Internet Engineering Task Force
INTERNET-DRAFT
Expires: October 12, 2001

Jun-ichiro Hagino
Research Laboratory, IIJ
April 12, 2001

IPv6 multihoming support at site exit routers
draft-ietf-ipngwg-ipv6-2260-01.txt

Status of this Memo

Abstract

The document describes a mechanism for basic IPv6 multihoming support, and its operational requirements. The mechanism can be combined with more sophisticated (or complex) multihoming support mechanisms, and can be used as a foundation for other mechanisms. The document is largely based on RFC2260 [Bates, 1998] by Tony Bates.

1. Problem

IPv6 specifications try to decrease the number of backbone routes, to cope with possible memory overflow problem in the backbone routers. To achieve this, the IPv6 addressing architecture [Hinden, 1998] only allows the use of aggregatable addresses. Also, IPv6 network administration rules [Durand, 1999] do not allow non-aggregatable routing announcements to the backbone.

In IPv4, a multihomed site uses either of the following technique to achieve better reachability:

- Obtain a portable IPv4 address prefix, and announce it from multiple upstream providers.

- Obtain single IPv4 address prefix from ISP A, and announce it from multiple upstream providers the site is connected to.

The above two methodologies are not available in IPv6, but on the other hand IPv6 sites and hosts may obtain multiple simultaneous address prefixes to achieve the same result.

The document provides a way to configure site exit routers and ISP routers, so that the site can achieve better reachability from multihomed connectivity, without violating IPv6 rules. Since the technique uses already-defined routing protocol (BGP or RIPng) and tunnelling of IPv6 packets, the document introduces no new protocol standard.

The document is largely based on RFC2260 [Bates, 1998] by Tony Bates.

2. Goals and non-goals

The goal of this document is to achieve better packet delivery from a site to the outside, or from the outside to the site, even when some of the site exit links are down.

Non goals are:

- Choose the "best" exit link as possible. Note that there can be no common definition of the "best" exit link.

- Achieve load-balancing between multiple exit links.

3. Basic mechanisms

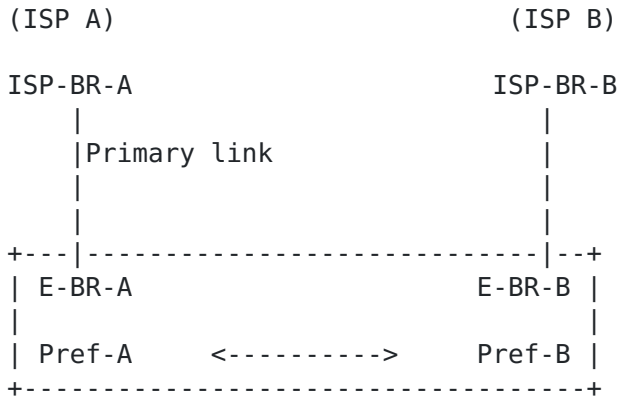
We use technique described in RFC2260 section 5.2 onto our configuration. To summarize, for IPv4-only networks, RFC2260 says that:

- We assume that our site is connected to 2 ISPs, ISP-A and ISP-B.

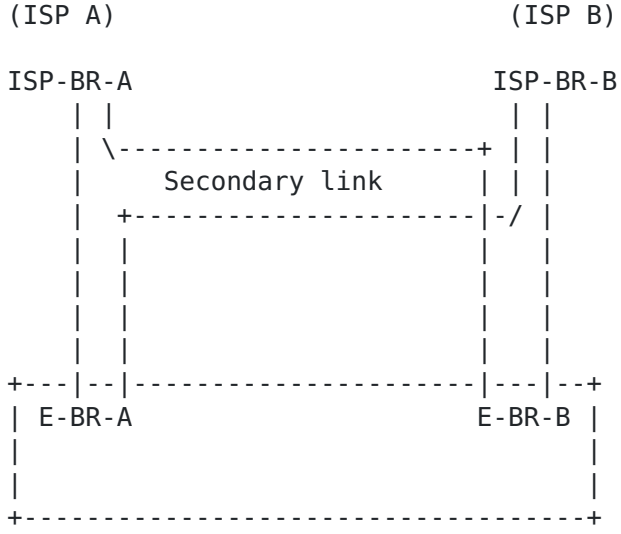
- We are assigned IP address prefix, Pref-A and Pref-B, from ISP-A and ISP-B respectively. Hosts near ISP-A will get an address from Pref-A, and vice versa.

- In the site, we locally exchanage routes for Pref-A and Pref-B, so that hosts in the site can communicate with each other without using external link.

o ISP-A and our site is connected by ``primary link'' between ISP router ISP-BR-A and our router E-BR-A. ISP B and our site is connected by primary link between ISP router ISP-BR-B and our router E-BR-B.



o Establish a secondary link, between ISP-BR-A and E-BR-B, and ISP-BR-B and E-BR-A, respectively. Secondary link usually is IP-over-IP tunnel. It is important to have secondary link on top of different medium than primary link, so that one of them survives link failure. For example, secondary link between ISP-BR-A and E-BR-B should go through different medium than primary link between ISP-BR-A and E-BR-A. If secondary link is an IPv4-over-IPv4 tunnel, tunnel endpoint at E-BR-A needs to be an address in Pref-A, not in Pref-B (tunnelled packet needs to travel from ISP-BR-B to E-BR-A, over the primary link between ISP-BR-A and E-BR-A).



o For inbound packets, E-BR-A will advertise (1) Pref-A toward ISP-BR-A with strong preference over primary link, and (2) Pref-B toward ISP-BR-B with weak preference over secondary link. Similarly, E-BR-B will advertise (1) Pref-B toward ISP-BR-B with strong preference over

primary link, and (2) Pref-A toward ISP-BR-A with weak preference over secondary link.
Note that we always announce Pref-A to ISP-BR-A, and Pref-B to ISP-BR-B.

o For outbound packets, ISP-BR-A will advertise (1) default route (or specific routes) toward E-BR-A with strong preference over primary link, and (2) default route (or specific routes) toward E-BR-B with weak preference over secondary link. Similarly, ISP-BR-B will advertise (1) default route (or specific routes) toward E-BR-B with strong preference over primary link, and (2) default route (or specific routes) toward E-BR-A with weak preference over secondary link.

Under this configuration, both inbound and outbound packet can survive link failure on either side. Routing information with weak preference will be available as backup, for both inbound and outbound cases.

4. Extensions for IPv6

RFC2260 is written for IPv4 and BGP. With IPv6 and BGP4+, or IPv6 and RIPng, similar result can be achieved, without violating IPv6 addressing/routing rules.

4.1. IPv6 rule conformance

In RFC2260, we announce Pref-A toward ISP-BR-A only, and Pref-B toward ISP-BR-B only. Therefore, there will be no extra routing announcement to the outside of the site. This conforms to the aggregation requirement in IPv6 documents. Also, RFC2260 does not require portable addresses.

4.2. Address assignment to the nodes

In IPv4, it is usually assumed that a node will be assigned single IPv4 address. Therefore, RFC2260 assumed that addresses from Pref-A will be assigned to nodes near E-BR-A, and vice versa (second bullet in the previous section).

With IPv6, multiple IPv6 addresses can be assigned to a node. So we can assign (1) one address from Pref-A, (2) one address from Pref-B, or (3) two addresses from both address prefixes, to a single node in the site.

This will allow more flexibility in node configuration. However, this may make source address selection on a node more complex. Source address selection itself is out of scope of the document.

4.3. Configuration of links

With IPv6, primary link can be IPv6 native connectivity, RFC1933 [Gilligan, 1996] IPv6-over-IPv4 configured tunnel, 6to4 [Carpenter, 2000] IPv6-over-IPv4 encapsulation, or some others.

If tunnel-based connectivity is used in some of primary links, administrators may want to avoid IPv6-over-IPv6 tunnels for secondary links. For example, if:

- primary links to ISP-A and ISP-B are RFC1933 IPv6-over-IPv4 tunnels, and
- ISP-A, ISP-B and the site have IPv4 connectivity with each other,

it makes no sense to configure a secondary link by IPv6-over-IPv6 tunnel, since it will actually be IPv6-over-IPv6-over-IPv4 tunnel. In this case, IPv6-over-IPv4 tunnel should be used for secondary link. IPv6-over-IPv4 configuration has a big advantage against IPv6-over-IPv6-over-IPv4 configuration, as secondary link will be able to have the same path MTU than the primary link.

4.4. Using RFC2260 with IPv6 and BGP4+

RFC2260 approach on top of IPv6 will work fine as documented in RFC2260. There will be no extra twists necessary.

4.5. Using RFC2260 with IPv6 and RIPng

It is possible to run RFC2260-like configuration with RIPng [Malkin, 1997] , with careful control of metric. Routers in the figure needs to increase RIPng metric on secondary link, to make primary link a preferred path.

If we denote the RIPng metric for route announcement, from router R1 toward router R2, as metric(R1, R2), the invariants that must hold are:

- metric(E-BR-A, ISP-BR-A) < metric(E-BR-B, ISP-BR-A)
- metric(E-BR-B, ISP-BR-B) < metric(E-BR-A, ISP-BR-B)
- metric(ISP-BR-A, E-BR-A) < metric(ISP-BR-A, E-BR-B)
- metric(ISP-BR-B, E-BR-B) < metric(ISP-BR-B, E-BR-A)

Note that smaller metric means stronger route in RIPng.

5. Issues with ingress filters in ISP

If the upstream ISP imposes ingress filters [Ferguson, 1998] to outbound traffic, story becomes much more complex. A packet with source address taken from Pref-A must go out from ISP-BR-A. Similarly, a packet with source address taken from Pref-B must go out from ISP-BR-B. Since none of the routers in the site network will route packets based on source address, packets can easily be routed to incorrect border router.

One possible way is to negotiate with both ISPs, to allow both Pref-B and Pref-A to be used as source address. This approach does not work if upstream ISP of ISP-A imposes ingress filtering. Since there will be multiple levels of ISP on top of ISP-A, it will be hard to understand which upstream ISP imposes the filter. In reality, this problem will be very rare, as ingress filter is not suitable for use in large ISPs where smaller ISPs are connected beneath.

Another possibility is to use source-based routing at E-BR-A and E-BR-B. Here we assume that IPv6-over-IPv6 tunnel is used for secondary links. When an outbound packet arrives to E-BR-A with source address in Pref-B, E-BR-A will forward it to secondary link (tunnel to ISP-BR-B) based on source-based routing decision. The packet will look like this:

- Outer IPv6 header: source = address of E-BR-A in Pref-A, dest = ISP-BR-B

- Inner IPv6 header: source = address in Pref-B, dest = final dest

Tunneled packet will travel across ISP-BR-A toward ISP-BR-B. The packet can go through ingress filter at ISP-BR-A, since it has outer IPv6 source address in Pref-A. Packet will reach ISP-BR-B and decapsulated before ingress filter is applied. Decapsulated packet can go through ingress filter at ISP-BR-B, since it now has source address in Pref-B (from inner IPv6 header). Notice the following facts when configuring this:

- Not every router implements source-based routing.

- The interaction between normal routing and source-based routing at E-BR-A (and/or E-BR-B) varies by router implementations.

- At ISP-BR-B (and/or ISP-BR-A), the interaction between tunnel egress processing and filtering rules varies by router implementations and filter configurations.

6. Observations

The document discussed the cases where a site has two upstream ISPs. The document can easily be extended to the cases where there are 3 or more upstream ISPs.

If you have many upstream providers, you would not make all ISPs backup each other, as it requires O(N^2) tunnels for N ISPs. Rather, it is better to make N/2 pairs of ISPs, and let each pair of ISP backup each other. It is important to pick pairs which are unlikely to be down simultaneously. In this way, number of tunnels will be O(N).

Suppose that the site is very large and it has ISP links in very distant locations, such as in the United States and in Japan. In such case, it is wiser to use this technique only among ISP links in the US, and only among ISP links in Japan. If you use this technique between ISP link A in the US and ISP link B in Japan, the secondary link makes packets travel very long path, for example, from host in the site in the US, to E-BR-B in Japan, to ISP-BR-B (again in Japan), and then to the final destination in the US. This may not make sense for actual use, due to excessive delay.

Similarly, in a large site, addresses must be assigned to end nodes with great care, to minimize delays due to extra path packets may travel. It may be wiser to avoid assigning an address in a prefix assigned from Japanese ISP, to an end node in the US.

If one of primary link is down for a long time, administrators may want to control source address selection on end hosts so that secondary link is less likely to be used. This can be achieved by marking unwanted prefix as deprecated. Suppose the primary link toward ISP-A has been down. You will issue router advertisement [Thomson, 1998; Narten, 1998] packets from routers, with preferred lifetime set to 0 in prefix information option for Pref-A. End hosts will consider addresses in Pref-A as deprecated, and will not use any of them as source address for future connections. If an end host in the site makes new connection to outside, the host will use an address in Pref-B as source address, and reply packet to the end host will travel primary link from ISP-BR-B toward E-BR-B.

Some of non-goals (such as "best" exit link selection) can be achieved by combining technique described in this document, with some other techniques. One example of the technique would be the source/destination address selection heuristics on the end nodes.

7. Security considerations

The configuration described in the document introduces no new security problem.

If primary links toward ISP-A and ISP-B have different security characteristics (like encrypted link and non-encrypted link), administrators needs to be careful setting up secondary links tunneled on them. Packets may travel unwanted path, if secondary links are configured without care.

References

Bates, 1998.
T. Bates and Y. Rekhter, "Scalable Support for Multi-homed Multi-provider Connectivity" in RFC2260 (January 1998). ftp://ftp.isi.edu/in-notes/rfc2260.txt.

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

Durand, 1999.
A. Durand and B. Buclin, "6Bone Routing Practice" in RFC2546 (March 1999). ftp://ftp.isi.edu/in-notes/rfc2546.txt.

Gilligan, 1996.
R. Gilligan and E. Nordmark, "Transition Mechanisms for IPv6 Hosts and Routers" in RFC1933 (April 1996). ftp://ftp.isi.edu/in-notes/rfc1933.txt.

Carpenter, 2000.
Brian Carpenter and Keith Moore, "Connection of IPv6 Domains via IPv4 Clouds without Explicit Tunnels" in draft-ietf-ngtrans-6to4-06.txt (June 2000). work in progress.

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

Ferguson, 1998.
P. Ferguson and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing" in RFC2267 (January 1998). ftp://ftp.isi.edu/in-notes/rfc2267.txt.

Thomson, 1998.
S. Thomson and T. Narten, "IPv6 Stateless Address Autoconfiguration" in RFC2462 (December 1998). ftp://ftp.isi.edu/in-notes/rfc2462.txt.

Narten, 1998.
T. Narten, E. Nordmark, and W. Simpson, "Neighbor Discovery for IP Version 6 (IPv6)" in RFC2461 (December 1998). ftp://ftp.isi.edu/in-notes/rfc2461.txt.

Acknowledgements

The document was made possible by cooperation from people in ipngwg multihoming design team, people in KAME project and George Tsirtsis.

Author's address

Jun-ichiro Hagino
Research Laboratory, Internet Initiative Japan Inc.
Takebashi Yasuda Bldg.,
3-13 Kanda Nishiki-cho,
Chiyoda-ku,Tokyo 101-0054, JAPAN
Tel: +81-3-5259-6350
Fax: +81-3-5259-6351
email: itojun@iijlab.net