Network Working Group
Request for Comments: 3178
Category: Informational

J. Hagino
Research Laboratory, IIJ
H. Snyder
Vail Systems

October 2001

IPv6 Multihoming Support at Site Exit Routers

Status of this Memo

Copyright Notice

Abstract

The document describes a mechanism for basic IPv6 multihoming support, and its operational requirements. Unlike currently-practiced IPv4 multihoming, the technique does not impact the worldwide routing table size, nor IGP (Interior Gateway Protocol) routing table size in upstream ISPs. 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 RFC 2260 by Tony Bates.

1. Problem

Routing table size has been a major issue for both IPv4 and IPv6. As IPv6 addresses are 4 times larger in bit width than IPv4, the routing table size issue would have more serious negative effects on router memory usage, as well as routing table lookup performance. To cope with this problem, the IPv6 addressing architecture [Hinden, 1998] is designed to take advantage of aggregated routing announcements to reduce the number of routes in default-free zone. Also, 6bone operation guideline [Rockell, 2000] (which is the currently-practiced guideline for IPv6 network operation) suggests that ASes not announce non-aggregatable announcements to the default-free zone, if there is no special agreement with the peer.

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

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

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

Since the above two methodologies effectively inject additional routes to the worldwide routing table, they have negative impact on the worldwide routing table size issue. They also are not compatible with current IPv6 operational practice.

This document provides a way to configure site exit routers and ISP routers, so that the site can achieve better reachability from multihomed connectivity, without impacting worldwide routing table size issues. The technique uses multiple distinct IPv6 address prefixes, assigned from multiple upstream ISPs. The technique uses an already-defined routing protocol (BGP or RIPng) and tunneling of IPv6 packets; therefore, this document introduces no new protocol standard (the document describes how to operate the configuration).

This document is largely based on RFC 2260 [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.

- Cope with breakage of any of the upstream ISPs.

3. Basic mechanisms

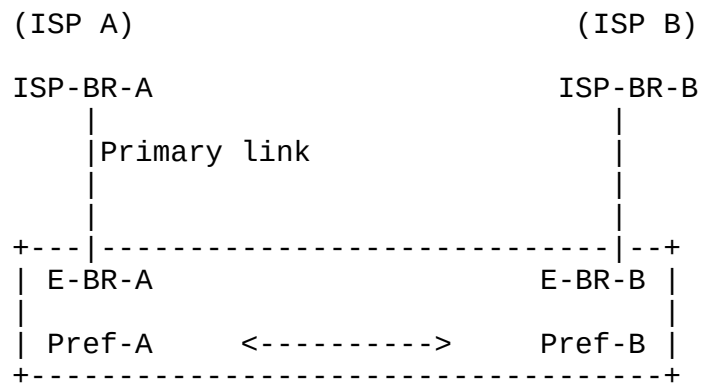
We use the technique described in RFC 2260 section 5.2 in our configuration. To summarize, for IPv4-only networks, RFC 2260 says that:

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

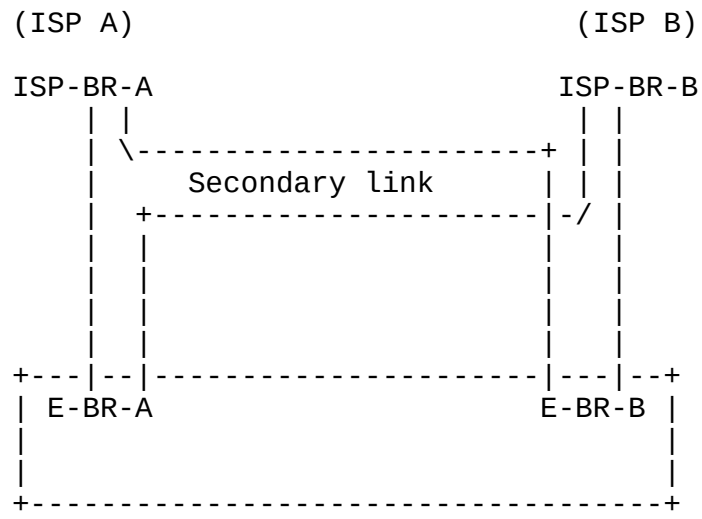
- We are assigned IP address prefixes, Pref-A and Pref-B, from ISP-A and ISP-B respectively. Hosts near ISP-A will get an address from Pref-A, and vice versa.

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

- ISP-A and our site are connected by a "primary link" between ISP router ISP-BR-A and our router E-BR-A. ISP B and our site are connected by a primary link between ISP router ISP-BR-B and our router E-BR-B.



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



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

  Note that we always announce Pref-A to ISP-BR-A, and Pref-B to ISP-BR-B.

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

Under this configuration, both inbound and outbound packets 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

RFC 2260 is written for IPv4 and BGP. With IPv6 and BGP4+, or IPv6 and RIPng, similar results can be achieved, without impacting worldwide IPv6 routing table size.

4.1. IPv6 rule conformance

In RFC 2260, 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 meets the suggestions in 6bone aggregation guidelines [Rockell, 2000]. Also, RFC 2260 does not require portable addresses.

4.2. Address assignment to the nodes

In IPv4, it is usually assumed that a node will be assigned a single IPv4 address. Therefore, RFC 2260 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) addresses from both prefixes, to a single node in the site. This will allow more flexibility in node configuration.

When multiple IPv6 global addresses are assigned to an IPv6 node, source address selection must take place on packet transmissions. Source address selection itself is out of scope of the document. Refer to a separate draft [Draves, 2001] for more discussions.

One simplifying approach is to place the site's Internet hosts on separate subnets, one with addresses in Pref-A and connected to E-BR-A, the other having addresses in Pref-B and connected to E-BR-B. This approach generalizes to having E-BR-A and E-BR-B at different sites, where site A and site B have links to the Internet and to each other.

4.3. Configuration of links

With IPv6, the primary link can be IPv6 native connectivity, RFC 2893 [Gilligan, 2000] 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 RFC 2893 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.

In the figure, ISP-BR-A and E-BR-A are both single points of failure for inbound traffic to Pref-A. This could be remedied by using different routers for primary vs. backup links.

4.4. Using RFC 2260 with IPv6 and BGP4+

The RFC 2260 approach on top of IPv6 will work fine as documented in RFC 2260. There will be no extra twists necessary. Since the multihomed site is not doing transit, variations are possible that do not require it to have a public AS number.

4.5. Using RFC 2260 with IPv6 and RIPng

It is possible to run an RFC 2260-like configuration with RIPng [Malkin, 1997] , with careful control of metric. Routers in the figure need to increase RIPng metric on the secondary link, to make the 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, the 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 the 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

A 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. The packet will reach ISP-BR-B and be 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 ISPs

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 a 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 a very long path, for example, from a 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 the primary links 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 the 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 a new connection to outside, the host will use an address in Pref-B as source address, and the reply packet to the end host will travel the primary link from ISP-BR-B toward E-BR-B. A great care must be taken when you try to automate this by using router renumbering protocols [Crawford, 2000] , as the approach could lead your site into very unstable state if any of the links flap. The author does not recommend to automate it.

Some of non-goals (such as "best" exit link selection) can be achieved by combining the technique described in this document, with some other techniques. One example of the technique would be the source/destination address selection [Draves, 2001] on the end nodes.

7. Operational experiences

Hal Snyder has been running the technique, with two upstream ISPs (lava.net and iijlab), using 2 RFC 2893 IPv6-over-IPv4 tunnels to each of them (in total 4 tunnels), and BGP4+ peering over them.

As expected, when the primary links goes down the routing switches to the secondary link within BGP hold time, i.e., we see approximately the relations:

- (hold time - keepalive time) < failover time
- failover time < hold time
- failback time < keepalive time

This has been tested with keepalive and hold times from as low as 3 and 10 seconds respectively, up to 60 and 180 seconds respectively.

The routing change will affect ISP-BR-A (or B) only. Because route instability is not propagated beyond one ISP, it should be feasible to use lower hold and keepalive times than in a conventional IPv4 setting. If primary and backup links terminate on the same router at the ISP, then failover from primary to backup link need not affect reachability information upstream of that router.

Many of the existing IPv6 networks (connected to worldwide 6bone) are assigned multiple IPv6 prefixes from multiple upstreams. In many cases people assign global IPv6 addresses generated from multiple address prefixes. There has been almost no problems raised about complication due to source address selection.

8. 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 need to be careful setting up secondary links tunneled on them. Packets may travel an unwanted path, if secondary links are configured without care.

References

[Bates, 1998] Bates, T. and Y. Rekhter, "Scalable Support for Multi-homed Multi-provider Connectivity", RFC 2260, January 1998.

[Hinden, 1998] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 2373, July 1998.

[Rockell, 2000] Rockell, R. and B. Fink, "6Bone Backbone Routing Guidelines", RFC 2772, February 2000.

[Draves, 2001] Draves, R., "Default Address Selection for IPv6", Work in Progress.

[Gilligan, 2000] Gilligan, R. and E. Nordmark, "Transition Mechanisms for IPv6 Hosts and Routers", RFC 2893, August 2000.

[Carpenter, 2000] Carpenter, B. and K. Moore, "Connection of IPv6 Domains via IPv4 Clouds", RFC 3056, February 2001.

[Malkin, 1997] Malkin, G. and R. Minnear, "RIPng for IPv6", RFC 2080, January 1997.

[Ferguson, 1998] Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", RFC 2267, January 1998.

[Thomson, 1998] Thomson, S. and T. Narten, "IPv6 Stateless Address Autoconfiguration", RFC 2462, December 1998.

[Narten, 1998] Narten, T., Nordmark, E. and W. Simpson, "Neighbor Discovery for IP Version 6 (IPv6)", RFC 2461, December 1998.

[Crawford, 2000] Crawford, M., "Router Renumbering for IPv6", RFC 2894, August 2000.

Acknowledgements

The document was made possible by cooperation from people participated in JEPG-IP IPv6 multihoming study meeting (1999), people in ipngwg multihoming design team, people in WIDE/KAME project and George Tsirtsis.

Authors' Addresses

Jun-ichiro itojun Hagino
Research Laboratory, Internet Initiative Japan Inc.
Takebashi Yasuda Bldg.,
3-13 Kanda Nishiki-cho,
Chiyoda-ku, Tokyo 101-0054, JAPAN

Phone: +81-3-5259-6350
Fax: +81-3-5259-6351
EMail: itojun@iijlab.net

Hal Snyder
Vail Systems, Inc.
570 Lake Cook Rd, Ste 408
Deerfield, IL 60015, US

Phone: +1-312-360-8245
EMail: hal@vailsys.com

Full Copyright Statement

Acknowledgement

Funding for the RFC Editor function is currently provided by the Internet Society.