on a
link between two Autonomous systems or on an Internet Exchange point.
This memo proposes changes of the protocol defined in RFC2545, and some
guidelines of implementation and operation to enable E-BGP sessions
using IPv6 link-local addresses.

1. Introduction

RFC2858 [Bates, 2000] defines multiprotocol extension framework to BGP-4
[Rekhter, 1995] and IPv6 specific definition is described in RFC2545
[Marques, 1999] . RFC2545 also describes that the transport of BGP-4
can be either over IPv4 or IPv6.

It is intuitive that a E-BGP session is specified by a couple of global
addresses because only that style has been feasible solution in IPv4.
In IPv4, it is possible to use a private address [Rekhter, 1996] ,
however, the private address and the global address have syntactically
no difference. In IPv4, it is necessary to allocate an address space to
a link which connects more than one Autonomous Systems. Such address
space is offered by one of the Autonomous Systems on the link or by a
third party such as ep.net [Manning, EP.NET] project.

In IPv6, notion of address scope has been introduced [Hinden, 1998] . A
link-local address has a scope which is only valid within a single link.
It is independent from a global address and it remain unchanged in the
event of global address renumbering. Due to this property, several ICMP
messages are defined so that their source addresses are selected from
link-local addresses [Conta, 1998] . RIPng [Malkin, 1997] also uses
link-local address for the routing information exchanges.

While this memo describes BGP peering with link-local addresses, it does
not describe that it must not use a global prefix on each IX. It
encourages every implementation support it. We may introduce further
guidelines for the global prefix assignment on each IX after we have
enough experience on BGP peering with link-local addresses.

[2]. Link-local E-BGP Session

Because an E-BGP session, which connects different ASes, is assumed to
span over a single link, it is possible to use a pair of link-local
addresses to specify the E-BGP session. We call such an E-BGP session
as a link-local E-BGP session. In this case, it is not necessary to
assign global IPv6 addresses to the link. As IPv6 has a huge address
space, address conservation may not be necessary. However, routing
space is limited under current routing architecture and wasting routing
table entries should be avoided as much as possible.

Note: It is possible to specify an BGP session by a combination of a
link-local address and a global address. This is no benefit in this
configuration and such configuration is out of scope of this memo.

Note: It is possible to specify I-BGP session by a couple of link-local
addresses. However, I-BGP sessions generally span multiple hops, and
the number of subnets internal to an AS can grow up to 65,536, it is
natural to use global address. So the following discussions do not
apply to I-BGP sessions.

In some cases, an E-BGP session may span multiple hops, which is known
as ebgp-multihop. In this case, it must to use global or site-local
addresses rather than link-local addresses. So the following
discussions do not apply.

It is possible to exchange IPv6 reachability information in BGP4+
connection using IPv4 transport. If the outbound interface of the E-BGP
session has no IPv6 address, it is almost identical to a link-local E-

BGP session in a sense that no global IPv6 address is specified. With a
few exception, following discussions will also apply to this case.

By using link-local addresses to specify E-BGP sessions, there are
following merits:

o It is possible to make DMZ (demilitarized zone) more neutral. None of
  the ASes connecting to a DMZ has to offer global address space.

o The announcement of the DMZ prefix is not recommended. However, if
  some AS advertises the prefix accidentally, the next-hop calculation
  in other ASes might be interfered with this announcement and might
  result in unexpected routing and congestion. By link-local E-BGP
  sessions, this kind of trouble is inherently avoidable.

Note: When an IPv6 node originates packets with global IPv6 destination
addresses, including ICMP echo reply and Time Exceed, through an
interface without any global IPv6 address, it will pick one of the
global address attached to it [Draves, 2001] .

3. Protocol Issues

According to RFC2545, the next hop field in the MP_REACH_NLRI attribute
is defined so that either of the following are acceptable:

o One global IPv6 address (Length of the next hop network address is 16)

o One global IPv6 address and one link-local IPv6 address (Length of the
  next hop Network Address is 32)

Based on the definition in RFC2545, an global IPv6 address should be
specified in the next hop field regardless of the transport of the E-BGP
session. When a link-local address is specified in the next hop field,
it must be valid on the shared link so that the other end of E-BGP
session can forward the traffic to that link-local address.

RFC2545 requires that an IPv6 global address must be specified in a part
of the next hop field. However, in a link-local E-BGP session, the
global IPv6 address is not necessary. It is possible to propose a
modification to RFC2545 to relax a rule in the next hop field, however,
it may cause a serious interoperability problem. Therefore, this memo
proposes the following changes in the behavior of each BGP speaker
without changing the packet format on the wire:

A) BGP-4 speakers should ignore the IPv6 global address specified in the
   next hop field in a MP_REACH_NLRI attribute learned over a link-local
   E-BGP session, provided if it does not match with one of the IPv6
   global prefixes assigned to the link. (It is possible to assign an
   IPv6 global prefix to an IX and have a link-local E-BGP session.)

B) BGP-4 speakers should not cease the link-local E-BGP session by
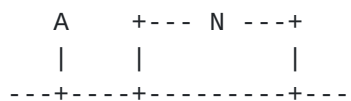   receiving a MP_REACH_NLRI information with a next hop field including

only a link-local IPv6 address.
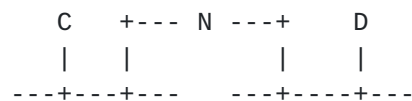
4. Implementation Issues

Any BGP-4 implementations which are capable to handle IPv6 reachability
information should be able to use IPv6 transport for BGP sessions.
Furthermore, those implementations should also be able to establish
link-local E-BGP sessions.

A link-local address has a scope which is limited on a single link. So
an IPv6 node with more than one interfaces may see the same link-local
address on each interface. There are two such cases:

1) The node has more than one interfaces attached to the link.

2) Different hosts which happen to have the same link-local address.

```
        A    +--- N ---+              C    +--- N ---+    D
        |    |         |              |    |         |    |
    ---+----+---------+---        ---+---+---    ---+----+---
         Case 1                  Case 2: if C's link-local address is
                                         the same as D's link-local address
```

As these configurations are not exceptional, all IPv6 nodes (especially
routers and IPv6 hosts which have more than one interfaces) should be
able to handle such cases:

A) Configuration language is properly designed so that different peers
   with the same link-local address can be distinguishable. Specifying
   the link-local address destination including <zone_id> is a good idea
   [Deering, 2001] .

B) BGP-4 implementations should be able to configure, at least per
   session basis, which IPv6 link-local address is to be filled in the
   next hop field of MP_REACH_NLRI attributes. This also helps when
   more than one interfaces are attached to the same link for the
   purpose of manual load distribution.

C) BGP-4 implementations should be able to configure, at least per
   session basis, which IPv6 global address is to be filled in the next
   hop field of MP_REACH_NLRI attributes. In the case of link-local E-
   BGP sessions, arbitrary IPv6 address can be specified while in a
   typical case an IPv6 global address attached to the node (other
   interface or loopback) will be specified.

D) If more than one interfaces are attached to the same link, the router
   must not be confused in terms of neighbor discovery context. This
   also applies a case where the peer node has multiple link-local
   addresses on the same interface (they share the same layer-2

destination address).

Note that D) should not be written in here but to IPv6 host/router
requirement documents. Until they are established, it remain here not
to be forgotten.

5. Operational Hints

5.1. Example at NSPIXP-6

In IPv6, a link-local address derived from the EUI-64 address of the
interface is used in most cases. It is not recommended to use this
EUI-64 derived link-local address in BGP because extensive BGP
reconfiguration will be necessary when the hardware is replaced due to
failure or to upgrade. In a E-BGP speaker, large amount of
configuration is necessary including routing protocols configuration,
prefix filters, and packet filters. So there is no reason to persist
auto configuration of the link-local address in BGP speakers.

In NSPIXP-6 [Kato, NSPIXP-6] , it is suggested to use link-local
addresses of the following format:

    fe80::ASN:ID

where ASN is the AS number, and ID is arbitrary 16bit number defined by
the AS and it identifies more than one routers from single AS attached
to the IX. This technique is similar to the Net39 experiment in 1995
[IANA, 1995] . For example, one of the router of AS2500 may use

    fe80::9c4:12

When 32bit AS number is in use, it can naturally be extended to

    fe80::ASN-high:ASN-low:ID

Not all of the implementations support IPv6 link-local E-BGP, the above
form of E-BGP configurations are used among a limited set of the routers
on NSPIXP-6 as of writing.

5.2. Network administration issue

By using link-local addresses at an IX, it might introduce some
difficulty on troubleshooting that it is not possible to ping the
router's interface on the IX from remote locations. As announcing
router advertisement is highly discouraged on an IX, every router may
have IPv6 arbitrary (different) global address. So each AS may
independently allocate a /64 prefix on the IX from the AS's available
address space and assign the interfaces of the routers of the the AS.
Because BGP4+ peering is specified by link-local addresses, assignment
of independent global prefixes does not affect the peering.

[6](). Address Allocation Guidelines

There are discussions regarding global address assignment to an IX.
Recent discussions in APNIC and RIPE/NCC communities converge to assign
a small global address space (in the case of APNIC a /64) to an IX.
ep.net assigns a /64 to participating IXes.

With this memo, the authors are not immediately against the assignment
of global IPv6 addresses to the link shared by more than one E-BGP
speakers. This is because there isn't enough experience of the link-
local E-BGP sessions and because not all of the implementations support
this functionality at the present time. It is suggested to review this
particular guideline in future, when there is more experience with this
technique.

[7](). Security Consideration

Security issues are not discussed in this memo.

[8](). IANA Consideration

There is no IANA consideration in this memo.

[9](). Acknowledgement

The authors would like to thank Jun-ichiro itojun Hagino, Tatsuya
Jinmei, Chris Fletcher, Jake Khuon, Barbara Roseman for valuable
comments and suggestions.

Authors' addresses

    Akira Kato
    USC/ISI & WIDE Project
    4676 Admiralty Way
    Marina del Rey, CA 90292, USA
    Tel: +1 310-448-9327
    Email: kato@wide.ad.jp

    Bill Manning
    USC/ISI
    4676 Admiralty Way
    Marina del Rey, CA 90292, USA
    Tel: +1 310-822-1511
    Email: bmanning@isi.edu

References

Bates, 2000.

T. Bates, Y. Rekhter, R. Chandra, and D. Katz, "Multiprotocol Extensions for BGP-4" in RFC2858 (June 2000). ftp://ftp.isi.edu/in-

notes/rfc2858.txt.

Rekhter, 1995.
Y. Rekhter and T. Li, "A Border Gateway Protocol 4 (BGP-4)" in RFC1771
(March 1995). ftp://ftp.isi.edu/in-notes/rfc1771.txt.

Marques, 1999.
P. Marques and F. Dupont, "Use of BGP-4 Multiprotocol Extensions for
IPv6 Inter-Domain Routing" in RFC2545 (March 1999).
ftp://ftp.isi.edu/in-notes/rfc2545.txt.

Rekhter, 1996.
Y. Rekhter, B. Moskowitz, D. Karrenberg, G. J. de Groot, and E. Lear,
Address Allocation for Private Internets (February 1996).
ftp://ftp.isi.edu/in-notes/rfc1918.txt.

Manning, EP.NET.
Bill Manning, EP.NET (EP.NET). http://www.ep.net/.

Hinden, 1998.
R. Hinden and S. Deering, "IP Version 6 Addressing Architecture" in
RFC2373 (July 1998). ftp://ftp.isi.edu/in-notes/rfc2373.txt.

Conta, 1998.
A. Conta and S. Deering, "Internet Control Message Protocol (ICMPv6) for
the Internet Protocol Version 6 (IPv6) Specification" in RFC2463
(December 1998). ftp://ftp.isi.edu/in-notes/rfc2463.txt.

Malkin, 1997.
G. Malkin and R. Minnear, "RIPng for IPv6" in RFC2080 (January 1997).
ftp://ftp.isi.edu/in-notes/rfc2080.txt.

Draves, 2001.
R. Draves, "Default Address Selection for IPv6" in draft-ietf-ipngwg-
default-addr-select-05.txt (June 2001). work in progress material.

Deering, 2001.
S. Deering, B. Haberman, T. Jinmei, E. Nordmark, A. Onoe, and B. Zill,
"IPv6 Scoped Address Architecture" in draft-ietf-ipngwg-scoping-
arch-02.txt (March 2001). work in progress material.

Kato, NSPIXP-6.
Akira Kato, NSPIXP-6 (NSPIXP-6). http://www.wide.ad.jp/nspixp6/.

IANA, 1995.
IANA, "Class A Subnet Experiment" in RFC1797 (April 1995).
ftp://ftp.isi.edu/inotes/rfc1797.txt.