

Scalable Support for Multi-homed Multi-provider Connectivity

Status of this Memo

This memo provides information for the Internet community. It does not specify an Internet standard of any kind. Distribution of this memo is unlimited.

Copyright Notice

Copyright (C) The Internet Society (1998). All Rights Reserved.

2. Abstract

This document describes addressing and routing strategies for multi-homed enterprises attached to multiple Internet Service Providers (ISPs) that are intended to reduce the routing overhead due to these enterprises in the global Internet routing system.

3. Motivations

An enterprise may acquire its Internet connectivity from more than one Internet Service Provider (ISP) for some of the following reasons. Maintaining connectivity via more than one ISP could be viewed as a way to make connectivity to the Internet more reliable. This way when connectivity through one of the ISPs fails, connectivity via the other ISP(s) would enable the enterprise to preserve its connectivity to the Internet. In addition to providing more reliable connectivity, maintaining connectivity via more than one ISP could also allow the enterprise to distribute load among multiple connections. For enterprises that span wide geographical area this could also enable better (more optimal) routing.

The above considerations, combined with the decreasing prices for the Internet connectivity, motivate more and more enterprises to become multi-homed to multiple ISPs. At the same time, the routing overhead that such enterprises impose on the Internet routing system becomes more and more significant. Scaling the Internet, and being able to support a growing number of such enterprises demands mechanism(s) to contain this overhead. This document assumes that an approach where routers in the "default-free" zone of the Internet would be required

to maintain a route for every multi-homed enterprise that is connected to multiple ISPs does not provide an adequate scaling. Moreover, given the nature of the Internet, this document assumes that any approach to handle routing for such enterprises should minimize the amount of coordination among ISPs, and especially the ISPs that are not directly connected to these enterprises.

There is a difference of opinions on whether the driving factors behind multi-homing to multiple ISPs could be adequately addressed by multi-homing just to a single ISP, which would in turn eliminate the negative impact of multi-homing on the Internet routing system. Discussion of this topic is beyond the scope of this document.

The focus of this document is on the routing and addressing strategies that could reduce the routing overhead due to multi-homed enterprises connected to multiple ISPs in the Internet routing system.

The strategies described in this document are equally applicable to both IPv4 and IPv6.

4. Address allocation and assignment

A multi-homed enterprise connected to a set of ISPs would be allocated a block of addresses (address prefix) by each of these ISPs (an enterprise connected to N ISPs would get N different blocks). The address allocation from the ISPs to the enterprise would be based on the "address-lending" policy [RFC2008]. The allocated addresses then would be used for address assignment within the enterprise.

One possible address assignment plan that the enterprise could employ is to use the topological proximity of a node (host) to a particular ISP (to the interconnect between the enterprise and the ISP) as a criteria for selecting which of the address prefixes to use for address assignment to the node. A particular node (host) may be assigned address(es) out of a single prefix, or may have addresses from different prefixes.

5. Routing information exchange

The issue of routing information exchange between an enterprise and its ISPs is decomposed into the following components:

- a) reachability information that an enterprise border router advertises to a border router within an ISP
- b) reachability information that a border router within an ISP advertises to an enterprise border router

The primary focus of this document is on (a); (b) is covered only as needed by this document.

5.1. Advertising reachability information by enterprise border routers

When an enterprise border router connected to a particular ISP determines that the connectivity between the enterprise and the Internet is up through all of its ISPs, the router advertises (to the border router of that ISP) reachability to only the address prefix that the ISP allocated to the enterprise. This way in a steady state routes injected by the enterprise into its ISPs are aggregated by these ISPs, and are not propagated into the "default-free" zone of the Internet.

When an enterprise border router connected to a particular ISP determines that the connectivity between the enterprise and the Internet through one or more of its other ISPs is down, the router starts advertising reachability to the address prefixes that was allocated by these ISPs to the enterprise. This would result in injecting additional routing information into the "default-free" zone of the Internet. However, one could observe that the probability of all multi-homed enterprises in the Internet concurrently losing connectivity to the Internet through one or more of their ISPs is fairly small. Thus on average the number of additional routes in the "default-free" zone of the Internet due to multi-homed enterprises is expected to be a small fraction of the total number of such enterprises.

The approach described above is predicated on the assumption that an enterprise border router has a mechanism(s) by which it could determine (a) whether the connectivity to the Internet through some other border router of that enterprise is up or down, and (b) the address prefix that was allocated to the enterprise by the ISP connected to the other border router. One such possible mechanism could be provided by BGP [RFC1771]. In this case border routers within the enterprise would have an IBGP peering with each other. Whenever one border router determines that the intersection between the set of reachable destinations it receives via its EBGP (from its directly connected ISP) peerings and the set of reachable destinations it receives from another border router (in the same enterprise) via IBGP is empty, the border router would start advertising to its external peer reachability to the address prefix that was allocated to the enterprise by the ISP connected to the other border router. The other border router would advertise (via IBGP) the address prefix that was allocated to the enterprise by the ISP connected to that router. This approach is known as "auto route injection".

As an illustration consider an enterprise connected to two ISPs, ISP-A and ISP-B. Denote the enterprise border router that connects the enterprise to ISP-A as BR-A; denote the enterprise border router that connects the enterprise to ISP-B as BR-B. Denote the address prefix that ISP-A allocated to the enterprise as Pref-A; denote the address prefix that ISP-B allocated to the enterprise as Pref-B. When the set of routes BR-A receives from ISP-A (via EBGp) has a non-empty intersection with the set of routes BR-A receives from BR-B (via IBGP), BR-A advertises to ISP-A only the reachability to Pref-A. When the intersection becomes empty, BR-A would advertise to ISP-A reachability to both Pref-A and Pref-B. This would continue for as long as the intersection remains empty. Once the intersection becomes non-empty, BR-A would stop advertising reachability to Pref-B to ISP-A (but would still continue to advertise reachability to Pref-A to ISP-A). Figure 1 below describes this method graphically.

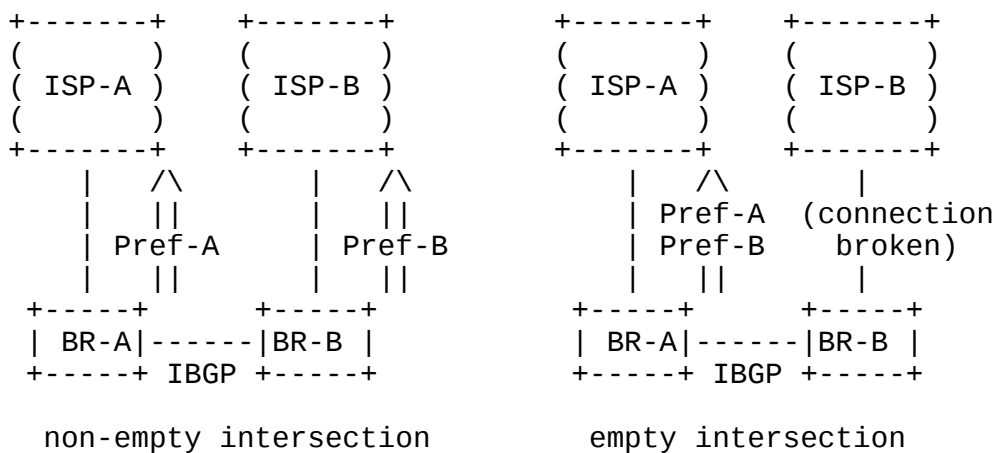


Figure 1: Reachability information advertised

Although strictly an implementation detail, calculating the intersection could potentially be a costly operation for a large set of routes. An alternate solution to this is to make use of a selected single (or more) address prefix received from an ISP (the ISP's backbone route for example) and configure the enterprise border router to perform auto route injection if the selected prefix is not present via IBGP. Let's suppose ISP-B has a well known address prefix, ISP-Pref-B for its backbone. ISP-B advertises this to BR-B and BR-B in turn advertises this via IBGP to BR-A. If BR-A sees a withdraw for ISP-Pref-B it advertises Pref-B to ISP-A.

The approach described in this section may produce less than the full Internet-wide connectivity in the presence of ISPs that filter out routes based on the length of their address prefixes. One could observe however, that this would be a problem regardless of how the enterprise would set up its routing and addressing.

5.2. Further improvements

The approach described in the previous section allows to significantly reduce the routing overhead in the "default-free" zone of the Internet due to multi-homed enterprises. The approach described in this section allows to completely eliminate this overhead.

An enterprise border router would maintain EBGP peering not just with the directly connected border router of an ISP, but with the border router(s) in one or more ISPs that have their border routers directly connected to the other border routers within the enterprise. We refer to such peering as "non-direct" EBGP.

An ISP that maintains both direct and non-direct EBGP peering with a particular enterprise would advertise the same set of routes over both of these peerings. An enterprise border router that maintains either direct or non-direct peering with an ISP advertises to that ISP reachability to the address prefix that was allocated by that ISP to the enterprise. Within the ISP routes received over direct peering should be preferred over routes received over non-direct peering. Likewise, within the enterprise routes received over direct peering should be preferred over routes received over non-direct peering.

Forwarding along a route received over non-direct peering should be accomplished via encapsulation [RFC1773].

As an illustration consider an enterprise connected to two ISPs, ISP-A and ISP-B. Denote the enterprise border router that connects the enterprise to ISP-A as E-BR-A, and the ISP-A border router that is connected to E-BR-A as ISP-BR-A; denote the enterprise border router that connects the enterprise to ISP-B as E-BR-B, and the ISP-B border router that is connected to E-BR-B as ISP-BR-B. Denote the address prefix that ISP-A allocated to the enterprise as Pref-A; denote the address prefix that ISP-B allocated to the enterprise as Pref-B. E-BR-A maintains direct EBGP peering with ISP-BR-A and advertises reachability to Pref-A over that peering. E-BR-A also maintain a non-direct EBGP peering with ISP-BR-B and advertises reachability to Pref-B over that peering. E-BR-B maintains direct EBGP peering with ISP-BR-B, and advertises reachability to Pref-B over that peering. E-BR-B also maintains a non-direct EBGP peering

with ISP-BR-A, and advertises reachability to Pref-A over that peering.

When connectivity between the enterprise and both of its ISPs (ISP-A and ISP-B) is up, traffic destined to hosts whose addresses were assigned out of Pref-A would flow through ISP-A to ISP-BR-A to E-BR-A, and then into the enterprise. Likewise, traffic destined to hosts whose addresses were assigned out of Pref-B would flow through ISP-B to ISP-BR-B to E-BR-B, and then into the enterprise. Now consider what would happen when connectivity between ISP-BR-B and E-BR-B goes down. In this case traffic to hosts whose addresses were assigned out of Pref-A would be handled as before. But traffic to hosts whose addresses were assigned out of Pref-B would flow through ISP-B to ISP-BR-B, ISP-BR-B would encapsulate this traffic and send it to E-BR-A, where the traffic will get decapsulated and then be sent into the enterprise. Figure 2 below describes this approach graphically.

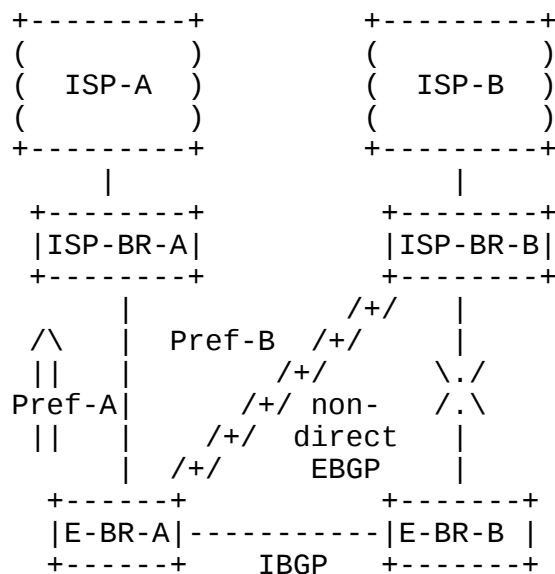


Figure 2: Reachability information advertised via non-direct EBGP

Observe that with this scheme there is no additional routing information due to multi-homed enterprises that has to be carried in the "default-free" zone of the Internet. In addition this scheme doesn't degrade in the presence of ISPs that filter out routes based on the length of their address prefixes.

Note that the set of routers within an ISP that maintain non-direct peering with the border routers within an enterprise doesn't have to be restricted to the ISP's border routers that have direct peering

with the enterprise's border routers. The non-direct peering could be maintained with any router within the ISP. Doing this could improve the overall robustness in the presence of failures within the ISP.

5.3. Combining the two

One could observe that while the approach described in Section 5.2 allows to completely eliminate the routing overhead due to multi-homed enterprises in the "default-free" zone of the Internet, it may result in a suboptimal routing in the presence of link failures. The sub-optimality could be reduced by combining the approach described in Section 5.2 with a slightly modified version of the approach described in Section 5.1. The modification consists of constraining the scope of propagation of additional routes that are advertised by an enterprise border router when the router detects problems with the Internet connectivity through its other border routers. A way to constrain the scope is by using the BGP Community attribute [RFC1997].

5.4. Better (more optimal) routing in steady state

The approach described in this document assumes that in a steady state an enterprise border router would advertise to a directly connected ISP border router only the reachability to the address prefix that this ISP allocated to the enterprise. As a result, traffic originated by other enterprises connected to that ISP and destined to the parts of the enterprise numbered out of other address prefixes would not enter the enterprise at this border router, resulting in potentially suboptimal paths. To improve the situation the border router may (in steady state) advertise reachability not only to the address prefix that was allocated by the ISP that the router is directly connected to, but to the address prefixes allocated by some other ISPs (directly connected to some other border routers within the enterprise). Distribution of such advertisements should be carefully constrained, or otherwise this may result in significant additional routing information that would need to be maintained in the "default-free" part of the Internet. A way to constrain the distribution of such advertisements is by using the BGP Community attribute [RFC1997].

6. Comparison with other approaches

CIDR [RFC1518] proposes several possible address allocation strategies for multi-homed enterprises that are connected to multiple ISPs. The following briefly reviews the alternatives being used today, and compares them with the approaches described above.

6.1. Solution 1

One possible solution suggested in [RFC1518] is for each multi-homed enterprise to obtain its IP address space independently from the ISPs to which it is attached. This allows each multi-homed enterprise to base its IP assignments on a single prefix, and to thereby summarize the set of all IP addresses reachable within that enterprise via a single prefix. The disadvantage of this approach is that since the IP address for that enterprise has no relationship to the addresses of any particular ISPs, the reachability information advertised by the enterprise is not aggregatable with any, but default route. results in the routing overhead in the "default-free" zone of the Internet of $O(N)$, where N is the total number of multi-homed enterprises across the whole Internet that are connected to multiple ISPs.

As a result, this approach can't be viewed as a viable alternative for all, but the enterprises that provide high enough degree of addressing information aggregation. Since by definition the number of such enterprises is likely to be fairly small, this approach isn't viable for most of the multi-homed enterprises connected to multiple ISPs.

6.2. Solution 2

Another possible solution suggested in [RFC1518] is to assign each multi-homed enterprise a single address prefix, based on one of its connections to one of its ISPs. Other ISPs to which the multi-homed enterprise is attached maintain a routing table entry for the organization, but are extremely selective in terms of which other ISPs are told of this route and would need to perform "proxy" aggregation. Most of the complexity associated with this approach is due to the need to perform "proxy" aggregation, which in turn requires additional inter-ISP coordination and more complex router configuration.

7. Discussion

The approach described in this document assumes that addresses that an enterprise would use are allocated based on the "address lending" policy. Consequently, whenever an enterprise changes its ISP, the enterprise would need to renumber part of its network that was numbered out of the address block that the ISP allocated to the enterprise. However, these issues are not specific to multihoming and should be considered accepted practice in today's Internet. The approach described in this document effectively eliminates any distinction between single-home and multi-homed enterprise with respect to the impact of changing ISPs on renumbering.

The approach described in this document also requires careful address assignment within an enterprise, as address assignment impacts traffic distribution among multiple connections between an enterprise and its ISPs.

Both the issue of address assignment and renumbering could be addressed by the appropriate use of network address translation (NAT). The use of NAT for multi-homed enterprises is beyond the scope of this document.

Use of auto route injection (as described in Section 5.1) increases the number of routers in the default-free zone of the Internet that could be affected by changes in the connectivity of multi-homed enterprises, as compared to the use of provider-independent addresses (as described in Section 6.1). Specifically, with auto route injection when a multi-homed enterprise loses its connectivity through one of its ISPs, the auto injected route has to be propagated to all the routers in the default-free zone of the Internet. In contrast, when an enterprise uses provider-independent addresses, only some (but not all) of the routers in the default-free zone would see changes in routing when the enterprise loses its connectivity through one of its ISPs.

To suppress excessive routing load due to link flapping the auto injected route has to be advertised until the connectivity via the other connection (that was previously down and that triggered auto route injection) becomes stable.

Use of the non-direct EBGp approach (as described in Section 5.2) allows to eliminate route flapping due to multi-homed enterprises in the default-free zone of the Internet. That is the non-direct EBGp approach has better properties with respect to routing stability than the use of provider-independent addresses (as described in Section 6.1).

8. Applications to multi-homed ISPs

The approach described in this document could be applicable to a small to medium size ISP that is connected to several upstream ISPs. The ISP would acquire blocks of addresses (address prefixes) from its upstream ISPs, and would use these addresses for allocations to its customers. Either auto route injection, or the non-direct EBGp approach, or a combination of both could be used by the ISP when peering with its upstream ISPs. Doing this would provide routability for the customers of such ISP, without adversely affecting the overall scalability of the Internet routing system.

9. Security Considerations

Since the non-direct EBGp approach (as described in Section 5.2) requires EBGp sessions between routers that are more than one IP hop from each other, routers that maintain these sessions should use an appropriate authentication mechanism(s) for BGP peer authentication.

Security issues related to the IBGP peering, as well as the EBGp peering between routers that are one IP hop from each other are outside the scope of this document.

10. Acknowledgments

The authors of this document do not make any claims on the originality of the ideas described in this document. Anyone who thought about these ideas before should be given all due credit.

11. References

[RFC1518]

Rekhter, Y., and T. Li, "An Architecture for IP Address Allocation with CIDR", RFC 1518, September 1993.

[RFC1771]

Rekhter, Y., and T. Li, "A Border Gateway Protocol 4 (BGP-4)", RFC 1771, March 1995.

[RFC1773]

Hanks, S., Li, T., Farinacci, T., and P. Traina, "Generic Routing Encapsulation over IPv4 networks", RFC 1773, October 1994.

[RFC1918]

Rekhter, Y., Moskowitz, B., Karrenberg, D., de Groot G.J., and E. Lear, "Address Allocation for Private Internets", RFC 1918, February 1996.

[RFC1997]

Chandra, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, August 1996.

[RFC2008]

Rekhter, Y., and T. Li, "Implications of Various Address Allocation Policies for Internet Routing", BCP 7, RFC 2008, October 1996.

12. Authors' Addresses

Tony Bates
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134

E-Mail: tbates@cisco.com

Yakov Rekhter
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134
E-Mail: yakov@cisco.com

13. Full Copyright Statement

Copyright (C) The Internet Society (1998). All Rights Reserved.

This document and translations of it may be copied and furnished to others, and derivative works that comment on or otherwise explain it or assist in its implementation may be prepared, copied, published and distributed, in whole or in part, without restriction of any kind, provided that the above copyright notice and this paragraph are included on all such copies and derivative works. However, this document itself may not be modified in any way, such as by removing the copyright notice or references to the Internet Society or other Internet organizations, except as needed for the purpose of developing Internet standards in which case the procedures for copyrights defined in the Internet Standards process must be followed, or as required to translate it into languages other than English.

The limited permissions granted above are perpetual and will not be revoked by the Internet Society or its successors or assigns.

This document and the information contained herein is provided on an "AS IS" basis and THE INTERNET SOCIETY AND THE INTERNET ENGINEERING TASK FORCE DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

