Network Working Group
Internet-Draft
Expires: April 5, 2006

M. Bagnulo
UC3M

October 2, 2005

Address selection in multihomed environments
draft-bagnulo-shim6-addr-selection-00

Status of this Memo

Copyright Notice

Abstract

This note describes mechanisms for address selection in hosts within a multihomed site. The goal of these mechanisms is to allow hosts within the multihomed site to establish new communications after an outage. The presented mechanisms are not by themselves a complete multihoming solution but rather a component of such solution.

Table of Contents

1. Introduction 3
2. Reference topology 4
3. The problem: address selection after failures 5
4. Goals and non-goals 7
4.1. Goal 7
4.2. Non goals 7
5. Proposed mechanisms 8
5.1. Proactive mechanisms 8
5.1.1. Direct link failures 8
5.1.2. Other failure modes 9
5.2. Reactive mechanisms 11
6. Future steps 14
7. Acknowledgments 15
8. References 15
Author's Address 16
Intellectual Property and Copyright Statements 17

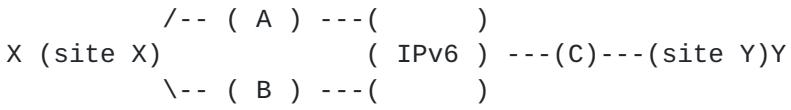
1. Introduction

A way to solve the issue of site multihoming is to have a separate site prefix for each connection of the site, and to derive as many addresses for each hosts. This approach to multi-homing has the advantage of minimal impact on the inter-domain routing fabric, since each site prefix can be aggregated within the larger prefix of a specific provider; however, it opens a number of issues, that have to be addressed in order to provide a multihoming solution compatible with such addressing scheme.

In this memo we will present a set of mechanisms that enable the hosts within the multihomed site to select the addresses in order to be able to establish new communications after an outage. The presented mechanisms are not by themselves a complete multihoming solution but rather a component of such solution.

2. Reference topology

In the following discussion, we will use this reference topology:



The topology features two hosts, X and Y. The site of X is multihomed while the site of Y is single homed. Host X has to global IPv6 addresses, which we will note "A:X" and "B:X", formed by combined the prefixes allocated by ISP A and B to "site X" with the host identifier of X. Y has only one address "C:Y".

We assume that Y, when it starts engaging communication with X, has learned the addresses A:X and B:X, for example because they were published in the DNS. We do not assume that the DNS is dynamic: there will be situations in which both A:X and B:X are published, while in fact only one is reachable. We assume that X, when it receives packets from Y, has only access to information contained in the packet coming from Y, e.g. the source address; we do not assume that X can retrieve by external means the set of addresses associated to Y. similar assumptions are made when X is initiating the communication, only that in this case, a single address i.e. C:Y is published in the DNS

In this scenario, both ISPA and ISPB are performing ingress filtering and have not relaxed the source address checks. So, we assume that an ingress filtering compatibility mechanism [7] is available in the multihomed site (Site X) so that packets are forwarded through the ISP that corresponds to the source address prefix included in the packet by the host.

3. The problem: address selection after failures

In case that a failure occurs in one of the ISPs of the multihomed site, it may not be possible to establish a new communication towards a destination outside the site using the addresses derived from the prefix of the ISP affected by the failure. For instance, in the case that the link between ISPA and the Internet fails, it will not be possible to establish a communication between X and Y using address A:X. In this case, any communication involving this address will fail because:

- If Y tries to establish a communication with X using A:X as a destination address, packets would be discarded because there is no path available from the Internet to ISPA.

- If X tries to communicate with Y using A:X as a source address, packets will be routed through ISPA in order to comply with ingress filters, and because ISPA has no link available wit the rest of the Internet, the packet will be discarded (it should be noted that even if the packet could make it to Y, reply packets from Y to X would contain A:X as a destination address, which is unreachable from Y).

So, in order to establish a communication between X and Y when a failure has occurred in ISPA, the address derived from ISPA block i.e. A:X, must not be used for the communication.

The solution for this problem has to be provided by the address selection mechanisms. In particular, when the communication is established from the host Y to the host X, the solution has to be provided by the destination address selection mechanism at host Y and when the communication is established from the host X to the host Y, the solution has to be provided by the source address selection mechanism at host X. Default address selection for IPv6 hosts is specified in RFC 3484 [5]

We will next analyze the support provided by RFC 3484 when the communication is established from host Y to host X. In this case, host Y has two possible destination addresses A:X and B:X. Without any additional knowledge, both addresses are equivalent to host Y, so the default destination address selection mechanism will return a list of the two addresses ordered as they were returned by the resolver. It may occur that A:X is first. In this case, host Y will use A:X to reach host X and it will fail. At this point, RFC 3484 states that if there are other destination addresses available, the application should retry to establish the communication, using the next address in the list. If the application retries with address B:X, the communication will be established successfully.

In conclusion, the current destination address selection mechanism is enough to deal with this situation (as long as applications retry with all the addresses).

Next, we will analyze the support provided by RFC 3484 when the communication is established from host X to host Y. In this case, destination address selection performed in host X is trivial, since there is only one address available for Y (C:Y). Source address selection mechanism as specified in RFC 3484 will not prefer any of the two source addresses if no additional information is available, so any of the addresses can be selected as source address. In the case that address A:X is selected, the communication will fail. In this case there are no alternative destination address to retry with, so the communication will definitely fail.

In conclusion, the source address selection mechanism defined in RFC 3484 is not enough to support this scenario. This memo defines mechanisms to provide a solution for this case.

4. Goals and non-goals

4.1. Goal

The goal of this memo is to provide mechanisms to complement the source address selection of the host within the multihomed that allow it to succesfully establish new communications after an outage, without requiring any modification to the hosts outside the multihomed site.

4.2. Non goals

It is not a goal of this memo to modify hosts located outside the multihomed site.

It is not the goal of this memo to define a complete multihoming solution, but rather to define a component of such solution.

5. Proposed mechanisms

There are two types of mechanisms that can be used to enable the host within the multihomed site to select the appropriate source address: proactive mechanisms, where the host is notified about which source address prefixes should be used for the different destinations and reactive mechanisms, which are based on the trial and error approach, where the hosts try with different source address prefixes and detect which one is available.

5.1. Proactive mechanisms

In this case, two mechanisms are needed: first, a mechanism to detect the outage and then a mechanisms to inform the host about which prefixes should be used in the source address for the different destinations. While this may seem a lot of information to be stored in the host, it should be noted that in the default case, when no outage has occurred, all prefixes can be used to reach all the destinations, so no information is needed. When an outage occurs, the host must be informed of which source address prefix should not be used to reach a certain set of destinations. Depending on the type of outage, the amount of information may vary.

We will first present mechanisms that can be used when the outage occurs in the direct link between the multihomed site and the ISP and then we will present a more general approach.

5.1.1. Direct link failures

In the case that the failure has occurred affecting the direct link between the multihomed site and one of its ISP (let's say ISPA), the prefix associated with that ISP should not be used for any new external communication, since the prefix is unreachable from any other part of the Internet.

There are several mechanisms that can be used to detect the outage. For instance, if any routing protocol is used between the ISP and the multihomed site, such protocol can be used to detect the outage. In any case, it is possible to use a periodic ICMP echo request for detecting the outage on direct links.

In addition, we must establish a communication channel that quickly signals these failures to the hosts. The logical channel to use is the "router advertisement" message, which the routers use to communicate to hosts the list of available prefixes. We propose here to use the "preferred" and "valid" flags in these prefixes to signal to hosts the addresses that should, or should not, be used as source address at any given time.

The router advertisement messages are defined in [1] ; their use for address configuration is defined in [2] . As specified in [1] , the router advertisements include a variable number of Prefix Information parameters. Each Prefix Information parameter specifies:

* an address prefix value,
* an "on-link" flag, used in neighbor discovery,
* an "autonomous" flag, used for autonomous address configuration,
* the "valid" lifetime,
* the "preferred" lifetime.

We propose to use the "preferred" lifetime to indicate whether addresses derived from the prefix can be used as source address in multihomed networks. When a prefix is temporarily not available routers MUST advertise a null preferred lifetime for that prefix.

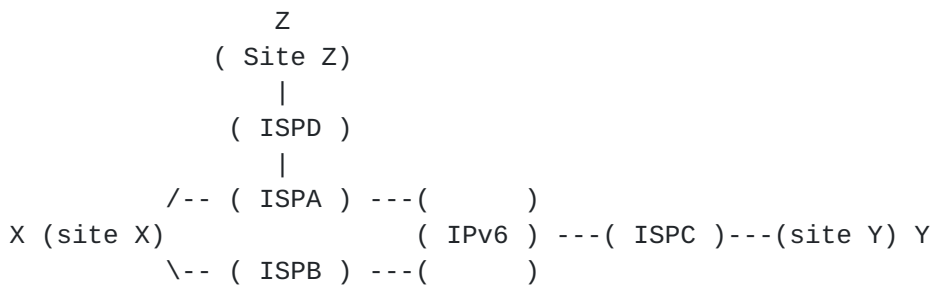
In conformance with section 5.5.4 of [1] , the hosts will notice that the formerly preferred address becomes deprecated when its preferred lifetime expires. They will not use the deprecated addresses in new communications if an alternate (non-deprecated) address is available and has sufficient scope.

This solution is sufficient when the site is composed of a single link; for more complex sites with multiple subnets, we need a mechanism with a broader scope than Router Advertisement. There are two available candidates that provide the required fucntionality: the Router Renumbering protocol [3] and the prefix delegation option defined for DHCP [8].

In order to advertise a null preferred lifetime for a specific prefix, the sites routers must be able to learn about that prefix. Any of the two proposed protocols, Router Renumbering or DHCP Prefix Delegation can be used to pass this information. These protocols allow an authorized agent, in that case the egress site, to update the list of prefixes advertised by the site's routers. The protocols can be used to change parameters associated to a prefix, such as the preferred lifetime.

5.1.2. Other failure modes

There are other failures modes where there are some parts of the Internet reachable using one prefix and other parts of the Internet are reachable using a different prefix. For instance, in the next figure, when a failure occurs in the link between ISPA and the rest of the Internet, host X should use address B:X to reach host C:Y, but host X should use address A:X to reach D:Z.



In this scenario, ISPD has its own prefix Prefix D and it obtains Internet connectivity through ISPA. So, in this case, deprecating the prefix associated with ISPA would preclude communication with Site Z.

In this scenario a mechanism like NAROS [6] can be used. In this mechanism, a server acquires the reachability information available in the inter-domain routing system using BGP. This means that there are BGP sessions established between each site exit router and the border router of the correspondent ISP, through which the site exit router obtains interdomain routing information. The server establishes IBGP sessions with each site exit router, so it discovers which destinations are reachable through each ISP. In the case there is no outage, all destinations are reachable through all the ISPs, so no information has to be propagated to the hosts. In case of a failure, a set of destinations will become unreachable through one (o more) ISP(s). In this case, the server has to inform the hosts about which prefix to use for the different destinations. The host can store this information in the policy table defined in RFC 3484. This policy table allows the host to prefer a certain source address for a given set of destinations.

The missing part is a mechanism to convey the information from the server to the host. A possible option is to define a new DHCP option to transmit policy table information.

In the example above, the mechanism would work as follows. Before the outage, the site exit router of site X are obtaining a default route from both ISPs. The hosts have the default policy table configured.

When an outage occurs in the link between ISPA and the Internet, ISPB still announces a default route, while ISPA will announce only a route to prefix A and to prefix D (and not a default route). This information is transmitted to the server that will craft the new policy table. Because of the longest prefix match rule of source address selection algorithm defined in RFC 3484, A:X will be the preferred source address when sending packet to destinations

containing prefix A. So, the new policy table must inform that for destinations containing prefix D the prefix A should be used in the source address. Additionally the policy table must reflect that for the rest of destinations prefix B should be preferred. This is achieved by adding the following entries to the policy table:

Prefix	Precedence	Label
Prefix A	10	11
Prefix D	10	11
::/0	10	12
Prefix B	10	12

Because both entries have the same Label, Rule 6 (Prefer matching label) of the source address selection algorithm will select the source address containing prefix A when sending packets to destinations containing Prefix D. Also because rule 6, for the rest of the destinations, the source address containing Prefix B will be preferred. The new policy table is transmitted to the hosts using the new DHCP option.

It is also possible to enable the multihomed host to query the server to obtain the source address prefix information as proposed in [6] . In this case, each time the multihomed node initiates a communication with a new destiantion, it would query the server for the proper prefix to include in the source address. The server, would reply based on the routing information it has gathered through BGP.

5.2. Reactive mechanisms

In this approach, the host will try with different source addresses until the communication is established. After the communication is established, the information about which source address to use for a given destination is stored in a Source Address Cache. Each entry of the cache contains for a Destination Address, information about the corresponding Source Address and its lifetime.

The source address cache entry creation process is the following:

When the host receives a packet containing a Source Address SA and a Destination Address DA, the Exit Path Cache is searched for an entry that contains SA Destination Address field.

- If such entry is found, the Source Address is verified.

-- If the Source Address contains DA, then the lifetime of the entry is extended.

-- If the Source Address is other than DA, then the cache entry is updated and DA is stored in the Source Address field. The lifetime of the entry is extended.

- If the entry is not found, the entry is created, with SA as the Destination Address and DA as the Source Address. The entry is blocked from changes for a period of Tb (TBD) seconds to avoid that multiple answers produce suboptimal behavior.

There are different retrial strategies that can be used in this approach.

The first option is to simply let the upper layers to handle retrials. In this case, the IP layer only has to make sure that a different source address is used in each retrial as long as there is not a preferred source address in the source address cache. So, in this approach, when the IP layer receives a packet, it searches the Source Address Cache for a preferred source address associated to the selected destination. If no entry exists for that destination address, one of the available source addresses is selected randomly. If reply packets arrive, an entry will be created in the Source Address Cache for that destination address. If no reply packets arrive, no entry will be created, so when the upper layer protocol resends the packet, a new source address will be used.

The second option would be that the IP layer handles retrials. In this case, the IP layer stores the packet and retries until a Source Address Cache entry is created for that destination. This approach imposes that the IP layer has to store the packets sent to destinations that still don't have a Source Address Cache entry created. It should be noted that the IP layer is incapable of recognizing if incoming packets are actually replies to the packets sent, since the IP layer knowledge is limited to the source and destination address pair. This means that the retransmission service provided by the IP layer won't be reliable. Additionally, this approach requires the definition of timeouts after which the IP layer will resend the packets. Since the IP layer has no information about the round trip times or reply time of the involved applications, the selection of this timer will be arbitrary.

The third option would be to try with all possible source address simultaneously. In this approach, when the IP layer receives a packet from the upper layer, it searches the Source Address Cache for the destination of the packet. If an entry is found for that destination, the source address contained in the Source Address Cache is used. If no entry is found for that destination, the IP layer sends as many packets as different source address are available, each one with a different address. In this case, the IP layer doesn't

need to store any packets and it doesn't rely retransmission by upper layer protocols. What is more, this approach would provide the selection of the "best path" (under certain criteria), since the destination address of the first reply packet could be used as source address, selecting the fastest path available. The main drawback of this approach is the additional load imposed by duplicated packets.

6. Future steps

This memo presents multiple possible approaches to select address for initiating new communications after an outage in multihomed environments. At this point, the goal of the memo is to foster discussion about the benefits and drawbacks of each approach, so that eventually a set of mechanisms can be selected.

7. Acknowledgments

This memo contains parts of a previous work entitled "Host-Centric IPv6 Multihoming" that benefited from comments from Alberto Garcia Martinez, Cedric de Launois, Brian Carpenter, Dave Crocker, Xiaowei Yang and Erik Nordmark.

8. References

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

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

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

[4] Abley, J., Black, B., and V. Gill, "Goals for IPv6 Site-Multihoming Architectures", RFC 3582, August 2003.

[5] Draves, R., "Default Address Selection for Internet Protocol version 6 (IPv6)", RFC 3484, February 2003.

[6] de Launois, C. and O. Bonaventure, "NAROS : Host-Centric IPv6 Multihoming with Traffic Engineering", ID draft-de-launois-multi6-naros-00.txt, May 2003.

[7] Huitema, C. and R. Draves, "Ingress Filtering compatibility for IPv6 multihomed sites", ID draft-huitema-shim6-ingress-filtering-00.txt, October 2005.

[8] Troan, O. and R. Droms, "IPv6 Prefix Options for Dynamic Host Configuration Protocol (DHCP) version 6", RFC 3633, December 2003.

Author's Address

Marcelo Bagnulo
Universidad Carlos III de Madrid
Av. Universidad 30
Leganes, Madrid 28911
SPAIN

Phone: 34 91 6249500
Email: marcelo@it.uc3m.es
URI: http://www.it.uc3m.es

Intellectual Property Statement

Disclaimer of Validity

Copyright Statement

Acknowledgment

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