Network Working Group
Internet-Draft
Expires: August 10, 2004

C. Huitema
R. Draves
Microsoft
M. Bagnulo
UC3M

February 10, 2004

Host-Centric IPv6 Multihoming
draft-huitema-multi6-hosts-03

Status of this Memo

Copyright Notice

Abstract

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, which we call Host-Centric IPv6 Multihoming, 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, which we will discuss in this memo, including the problem created by the interaction between ingress filtering and multihoming. We then propose a set of specific solutions that enable host centric multihoming, and we review how

these solutions meet the goals of IPv6 site multihoming.

Table of Contents

1. Introduction

There are two basic forms of multihoming, multiple interfaces per host and multiple site connections shared by many hosts. This working group is specifically concerned with site multi-homing. 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 can in fact be supported by a combination of router renumbering (RFC2894) and Stateless Address Autoconfiguration (RFC2462). This approach to multi-homing, which we call Host-Centric IPv6 Multihoming, 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, which we will discuss in this memo, including the problem created by the interaction between ingress filtering and multihoming. We then propose a set of specific solutions that enable host centric multihoming, and we review how these solutions meet the goals of IPv6 site multihoming.

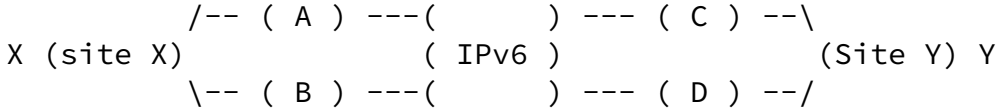
2. Notations

2.1 Requirements language

In this document, the key words "MAY", "MUST", "MUST NOT", "optional", "recommended", "SHOULD", and "SHOULD NOT", are to be interpreted as described in [13].

2.2 Reference topology

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



The topology features two hosts, X and Y, whose respective sites are both multi-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. Similarly, Y

has two addresses "C:Y" and "D:Y".

We assume that X, when it starts engaging communication with Y, has learned the addresses C:Y and D:Y, 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 C:Y and D:Y are published, while in fact only one is reachable. We assume that Y, when it receives packets from X, has only access to information contained in the packet coming from X, e.g. the source address; we do not assume that Y can retrieve by external means the set of addresses associated to X.

2.3 Site exit router

A site exit router is an IPv6 router managing at least one of the connections between a site and the Internet.

2.4 Ingress filtering

Ingress filtering refers to the verification of the source address of the IP packets at the periphery of the Internet, typically at the link between a customer and an ISP. This process, which is described in [9] is intended to thwart a class of denial of service attacks in which attackers hide their identity by using a "spoofed" source address.

2.5 Site exit anycast identifier

A 7 bit anycast identifier whose value is XX. [TBD IANA]

2.6 Site exit anycast address

An IPv6 anycast address built by combining an IPv6 address prefix allocated to the site with the site exit anycast identifier, according to the rules specified in [RFC2526]. The proposed used of this anycast address is detailed in section 5.

3. Host-Centric IPv6 Multihoming issues

Host-Centric IPv6 Multihoming forces hosts to choose the source and the destination address of the IPv6 packets, in a way that makes the best usage, or at least a reasonable usage, of the network resource. Hosts must first select a destination address, and will then perform source address selection. Source address selection must be consistent

with ingress filtering, which is sometime implemented at network interfaces: we call this the "site exit" issue. Destination address selection is often based on incomplete or obsolete information, which can be harmful if, for example, hosts fail to notice that one of the site's connections is suddenly made unavailable. In any case, we must also consider "low budget" hosts, and make sure that these hosts can get some benefits from multihoming without enduring too much cost.

3.1 Destination address selection

It is fairly common for hosts to have to choose between multiple destination addresses for a peer. TCP performs this choice when the connection is instantiated; SCTP may perform similar choices through the life-time of the connection; UDP may perform this choice either for each packet, or at the beginning of an association. We may debate whether hosts have sufficient information to perform a valid choice, and it is a complex debate. Some very simple appliances probably never will have any information; large servers potentially have tons of information available; personal computers are somewhere in between. It is not unrealistic to expect progress in this area, either by communication between the hosts and the routers, by sharing of experience between hosts, or maybe by innovative application design that would for example implement a file transfer by parallel retrieval of fraction of the file from multiple sources. At worst, a host can always try the proposed addresses one by one, and pick the first one that actually works -- not very elegant, but definitely workable.

3.2 Source address selection

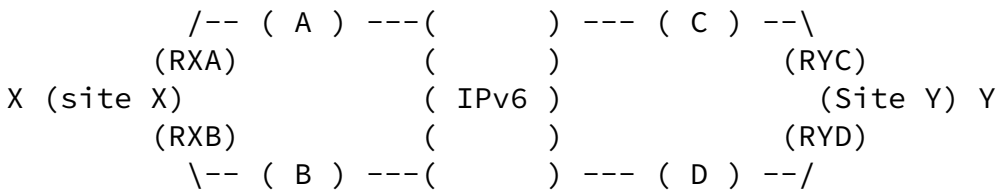
The source address selection in most hosts immediately follows the destination address selection. When a host has multiple interfaces, the normal procedure is to select the destination address, then identify the interface over which packets bound to that address will be routed, and finally select a source address associated to that interface. When the host has multiple addresses attached to an interface, which is the case with host centric IPv6 multihoming, the host could in theory pick any of these addresses, or at least any of those that have an appropriate scope. In our example topology, supposing that X has selected the destination address "C:Y", it can choose as source address either "A:X" or "B:X".

Choosing the source address will affect the reverse path of the connection, as the source address of the message will become the destination address of the responses. This may be a serious matter in asymmetric applications like web access, in which the bulk of the data is sent by the server to the client. If the multiple addresses correspond to different ISP, the hosts should normally pick the source address that will provide the best performances. As for destination address selection, we may expect that the host will have some knowledge of the effect of choosing one or the other address, for example by observing that web pages are served faster through one address than through the other.

3.3 The site exit issue

A special complication appears when the ISPs who serve the multihomed site perform "source address selection." In our generic configuration, we assume that X is served by ISP A and B, and thus can be reached by the addresses A:X and B:X. We also we assume that Y is served by ISP C and D, and thus can be reached by the addresses C:Y and D:Y. To communicate with Y, X will choose the destination address that appears to be easier to reach, for example D:Y; then, it will choose the source address that provides the most efficient reverse path, say A:X.

Suppose now that the ISP connections at Site X are managed by two different site exit routers, RXA and RXB, and that there is a similar configuration at Site Y, with routers RYC and RYD.



Within Site X, the interior routing will decide which of RXA or RXB is the preferred exit router for the destination "D:Y"; similarly, within Site Y, the interior routing will decide which of RYC or RYD is the preferred exit for destination A:X. If the chosen exit router at Site X is RXA, the packet will flow freely to RYD; If the chosen exit router at Site Y is RYD, the response will also flow freely. However, if the exit routers are RXB or RYC, and if the ISPs perform ingress filtering, we have a problem: ISP B sees a packet coming from RXB, whose source address does not match the prefix assigned by B to X; ISP C, similarly, sees a packet whose source address does not match the prefix assigned by that ISP to Y. If either of these ISPs decides to drop the packet, the communication will be broken.

3.4 Rapid reaction to topology changes

Network conditions change over time. In order to meet the performance requirement, we must allow the use of the best path at any time; In order to meet the "redundancy" requirement, we have to make sure that if a network connection breaks, the corresponding prefix is not used as either a source or a destination address.

We may assume that the destination address selection algorithm mentioned in 3.1 will naturally result in the selection by X of an appropriate address for Y; X may for example try in turn the addresses C:Y and D:Y, and retain the address for which a response comes back. However, we must make sure that X selects a source address that will be reachable: for example, if the link to ISP A fails, X must make sure that it uses as source address B:X, not A:X.

4. Analysis of solutions to the site exit issue

The site exit issue is caused by ingress filtering at the site egress. In this section, we will analyse four ways to solve the issue: somehow relax the source address check, implement source address dependent routing, ask hosts to pick "the right" source address, or ask routers to somehow rewrite the packets so that it can pass the source address checks. We will then compare the proposed solutions, in order to prepare a recommendation.

4.1 Relaxing the source address checks

An obvious way to avoid failures due to ingress filtering is to simply make sure that all the addresses used by the hosts of a given site will be considered acceptable by each of the site's providers. In our site X example, that would mean that provider A would accept addresses of the form "B:X" as valid, and that provider B will in turn accept addresses of the form "A:X" as valid.

One way to achieve this is simply to ask the service provider to turn off source address checks on the site connection. This requires a substantial amount of trust between the provider and the site, as source address checks are in effect delegated to the site routers. One possible way to achieve this trust is to make sure that the site routers, or possibly the site firewalls, meet a quality level specified by the provider.

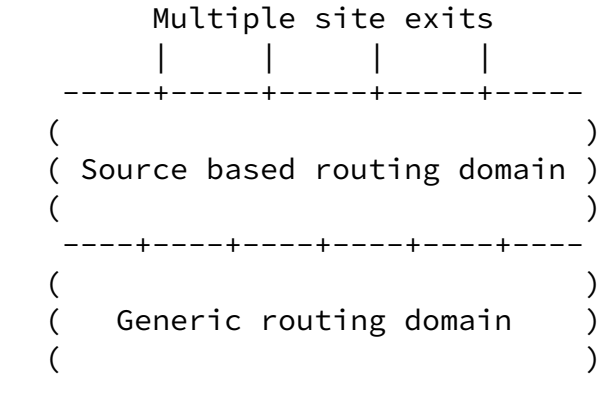
Another way to achieve this relaxed level of checking is to check source addresses against a list of "authorized prefixes" for the site connection, rather than simply the single prefix delegated by the provider. This solution requires that the site communicates the authorized prefixes to the provider, either through a management interface or through a routing protocol. This is obviously more complex than simply lifting the controls, and in fact ends up with a very similar requirement of trust: the provider has to believe that the site will transmit the right prefixes.

A special case occurs when the site exit is an "automatic tunnel" interface, such as 6to4 or Teredo. Source address control for automatic tunnels is delegated to the "other side" of the tunnel, in practice to a very large number of relay routers located across the Internet. Checks are based on the correlation between the IPv6 source address and the IPv4 source address used in the tunneling protocol. Asking each potential end of the tunnel to relax its checks is not realistic; in practice, this means that the exit routers will have to obtain the right to use as source address a privileged IPv4 address, such as the 6to4 anycast address; this will imply a negotiation with the provider of the IPv4 service.

In conclusion, relaxing the source address checks requires some form of explicit trust between the site and its providers. There is no doubt that this level of trust will exist in many cases; there is also no doubt that there will be many cases in which the provider is unwilling to grant this trust, particularly in the case of small sites, such as for example home networks dual-homes to a DSL provider and a cable network provider.

4.2 Source address dependent routing

Another solution to the site exit problem is to perform some kind of source routing within the site, so that the site exit is effectively a function of the source address in the packet. Current routers generally do not implement any kind of source address dependent routing; this implies that this solution would have to be "rolled in" progressively in the site, following a generic schema such as:



In this schema, all site exit routers are connected to a source based routing domain. Packets initiated in the generic routing domain and bound to an "out of site" address are passed to the nearest access point to the source based routing domain, using classic "hot potato" routing. The routers in the source based routing domain maintain as many parallel routing tables as there are valid source prefixes, and would choose a route that is a function of both the source and the destination address; the packets exit the site through the "right" router. There are multiple possible implementations of this general concept.

The simplest implementation is to have only one exit router for the site; this exit router chooses the exit link on the basis of the source address in the packet. This simple implementation might be adequate for very small sites, but introduces a single point of failure, and thus fails to meet the "redundancy" requirement of multihoming.

In the most complex set up, each router of the site would maintain as many parallel routing tables as there are valid source prefixes, and would choose a route that is a function of both the source and the destination address. This solution enables "shortest path" routing and can provide an arbitrary level of redundancy. However, maintaining parallel routing tables requires a massive re-engineering of routers and routing protocols, and thus would be hard to deploy in the short term.

A slightly less complex implementation is to connect all exit routers to the same link, e.g. to what is often referred to as the "DMZ" for the site. This solution requires that all routers connected to the DMZ are upgraded to perform source address based routing. This configuration is less fragile than a single router solution; however, the single link requirement seems to preclude "geographic redundancy" between the site exits. It does requires the re-engineering of some routers, but not necessarily all routers of the site. In practice, it could be a way to "phase in" the most complex setup described in the previous paragraph.

A much simpler alternative is to establish a mesh of "tunnels" between the site exit routers. A site exit router that receives a

packet bound for an out-of-site address would perform a source address check before forwarding the packet on one of its outgoing interfaces; if the source address check is positive, the packet will effectively be sent on the interface; if it is not, the packet would be "tunneled" to a more adequate router.

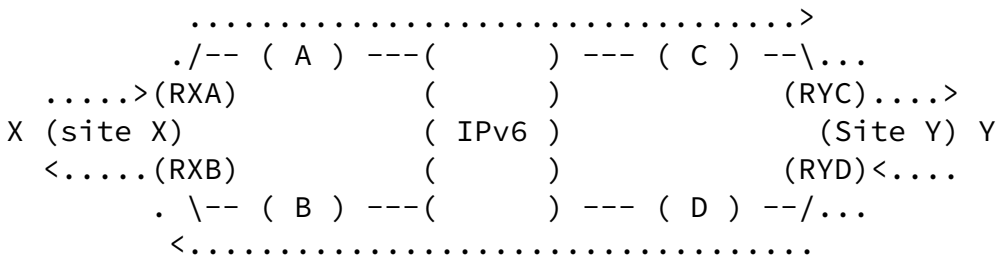
The main requirement of the tunneling alternative is that site-exit routers be able to perform address checks, and that each site exit router be able to associate to each valid site prefix the address of a corresponding site exit router. An obvious possibility is to configure prefixes and corresponding addresses in each router; it would however be preferable to derive these addresses automatically. A strong assumption of the IPv6 architecture is that all prefixes of a site will have the same length; it is thus possible to derive a prefix from the source address of a "misdirected" packet, by combining this prefix with a conventional suffix. The suffix should be chosen to not collide with the subnet numbers used in the site; a null value will be inadequate, since it could be matched by any router with knowledge of the prefix, not just the site exit router; a value of "all ones" could be adequate.

In order to enable tunneling, each router managing a site prefix will then inject a "host route" for its locally managed prefixes in the interior routing protocol. Site exit routers performing automatic tunneling can then use the standard routing procedures to detect whether the anycast address corresponding to the prefix in use is

reachable; they can automatically reject, rather than tunnel, packets whose source address does not correspond to a reachable anycast address.

An inconvenience of the set-up is that some packet will follow a less than direct path; we will see in the next section how that could be palliated by host based processing.

Source based routing allows for a large diversity between the site exits; it also allows for host based policy decision, since a host can influence the routing of a packet by choosing the appropriate source address. There is however one drawback of any source address based scheme, the impossibility to use "asymmetric" path between two sites:



Using source based routing implies that if the host X chooses the source address B:X, then its packets will exit through router RXB, never through RXA. This may provide lesser performances if a link is congested in one direction but not in the other. However, source based routing would allow four paths, A-C, A-D, B-C and B-D, thus providing an adequate redundancy and allowing a great deal of performance optimization.

4.3 Source address selection by the host

The site exit issue would be mitigated if the hosts chose a source address that would be compatible with the exit point chosen by the routing protocol, or alternatively if the host tunneled the packet directly to an adequate exit router.

The first alternative could be called "source address discovery". In many ways, source address discovery is similar to path MTU discovery. The two issues are similar: packets that do not meet some criteria fixed by the network are dropped; the host has to find the cause of the loss, and to take action in order to make sure that these packets will be accepted. In the path MTU case, the action is to use shorter packets; in the ingress filtering case, the action is to present a different source address.

To implement source address discovery, the hosts would have to introduce a "preferred source address" parameter in the "destination cache" mentioned in the Neighbor Discovery standard [3]. The primary purpose of the cache is to link a destination address to a next hop neighbor; it is also the repository of per-destination parameters such as the path MTU; it is the natural repository for the new parameter. The source prefix in the destination cache would be used during source address selection, to select an available interface

address that matches the prefix.

As for path MTU discovery, source address discovery requires that the hosts receive some information from the network. Such information can be conveyed in an ICMP Destination Unreachable error message with code 5 which means source address failed ingress policy [18]. The router is supposed to send such message when the packet is discarded because of ingress filtering issues. The error message contains information about the packet that triggered the error. However, the host will also need information about the source address prefix that should be used to pass the source address check. The proposed format of the error message does not includes such information. However, a proper choice of the source address by the router that generates the message can provide a good substitute. This means that the router that generates the error message will have to include the prefix that complies with the ingress filtering in the source address of the packet that carries the error message. The host will then select the source address to be used for the selected destination by performing a longest prefix match between the source address contained in the error message and the potential source addresses. In the absence of an explicit ICMP message, the hosts would have to rely on a trial an error process, noticing that packets get dropped and trying retransmissions with alternate source addresses; the experience of path MTU discovery shows that such processes are awkward and error prone.

An alternative to source address discovery is "exit router discovery", i.e. the discovery by the source of the preferred exit router for a given source address. This requires a slightly different change to the caches used in neighbor discovery, specifically the management of a "source exit cache" that associates a specific source address with an exit router, or maybe the combination of a destination address and a source address with an exit router. As with source address discovery, this would be learned through an ICMP message; this message would not be an error message, but rather a variation of the redirect message. After receiving such messages, the host will tunnel to the specified exit point the packets sent from the source address to the destination; the exit point will decapsulate these packets and send them over the appropriate exit link.

The "exit router discovery" procedure appears to be superior to the

"source address discovery." Both solutions require approximately the same amount of resource in the host, but the exit router discovery has two advantages: it enables hosts to actually specify the point of exit from the site, thus giving them a greater amount of "policy control".

We should note that neither "source address discovery" nor "exit router discovery" are implemented in current hosts. In order to accomplish the goal expressed in [7] that hosts implementing the current version of IPv6 can continue to operate in a multi-homed site, even if they would not take advantage of multihoming; in consequence, these procedures can only be used as an optional optimization.

4.4 Packet rewriting at exit router

In section 4.2, we explained how a site exit router that discovers that a packet bound out of the site has the "wrong" source address can route the packet to an alternative exit. Another way to pass the source the source address check is to modify the packet, which could in theory be done by replacing the source address or by encapsulating the packet using "IPv6 in IPv6".

In fact, replacing the source address is not necessarily a good idea, since this will remove information from the packet; it also requires some level of cooperation between the exit router and the host, if only to understand what alternative source addresses can be used by the host, if any.

One could preserve the information by encapsulating the packet in a new IPv6 header, using "IP in IP". The source address of the new header will be the address used by the router on the exit interface, the destination address will be the original destination, and the payload type will be set to "IPv6." After the insertion of the option, the outgoing packet will have the following values:

* outer IPv6 header source address: address of the site egress interface,

* outer IPv6 header destination address: from initial packet,

* outer IPv6 header payload type: "IPv6",

* inner IPv6 header source address: source address of initial packet,

* inner IPv6 header destination address: from initial packet,

* inner IPv6 header payload type: payload type of initial packet,

4.5 Comparison of the site exit solutions

The four solutions that we have reviewed have different advantages and inconveniences. The may differences are in terms of deployability, generality, redundancy, policy control, and impact on existing hosts, i.e. minimal implementations of IPv6 that would use only one of the available prefix for the site, that would not perform any more sophisticated logic than picking a destination address at random among multiple alternatives, and that would not understand any additional IPv6 option or any additional ICMP message.

The relaxation of source address checks detailed in 4.1 is easy to deploy, and would not affect minimal hosts. It is a perfectly reasonable solution for large sites, i.e. the sites that benefit of IPv4 multihoming today: it should not be more complex to convince a provider to relax address checks for a particular customer tomorrow, than to convince today a similar provider to advertise in its routing table the global IPv4 address of the site. If we choose this solution, we should choose its simplest implementation, i.e. one in which the provider completely delegates source address checks to the site's router or firewalls. This is however not a general solution, since we cannot expect all sites to convince every provider to relax their checks.

The rewriting at exit routers appears to be an inferior solution. It is not really easier to implement than the "tunneling" variation of source routing at the exit sites: if a router can detect that a source address does not pass the checks for a proposed interface, and if it can encapsulate the packet before forwarding it, then it could just as well tunnel the packet to the "correct" exit router for the site. Tunneling the packet to its final destination actually has a larger impact on the existing hosts than simply tunneling the packet to another router: we have to assume that the destination host is willing to accept tunneled packets, which is not an obvious proposition. Since the packet is tunneled, the destination host has to trust that the source address in the encapsulated packet is genuine; in the absence of an authentication header, this is risky proposition.

When the source address checks cannot be relaxed, the best solution is probably to perform some kind of source address based routing to the adequate exit router. In the long term, the IETF may develop internal routing protocols that take into account the source address as part of the "reachability information" for a set of destinations;

in the short term, there are no such protocols, and we have to rely on a tunneling mechanism between site exit routers.

Exit router discovery is a natural complement of the tunneling mechanism between site exit routers. When an exit router tunnels a misdirected packet towards another exit, it may send an appropriate "exit redirection" ICMP message. If the host is a minimal IPv6 host, the ICMP message will be ignored; further packets will continue using the same slightly sub-optimal path. On the other hand, if the host has been upgraded to take advantage of multi-homing, the packets will be tunneled to the appropriate exit router; they will follow a direct path to this router.

5. Analysis of solutions to provide rapid reaction to topology changes

In order to fulfill the "redundancy" requirement, a multihoming solution has to provide the means to identify the available exit paths towards a given destination and carry packets through it. In other words, a mechanism to detect unavailable exit paths is required, so that they are not used. We will analyze the different mechanisms to perform the path selection in two situations: path selection when establishing a new communication and path selection during the lifetime of a communication. These two problems are quite different, since the timing requirements are different in the two situations and also requirements imposed in the addresses to be used are different.

5.1 Path selection when establishing a new communication

5.1.1 Externally initiated communications

We will first analyze the mechanism used by hosts outside the multihomed site to select among the paths to the multi-homed site. We have already assumed that the multihomed site will have as many prefixes as ISPs, and that it will assign multiple addresses to every host that will benefit from multihoming. It is also assumed that those addresses will be announced through the DNS.

So, when an external host tries to establish a communication, it will first obtain all the host's addresses from the DNS. Then it will try to use one of them and if it succeeds the communication is established; and if not, it will try with the next address. Considering that each address is routed through one and only one provider, the selection of an address implies the selection of a provider, then it implies the selection of a incoming path to the

multihomed site. So, for external hosts, incoming path failure detection and incoming path selection is already being performed by the external host itself and the provided capabilities are enough to provide support to the multihomed environments. When the host within the multihomed site replies to the incoming packet, both the destination and the source addresses are already determined, so no selection has to be performed by the host in the multi-homed site. Moreover, since the incoming packet has reached the host within the multihomed site, and assuming that some mechanism to guarantee ingress filtering compatibility mechanism is being used, the exit path will be the same than the ingress path, so it is likely to be working properly.

5.1.2 Internally initiated communications

We will next analyze the mechanisms required within the multihomed

site to select among the multiple path connecting the site to the Internet. When a host within the multihomed site sends a packet to a given external destination address, the exit path through which the packet will be routed has to be selected. In order to select a path two mechanisms are needed: a mechanism to discover the available paths and a mechanism to route the packets through the path identified as available. We have two elements that may perform these tasks: the routing system and the host itself.

We will analyze the following possibilities:

The first possibility is to let the intra-site routing system perform both tasks.

The second possibility presented is to let the host do both tasks.

The third possibility is to use the routing system to identify the available paths and to use a mechanism in the host to force the routing of packets through the identified path.

The fourth possibility where the host identifies the available path and the routing system routes the packet through the path identified by the host doesn't seems a reasonable approach to us, so it will not be included in our analysis.

5.1.2.1 Exit path selection by the intra-site routing system

One possibility is to let the intra-site routing system perform the complete exit path selection mechanism. In order to do that, intra-site routing system requires information about which destinations are reachable through each one of the exit paths. This means that each one of the providers has to inform the multi-homed site which destinations are reachable through him. Normally, the BGP protocol is used for this task. From the multihomed site perspective, there are two difficulties with this approach: first, the amount of information that is to be injected in the intra-site routing system is important and second, running the BGP protocol is more than a trivial task. While there are some medium-big multihomed sites that certainly can deal easily with these two issues, other smaller multi-homed sites may not deal with them so easily (imagine for instance a site consisting a few PCs on a single Ethernet and a single router connected to the Internet through a DSL access and a cable access).

We can explore approaches to try to reduce the amount of routing information to be injected to the multi-homed site. The most aggressive approach would be to inject only a default route through each of the ISPs. This case works fine when one of the direct links

between the multihomed site and ISP fails, but, if we only want to provide protection for this specific case, RFC 3178 provides a solution that it is simpler overall since it deals with all the problems for this particular case (like ingress filtering, transport survivability, etc). So, since we are looking for a solution that provides better fault tolerance capabilities than RFC 3178, we need more information to be injected to the intra-site routing system.

We need then alternatives that allow us to obtain better fault tolerance. A possible approach is to filter the accepted routes based on the AS path length, as proposed in [12]. By this mechanisms, the multihomed site would only accept routes with an AS path attribute whose length is no longer than x ASes. This approach allows reducing the amount of routing information while still achieving a high level of fault tolerance. The value of x is a trade-off between the two of them. As more routing information is injected into the site (higher x), better path selection will be performed and better fault tolerant capabilities will be provided by the solution, but at the same time

more resources will be needed within the multihomed site. However, configuring filters raises the difficulty of running BGP, requiring additional BGP expertise in the end-site, making the adoption of this solution harder for small unmanaged sites.

5.1.2.2 Host based exit path selection

An alternative to intra-site routing system exit path selection is to move exit path selection to the host itself. In order to enable the host to perform the exit path selection, two mechanisms are needed: a mechanism to discover available paths and a mechanism to enable the host to force the routing of packets through the selected exit path, overruling intra-site routing system routing.

5.1.2.2.1 Mechanisms to force the routing of packets.

A possible mechanism to let the host force the path of the packets is to make a tunnel directly to the exit router. In order to do that, the host must be able to discover the address of the exit router. Using an "Exit Router Discovery" ICMP message as presented in section 4.3 would be an option. An alternative to tunnels could be the usage of routing headers. However this is considered an inferior solution since the routing header would be carried all along the path to the final destination even if it were not needed.

Another mechanism to enable the host to select the exit path is available when some form of source address dependent routing is used within the multihomed site. As it has been presented in section 4.2, if each exit ISP is associated with one of the available prefixes, and source address dependent routing is used, selecting the prefix to

be included in the source address implies the selection of the exit ISP through which the packet will be carried. So, source address dependent routing can be considered as an option to allow the host to select the exit path.

5.1.2.2.2 Mechanism to discover available and unavailable paths

A mechanism to identify available paths is just to let the host do trial and error procedure. That is, in order to reach a certain destination, the host tries every possible exit path. The procedure can be carried out either sequentially or in parallel, that is, the

host can try with every path simultaneously or it can try with one path and if the chosen path fails then it tries with the next one. The benefits and drawbacks of these two approaches are clear: the sequential procedure may take longer to find the available path, but the parallel procedure consumes more resources since multiple packets are sent every time an available path has to be discovered.

However, the implementation of a failure detection mechanism based on sending packets may be trickier than what it may seem. A possible approach could be to define a new protocol for detecting available paths that sends probe packets end to end. However, a solution that doesn't impose changes in hosts outside the multihomed site is preferred because it is easier to deploy. So, we have to use already available mechanisms. Among the available choices, we could use ICMP echo packets to detect path availability. The problem here is the wide adoption of ICMP filtering because security issues.

The other available option is to use data packets as probes. The main problem here is that not all applications are bi-directional, so there may be cases when no packets will return but the path is available. However, we consider that an important number of applications are bi-directional, so we will explore this possibility (Note that we are not considering the multicast case here, where the unidirectional flows are more common). So, a path failure detection mechanism based on data packets stores the exit path information corresponding to a destination address in a cache, the Exit Path Cache. The information contained in this cache depends on the mechanism that is used to force the routing of the packet by the host. If the tunnel mechanism is used, the address of the exit router and the source address to be included are cached. If source address based routing is used, only the source address to be used is cached.

So, when a packet is to be sent to a certain destination address, the Exit Path Cache is searched for an exit path corresponding to the destination address. If no exit path is found in the cache, the host has no knowledge about the available paths, so it has to start the failure detection procedure by sending packets through all the

available paths. As we have seen, such procedure can be performed sequentially or in parallel, but in any case packets will be sent through the available paths and the host will wait for replies. When the first reply is received (whether because packets through all

available paths have been sent simultaneously or because packets through different exit paths have been sent and a timeout has occurred, so the packet has been retransmitted through an alternative destination), the exit path used is stored in the Exit Path Cache and following packets are sent through the same exit path. Exit Path Cache entries have a finite lifetime. An Exit Path Cache entry lifetime is extended every time that a packet is received coming from the corresponding exit path. When an Exit Path Cache entry lifetime expires, the failure detection procedure is performed when new packets arrive for such destination.

A case that has to be considered is when no reply packets for a given destination are received from any exit path. Such behavior may have two causes: the application generates unidirectional traffic, so no packets are supposed to arrive or all the paths are down. In any of the two cases the mechanism can't do anything to select the exit path, so when such situation is detected, a random exit path has to be selected and used. So, an Exit Path Cache entry is generated with a random path and with a certain lifetime. When the lifetime expires, the failure detection mechanism is performed again, so that if the case was that all exit paths were down, the mechanism can detect when one of the paths is up again. Note that this would cause additional overhead for unidirectional applications, so the failure detection mechanism should not be performed very often i.e. the lifetime should not be very low.

5.1.2.3 Hybrid approach: Routing system based failure detection and host based exit path selection

An alternative approach is to obtain the information about available paths from the routing system but let the host to force the routing of packets through the identified exit path. The benefit of this approach is that the routing information injection into the intra-site routing system is not required because the exit path is selected by the host.

This hybrid approach requires a mechanism to convey the path availability information from the routing system to the hosts. Considering the amount of information involved, we consider that it is better to limit the path information stored in the hosts to the information concerning the paths that the host is currently using. There are two approaches that can be used at this point.

One possible approach is to define a new protocol to let the host to

query a server for the correct exit path to be used to reach a certain destination, for example as defined in [16]. The server would have to be configured with enough information to answer those queries. For instance the server has to know all the BGP information that is received from each one of the ISPs, the associated prefix and the address of the corresponding exit router. So, when a host wants to initiate a communication with an unknown destination address, it queries the server and obtains the exit path to be used. Then the host itself forces the packet to be routed through the exit path.

An alternative option is to let the exit routers to inform the host about the correct exit path to be used. In this case, only the exit routers are running BGP. So, when a host sends a packet to a new destination, it randomly selects the exit path. More likely, the host will randomly select a source address and won't tunnel the packet, so that the packet is carried to the default route. In case that the destination contained in the packet is not reachable through the ISP whose prefix has been included as source address, but the exit router knows because of BGP that it is reachable through an alternative exit router, the exit router will send an ICMP error message containing the exit path information back to the host.

A particular case of this approach can be used when the failure has occurred in the direct link. In this case, the exit router can detect the outage and considering that no destination is reachable through this ISP, simply deprecate the prefix. This approach is only an optimization since it does not address the general case.

5.2 Preserving established communications

Multiple solutions for preserving established communications have been proposed such as HIP, SIM, ODT, LIN6, MAST, NOID, CB64. Many of these approaches mainly focus on how to present an unchanged IP address to the upper layers through changes in the address used to actually reach the host. However, not only such mechanism is required in order to preserve established communications, but a mechanism to perform path selection is also required (both a mechanism to identify available paths and a mechanism to force the routing of packets through the identified path are required). Additionally, a mechanism to solve the site exit issue may be needed in those solutions.

Next versions of this document will include an analysis of which mechanism can be used to select paths during the lifetime of a communication and how such mechanisms can interoperate with the proposed solutions.

6. Integrating Solutions

In this section we will integrate solutions, combining a site exit issue solution with a path selection solution. Next, we will present what we consider to be the most natural combinations of a solution for the site exit issue and an exit path selection mechanism. Other combinations may be possible but they don't seem very natural so far. Perhaps future versions of this document will consider them if it seems appropriate.

6.1 Solution 1: Relaxing source address checks + Intra site routing system based exit path selection

The site exit issue is addressed relaxing the source address checks at the ISP, since the required level of trust exists between the site and the ISPs. The exit path selection is addressed using BGP and injecting some of the information into the intra-site routing system. Since BGP expertise is available, appropriate filters can be configured.

Requirements:

- enough level of trust between the site and the ISPs in order to relax the source address check

- enough expertise to run BGP and configure appropriate filters

- enough resources to import part of the global routing table into intra-site routing system

Suitable for: big/medium sites. The solution is not deemed suitable for small sites because of the required level of expertise and resources.

6.2 Solution 2: Source address Discovery + Intra site routing system based exit path selection

The host generates packet with one if its source address and then the packet is routed according the information available at the intra-site routing. When the packet reaches the site border router,

it verifies the source address. If the source address is compatible with the selected ISP, the packet is forwarded as it is, if not, the exit router discards the packet and sends an ICMP Destination Unreachable error message (code 5) back to host informing the appropriate source address for the selected destination.

Requirements:

- enough expertise to run BGP and configure appropriate filters

- enough resources to import part of the global routing table into intra-site routing system

- Required modifications: all hosts within the multihomed site have to be modified to implement the processing of the ICMP error in order to work properly. Communications initiated by hosts within the multihomed not implementing such processing will fail when selecting the wrong source address. Those hosts will not obtain even the level of service they would obtain in a single homed site.

Suitable for: big/medium sites The solution is not deemed suitable for small sites because of the required level of expertise and resources.

6.3 Solution 3: Packet Rewriting at site exit + Intra site routing system based exit path selection

The host generates packet with one of its source address and then the packet is routed according the information available at the intra-site routing. When the packet reaches the site border router, it verifies the source address. If the source address is compatible with the selected ISP, the packet is forwarded as it is, if not, the exit router rewrites the source address of the packet with a new prefix compatible with the exit ISP.

Requirements:

- enough expertise to run BGP and configure appropriate filters

- enough resources to import part of the global routing table into intra-site routing system

- Required modifications: all hosts within the multihomed site have to be modified to implement a mechanism to recognize reply packets with modified destination address as valid replies to the initial packet. Communications initiated by hosts within the multihomed site not implementing such mechanism will fail when using a source address that is rewritten at site exit. Those hosts will not obtain even the level of service they would obtain in a single homed site. Additional modification in applications and/or external hosts may be required.

Suitable for: big/medium sites The solution is not deemed suitable for small sites because of the required level of expertise and resources.

6.4 Solution 4: Host based exit path selection + source address based routing

In this case, an exit path is determined by the source address selected included in the packet. So, the host can force the routing of a packet through an exit path just by selecting the source address. So, the host determines which paths are available by sending packets with different source addresses. If reply packets arrive, the path associated with the destination address included in the reply packet is available, so the address is introduced in the site exit path cache.

Requirements:

- Configuration of source address dependent routing. This can be configured site wide or just at the site exit routers. In the second case, a mesh of tunnels between of the site exit router has also to be configured

- Required modifications: implementation of the path discovery mechanism in the hosts of the multihomed site in order to benefit from multihoming. Hosts not implementing such mechanism can configure a single source address and behave as they were in a single homed site. Source address dependent routing is supported by some router vendor.

Suitable for: small sites While this solution may be adopted by medium and big sites, those sites may prefer other type of solutions based on BGP because policy issues. This solutions relies on hosts to implement policing, since the hosts themselves perform the path selection. Other solutions based on BGP enable policy configuration on router or in central servers. This last option is considered to be more scalable with respect to the number of hosts within the site, making it more attractive for medium and big sites.

6.5 Solution 5: Host based exit path selection + site exit discovery

In this case, hosts within the multihomed site tries to discover the available exit path by generating packets with different source address. In the case that the exit ISP corresponds to the selected source address, the packet is forwarded through the ISP. If not, the packet is discarded and an ICMP error message containing the appropriate exit router is sent back to the host. The host then retries forcing the routing of the packet, tunneling it directly to the exit router. The host identifies available paths when it receives reply packets. The host then stores the information about the source address and optionally about the exit router to be used in the site exit path cache.

Requirements:

- Required modifications: implementation of the ICMP error generation mechanism in the routers. Implementation of the path discovery mechanism and the processing of the ICMP error in the hosts. Communications initiated by hosts within the multihomed not implementing such mechanism will fail when using an incompatible source address. Those hosts will not obtain even the level of service they would obtain in a single homed site.

Suitable for: small sites While this solution may be adopted by medium and big sites, those sites may prefer other type of solutions based on BGP because policy issues. This solutions relies on hosts to implement policing, since the hosts themselves perform the path selection. Other solutions based on BGP enable policy configuration on router or in central servers. This last option is considered to be more scalable with respect to the number of hosts within the site, making it more attractive for medium and big sites.

6.6 Solution 6: Hybrid approach + source address dependent routing

In this case, a server or the exit router has the information about the correct site exit router and source address to be used for a given destination. So, the host within the multihomed site contacts the server and obtains the correct site exit router and the appropriate source address. Then the hosts sends packets with the appropriate source address so that it routed trough the correct exit router,

Requirements:

- enough expertise to run BGP and configure appropriate filters

- enough resources to import part of the global routing table into intra-site routing system

- Required Modifications: the exit router or a server has to inform the host about the correct source address. So both the router and hosts has to be modified to implement the mechanism

Suitable for: medium and big sites The solution is not deemed suitable for small sites because of the required level of expertise.

6.7 Solution 7: Hybrid approach + source address selection by the host

In this case, a server or the exit router has the information about the correct site exit router and source address to be used for a given destination. So, the host within the multihomed site contacts

the server and obtains the correct site exit router and the appropriate source address. Then the host sends packets with the appropriate source address tunneling it to the correct exit router,

Requirements:

- enough expertise to run BGP and configure appropriate filters

- enough resources to import part of the global routing table into intra-site routing system

- Required Modifications: the exit router or a server has to inform

the host about the correct exit path. So both the router and hosts has to be modified to implement the mechanism

Suitable for: medium and big sites The solution is not deemed suitable for small sites because of the required level of expertise and resources.

6.8 Remaining combinations

The remaining combinations of site exit issue solutions with site exit path selections mechanism are considered no to naturally match together.

For instance, it is considered that sites that obtain the level of trust required from its provider to relax the source address checks will prefer to run BGP to obtain the available path information rather than using a host based or hybrid approach.

In source address dependent routing and in site exit discovery approaches to the site exit issue, it is the host itself who selects the exit path (using the source address or tunneling). This type of mechanism seems hardly compatible with intra-site routing system exit path selection, since it is no longer the intra-site routing system that selects the exit path but it is the host through the source address who performs that selection.

7. Proposed solution

In order to implement the host centric multihoming solution, we must solve the issues presented in the previous section. In this section, we will present the recommended ways to solve the site exit issue and

how to trigger rapid reactions to failures.

We will next present different solutions for different scenarios. As we have concluded from our analysis presented above, different solutions have different requirements and are then suitable for different type of scenarios. We will consider three different scenarios:

- multihoming for small sites

- multihoming for medium sites

- multihoming for big sites

7.1 Multihoming solutions for small sites

It is not likely that small sites can obtain some form of source address check relaxation from their IPSs, so an alternative solution to deal with the site exit issue is to be used. It is also considered that in general, small sites don't have neither the expertise nor the resources required to run BGP, so an alternative mechanism to react to topology changes is required. We think that the host based approach is the mechanism that better suits the requirements of the small sites.

We will next present a set of mechanism and tools to enable multihoming is small sites.

The goal is to propose a roadmap to adopt multihoming that preserves existent functionalities and adds new functionalities progressively. This would allow legacy systems to keep on working exactly the same way they did before multihoming adoption and then add new features to enable multihoming benefits

7.1.1 Step 1: preserving functionality for legacy hosts when becoming multihomed.

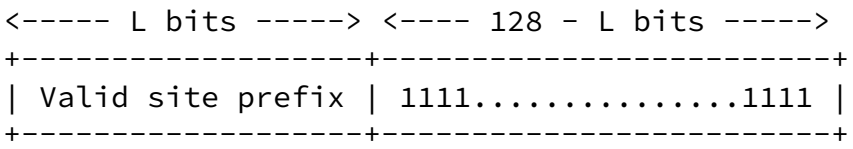
Suppose that a single homed site becomes multihomed. The problem here is that the site exit issue will affect communications of the newly multihomed site. So, the first step is to deploy a solution for the site exit issue as simple as possible that does not require updating the hosts of the site, and if possible does not requires updating other equipment. Using such solution, legacy hosts within the

multihomed site will work as if they were in a singlehomed site. That is, they will not obtain multihoming benefits, but at least they will not fail because of multihoming. This solution also allows then attaching legacy hosts to the site and they will work as if they were in a singlehomed site.

In line with the analysis presented in the previous section, our recommendation is to enable multihoming by establishing tunnels between the site exit routers. In order to implement this solution, we must define a way to convey the site exit addresses to the various routers in the site; the simplest solution, which we propose here, uses an anycast address that is arithmetically derived from the sites' prefixes.

7.1.1.1 Site exit anycast address

The site exit anycast address solution assumes that all of the sites prefixes have the same length L; it also assume that we can define a conventional "subnet" associated to the prefix. The proposed solution is to compose the anycast address by appending an "all 1" suffix to the site prefix:



-- Site exit anycast address --

Each site exit router that can forward to the outside packets whose source address is derived from a specific site prefix will advertise reachability of the corresponding site exit anycast address through the routing mechanism.

7.1.1.2 Tunneling to the appropriate exit

Site exit routers are expected to perform necessary source address checks before forwarding any packet on a site exit link. The site exit router must check the source address first, in order to avoid local packets being routed to a black hole. If the result of the check is positive, the packet will be forwarded. If the result is negative, the router will derive a "site exit anycast address" from the source address of the incoming packet. If the anycast address is unreachable, the incoming packet will have to be discarded. If the anycast address is reachable, the incoming packet will be tunneled towards that address.

It is accepted that tunneling mechanism introduces additional overhead, but it is recommended because its fast deployability. However, in the long run, a mechanism based on source address dependent routing will provide better performance. Such mechanism can be gradually deployed as presented in section 4.2

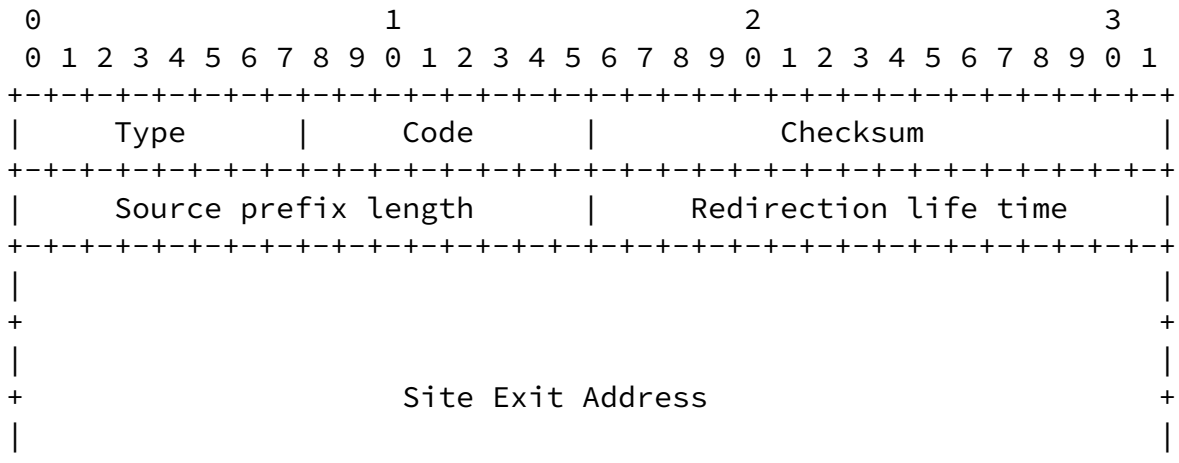
7.1.2 Step 2: Optimizations to enhance the multihoming support

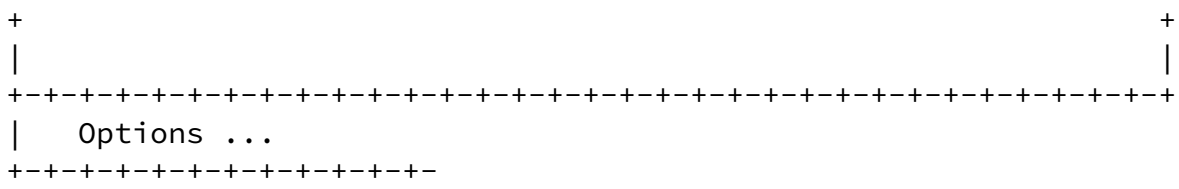
7.1.2.1 Site exit router discovery

A problem with the approach presented in the previous section is that packets may be routed through a suboptimal path, because they are initially routed towards a site exit router and if the source address doesn't match, this site exit will send the packet through a tunnel to an alternative exit router. This problem can be solved using a Site Exit Redirection ICMP message. This ICMP error would be generated by the site exit routers when they receive a packet whose source address doesn't match with the exit ISP. So, after forwarding the packet through a tunnel to the appropriate site exit, the router generates a Site Exit Redirection ICMP message to inform the originating host about the correct site exit router for this destination and source address combination.

7.1.2.1.1 Site Exit Redirection ICMP message

Routers send site exit redirect packets to inform a host of a better exit path on the path to a destination.





IP Fields:

Source Address: An address assigned to the router from which this message is sent.

Destination Address: The Source Address of the packet that triggered the redirect.

Hop Limit: 255

Authentication Header: If a Security Association for the IP Authentication Header exists between the sender and the destination address, then the sender SHOULD include this header.

ICMP Fields:

Type: TBD, IANA

Code: 0

Checksum: The ICMP checksum. See [14].

Prefix length: The length of the source address prefix, in bits, expressed as a 16 bit integer, transmitted in network order, i.e. most significant byte first.

Redirection lifetime: The number of seconds during which the redirection should remain in effect, expressed as a 16 bit integer, in network byte order.

Site Exit Address: An IP address of the preferred exit router to use for packets sent using as source address the IPv6 destination of this packet, or using any source address whose prefix matches the first "prefix length" bits of this packet's IPv6 destination.

Possible options:

Redirected Header: As much as possible of the IP packet that triggered the sending of the Redirect without making the redirect packet exceed 1280 octets.

7.1.2.1.2 Host behavior

Hosts can be programmed to perform "exit router discovery", i.e. associate to a source and destination address pair the address of the preferred exit router, and then tunnel packets directly to that exit router. Hosts will learn the address of the exit router through ICMP "Site Exit Redirection" messages.

Any redirection message poses a potential threat, since it can be used by third parties to misdirect and possibly capture traffic. In a secure set-up, hosts will establish a security association with the exit routers, and will only accept Site Exit Redirection messages that are properly secured by an authentication header. In the absence of a security association, the host may perform a number of checks before accepting a Site Exit Redirection ICMP message:

* check that the IPv6 source address corresponds to a local prefix;

* check that the prefix length has a realistic value, e.g. at least 48 bits;

* check that the Site Exit Address matches the site prefix being redirected;

* select a redirection life time that is the minimum of the ICMP value and a locally selected maximum duration.

Since site exit discovery is a routing optimization, hosts should balance the routing gain with the possible security risk.

7.1.2.2 Rapid reaction to failures

7.1.2.2.1 Direct link failures

In order to react to local failures, 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.

This solution is sufficient when the site is composed of a single link; for more complex site, we propose to use the "router renumbering" mechanism to maintain an up-to-date list of available prefixes.

7.1.2.2.1.1 Use of Router advertisements

The router advertisement messages are defined in [3]; their use for address configuration is defined in [4]. As specified in [3], 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 [3], 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.

Manipulating the preferred lifetime only solves part of our problem, since according to [3] the hosts should continue to use the "valid"

source address in existing communications. To actually maintain the transport sessions that used the now unavailable link, we will need additional host improvements.

7.1.2.2.1.2 Use of Router Renumbering

In order to advertise a null preferred lifetime for a specific prefix, the sites routers must be able to learn about that prefix. A possibility is to use the "Router renumbering" protocol [6][RFC2894] to pass this information. The protocol allows an authorized agent, in that case the egress site, to update the list of prefixes advertised by the site's routers. The protocol can be used to change parameters associated to a prefix, such as the preferred lifetime.

7.1.2.2.2 Reaction to other failures modes

Based on the analysis presented in section 5 and 6, we recommend the adoption of a host based exit path selection mechanism to enable the hosts within the multihoming site to react to topology changes.

Considering that we are recommending a solution for the site exit issue based on source address dependent routing, we can assume that the exit ISP is determined by the source address included in the packet. So, in order to force the routing of a packet through a particular ISP, the host only has to set the appropriate source address. As described in section 5, the proposed mechanism to identify available paths will be based on the trial and error

procedure.

The following mechanism is to be implemented in host in order to react to topology changes.

7.1.2.2.2.1 Exit Path Cache

An Exit Path Cache is created in the hosts. Each entry contain for a Destination Address, information about the corresponding Source Address and a lifetime.

Exit Path Cache entry creation process:

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. (would this break some apps?)

- 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 certain period to avoid that multiple answers used in the next section produce suboptimal behavior. (the other option would be not to modify existent (valid) cache entries when packets with a different DA are received)

7.1.2.2.2.2 Initiating new communications

When a host attempts to initiate a communication with a certain destination address D, it first verifies if an Exit Path Cache entry exists for that destination address D. If it does exists, the host obtains the source address to be used form it.

If no entry exists for that destination address D, the host generates multiple packets, one per available source address, sends them and sets a timer.

If a reply packet is received, the cache entry is created as described in the previous section. Following packets addressed to that destination will use the discovered source address if the applications does not sets the source address to be used.

If the timer expires before any packets containing D as source address are received, this may mean that there is no exit path available to reach destination D or that the application generates an unidirectional flow, so no packets are to be received. In any case, the issue cannot be addressed at this level, so the recommended behavior is that the host simply selects one source address S and use it for the packets addressed to destination D. In order to avoid the

procedure to restart, a Exit Path Cache entry has to be created for this destination address, containing the selected source address.

7.1.2.2.2.3 Preserving established communications

TBD

7.2 Multihoming solution for medium sites

Medium sites are likely to be capable of running BGP but they may not be able to obtain enough trust from their ISP to relax the source address checks. So, medium sites could use the mechanisms proposed for small sites, but they are likely to benefit by integrating BGP in the multihoming solution.

So, the recommended integration of BGP capabilities in the proposed small site solution basically affects the mechanism used to react to topology changes affecting non-direct links.

This means that the solution for the site exit issue recommended for medium sites is also a mesh of tunnels as presented in section 7.2.1 allowing a smooth transition to multihoming without interfering with the installed base within the multihomed site. Also, we recommend the usage of Neighbor Advertisement and Neighbor Renumbering to convey information about direct link outages by deprecating the correspondent prefix, as presented in section 7.2.2.1.1.

7.2.1 Reaction to topology changes

The proposed solution requires that the site exit routers run BGP with their correspondent providers. By doing so, exit router have information about reachable destinations through their directly connected ISP. Moreover, through IBGP sessions with the other site exit routers, they have information about reachable destination through the other ISPs.

An Exit Path Cache is created in the hosts. Each entry contains for each Destination Address, information about the corresponding Source Address and a lifetime.

Exit path Cache entries are created when the host receives a packet

as described in section 7.2.2.1.2.1

So when a host attempts to initiate a communication with a certain destination address D, it first verifies if an Exit Path Cache entry exists for that destination address D. If it does exist, the host obtains the source address to be used from it.

If no entry exists for that destination address D, the host just selects one of the possible source addresses and includes it in the packet.

When the packet reaches the site exit router the following situations are possible:

1. The destination is reachable through this site exit router and its directly connected ISP and the source address contains the prefix corresponding to the connected ISP. In this case, the site router forwards the packet through its directly connected ISP.

2. The source address contained in the packet corresponds to another exit router and the destination is reachable through the other site exit router (the one that corresponds to the source address). In this case, the router tunnels the packet to the correct site exit router and sends a Site Exit Redirection ICMP message (as defined in 7.2.2.1.1) back to the host, so that the host can send following packets directly to the correct exit router.

3. The destination is not reachable through the ISP that corresponds to the source address included in the packet, but it is reachable through another ISP. In this case, this packet has to be discarded and the host has to be informed that an alternative source address has to be used. The router then sends an ICMP Destination Unreachable Error message with code 5 (meaning source address failed ingress policy) back to the host, carrying in the source address the prefix that has to be used to reach the selected destination. A new Exit Path Cache entry is created containing the source and destination address.

4. The destination is unreachable through any of the ISPs, so the packet is discarded and an ICMP Destination Unreachable error message is sent back to the host.

7.3 Multihoming solution for big sites

A big site is likely to have enough expertise and resources available to run BGP. Also, it seems likely that a big site can obtain the required level of trust from its providers to relax the source address checks. So, big sites are likely to adopt a multihoming

solution based on these two mechanisms, the relaxation of source address checks and the usage of BGP and the intra-site routing system to select the exit path.

Source address check relaxation allows a big site to become multi-homing without prejudice to legacy hosts within the multi-homed site. Those hosts can still work properly as if they were in a single homed site.

BGP provides information about what path reaches the selected destination. However, in case that one of the ISPs is down, the corresponding address is unreachable, meaning that such address is not to be used as a source address by hosts within the multihomed site that establish new communications, because there is no route available for return packets. In this case, a similar (but simplified) mechanism to the one proposed in the previous section about reaction to topology changes for medium sites is to be used. This mechanism is simpler because no ingress filtering considerations are involved, so the situation described in point 2 in the section above is no longer relevant.

8. Evaluation of Host centric solution and Multihoming Goals

The MULTI6 working group has elaborated a list of goals for a multi-homing solution that is detailed in [7]. In this section, we will review how the host centric approach to IPv6 multihoming meets these goals, which are distributed in two subsections: matching the capabilities of IPv4 multihoming, and meeting additional goals.

8.1 Capabilities of IPv4 Multihoming

8.1.1 Redundancy

The solution presented here can provide protection against:

- Physical link failure,
- Logical link failure,
- Routing protocol failure,
- Transit provider failure, and
- Exchange failure.

Basic redundancy is provided by the availability of multiple addresses, that can be tried in turn, and by a reliance on classic destination based routing protocols. We assume that if an address is reachable, the routing protocol will find a path that leads to it; at worst, the host will have to perform several transmission trials, using different addresses, until the destination is reached.

On the reverse path, redundancy is based on the selection of an appropriate source address. The "preferred lifetime" mechanism allows even the simplest hosts to learn which addresses are robust enough to be used.

Destination and source selection provide a protection against a failure of the site access link, which is catalogued in the goals as Physical link failure, or Logical link failure. The availability of multiple destination addresses provides a protection against Routing protocol failure, Transit provider failure, and Exchange failure on the forward path: the communication will succeed if at least one of the destination addresses can be routed. The protection against such failures on the reverse path is provided if multiple source addresses are tried.

8.1.2 Load Sharing

An enterprise can distribute the inbound traffic by manipulating the "preference" associated to various addresses in the DNS, e.g. by using mechanisms such as MX records or SRV records.

8.1.3 Performance

Performance enhancements can be obtained by appropriate development of destination address selection and source address selection algorithms.

8.1.4 Policy

Classes of applications may be shifted to a specific provider by appropriate use of DNS records associated to specific services. For example, the NNTP traffic could be directed to the specific server "nntp.example.com", and the enterprise could decide to only advertise for that server an address provided by one of its providers.

The Policy table defined in [15] allows to prefer a certain source address rather than others. Considering that the source address determines the exit path, the policy table allows to express the preferred exit path.

8.1.5 Simplicity

Host centric multihoming is simple to deploy, since it does not require any cooperation between the site and its providers, or in

fact between the various providers. The main requirement is to advertise an up-to-date list of prefixes in the router advertisements; this can be automated using the router renumbering protocol.

8.1.6 Transport-Layer Survivability

TBD

8.2 Additional Goals

8.2.1 Scalability

The host centric multihoming system does not impose any unreasonable requirements on the routing system: the sites use multiple addresses, but each of these addresses can be aggregated under the prefixes of their respective providers.

8.2.2 Impact on Routers

In order to quickly signal to hosts any change in the sites' connectivity, the site routers should implement the "router renumbering" procedures, and the exit routers should be able to use that procedure if a physical or logical link becomes unavailable. Additionally, routers have to implement the new Site Exit redirection ICMP message and the proposed processing of the ICMP destination unreachable error message with code 5 (source address failed ingress policy).

8.2.3 Impact on Hosts

The solution does not destroy IPv6 connectivity for a legacy host implementing [1], [2], [5] and other basic IPv6 specifications current in January 2004. Such hosts may not be taking the full benefit from multihoming; in particular, their transport connections may not survive the failure of a site connection. However, the preferred lifetime mechanism guarantees that after a re- homing event, the new connections of these basic hosts will follow an available path.

Hosts will take better advantage of multi-homing if they implement better destination address and source address selection algorithms, exit router discovery. Each of these is a logically separate function that can be added to existing functions.

The solution does not require changes to the socket API and/or the transport layer; such changes may however be required if the host wants to implement a combined selection of the source and destination addresses, which is an optional additional function. The solution allows host or application change to enhance session survivability.

8.2.4 Interaction between Hosts and the Routing System

The interaction between a site's hosts and its routing system is limited to the normal processing of router advertisements.

Upgraded host will be able to obtain additional information from the routing system through the newly defined ICMP messages.

8.2.5 Operations and Management

It is possible to monitor and configure the multihoming system.

9. Things MULTI6 Developers should think about

This section contains the answers to the questions contained in [17].

9.1 The Answers

9.1.1 Routing

9.1.1.1 How will your solution solve the multihoming problem?

The Host-Centric Multihomig proposal addresses multiple of the multihoming issues. In particular, Host Centric multihoming proposal includes mechanisms to:

- Solve the site exit issue

- Select proper (reachable) addresses when establishing a communication

- Perform policing

The Host Centric multihoming proposal does not includes a proposal to preserve established communications through outages. However, The compatibility of Host Centric multihoming mechanisms with proposed solution to provide transport layer connection survivability will be analysed in future versions of the document.

9.1.1.2 Uniqueness

9.1.1.2.1 Does your solution address mobility?

No.

9.1.2 Identifiers and locators

9.1.2.1 Does your solution provide for a split between identifiers and locators?

No.

9.1.3 On The Wire

9.1.3.1 At what layer is your solution applied, and how?

All the proposed mechanisms work at the IP layer.

Is it applied in every packet?

No.

If so, what fields are used?

Some packets may require to be tunneled to the correct exit router, so an additional IPv6 header may be required.

9.1.3.2 Why is the layer you chose the correct one?

Host Centric Multihoming basically uses tools that are already available in some form in current implementations. While some modifications are required, the goal is to reuse as much of the existent mechanisms as possible. So, the layer used is the layer where these mechanisms already reside.

9.1.3.3 Does your solution expand the size of an IP packet?

The solution does not expand the size of the packet but is uses tunnels in some occasions, so we will analyze the impact of tunnels in fragmentation in this section.

Some packets require to be tunneled to the correct tunnel. Two type of tunneling is used:

- Tunnels between the site exit routers: when packets reach the exit router selected by the intra-site routing system, the exit router verifies whether the source address is compatible with ingress filtering defined by the directly connected provider. If not, the packet will be tunneled to the appropriate exit router. Such tunnelling imposes a reduced MTU. There are two was this can be handled. One option could be to announce a reduced MTU within the site, so that hosts just assume a 20 bytes smaller MTUs always and tunnel overhead doesn't impose additional fragmentation. The other option would be just to let the tunnel endpoint to fragment when needed.

- Tunnels between the host and the site exit router: when the host learns the appropriate exit router through the ICMP Site exit redirection message, the host will tunnel packet directly to the exit router. Again, in this case the host may need to fragment the packets because of the tunnel overhead. It should be noted that packets will only be tunnelled once, whether between exit router or from the host to the exit router. In no case a packet will be tunneled twice because of the multihoming solution. Now, if the option to announce a 20 byte smaller MTU within the site is adopted, it would be desirable that also the tunnels between the host and the exit router can use this reserved space. So an option could be to present the smaller MTU to the upper layers, but allow the tunnel interface to

send 20 bytes larger packets.

Summarizing, the solution will imply a 20 byte MTU reduction within the multihomed site.

This overhead can be eliminated by adopting a source address dependent routing within the site.

9.1.3.4 Will your solution add additional latency?

Small sites: two strategies for detecting available path when initiating communications are presented: sequential retrial of paths or path retrial in parallel. The first approach would impose and additional latency in the case that the first path is not available. The second option would introduce packet overhead but would not increase the latency.

Medium and Big sites: in this case, site exit router will have the information about destination address reachability. In the case that the destination address is not reachable through the ISP corresponding to the selected source address, the packet will be discarded and an increased latency will be generated. However, the introduced latency will reduced since the packet is discarded within the site.

9.1.3.5 Do you change the way fragmenting is handled?

No, see above.

9.1.3.6 Are there any changes to ICMP error semantics?

A new Site Exit redirection ICMP message is defined. see section 7.2.2.1

The processing of the ICMP destination unreachable error message with code 5 (source address failed ingress policy) will be modified according to the procedure described in section 4.3

9.1.4 Names, Hosts, Endpoints, or none of the above?

9.1.4.1 Please explain the relationship of your solution to DNS

Host Centric Multihoming does not introduce a new namespace nor separates locators from identifiers. No changes to the DNS are introduced.

9.1.4.2 If you are not using DNS...

No other mechanism is used.

9.1.4.3 Please explain interactions with 2-faced DNS

No changes are introduced to the DNS.

9.1.4.4 Does your solution require centralized registration?

No.

9.1.4.5 Have you checked for DNS circular dependencies?

No changes are introduced to the DNS.

9.1.4.6 How does a host know its identity?

No new identity is defined. The host learns its IP address by existent mechanisms.

9.1.4.7 What if a DNS server itself is multihomed?

Host Centric Multihomed can be used to provide multihoming benefits DNS. In order to benefit from multihoming, the DNS server has to implement the host mechanisms, just as any other host within the multihomed site that benefits from Host Centric Multihoming.

9.1.4.8 What additional load will be placed on DNS servers?

None.

9.1.4.9 Any upstream provider support required?

for small sites: none

For medium sites: running BGP with the site

For big sites: relaxing ingress filtering, running BGP with the site

9.1.4.10 What application/API changes are needed?

None. It is assumed that current applications are RFC 3484 compliant

9.1.4.11 Is this solution backward compatible with old IP version 6?

Yes.

Can it be deployed incrementally? Please describe how.

Incremental deployment is the major goal of the Host Centric Multihoming proposal. In order to enable incremental deployment, the following roadmap is proposed:

1- The first step is to preserve at least single homing functionalities when a single homed host that becomes multihomed. When a single site becomes multihomed, the site exit issue affects the communications of the hosts of the newly multihomed site. Establishing a mesh of tunnels between the site exit router in the case of the small and medium sites and relaxing the source address checks in the big sites overcomes this problems without imposing a general equipment upgrade. Moreover, the proposed solution only requires configuration of the specific devices. After this step, all the hosts within the multihomed site will work at least as if they were in a single homed site.

2- The second step is to enable some of the multihoming benefits with minor modifications. This would provide some degree of fault tolerance when the direct link between the site and its direct providers fails.

3- The third step is to enable most of the fault tolerance capabilities by upgrading the hosts to select the proper path. This step requires the upgrade of the hosts within the multihomed site.

Does your solution impose requirements on non-multihomed/non-mobile hosts?

No. The changes required by the solution are limited to the multihomed site.

What happens if someone plugs in a normal IPv6 node?

The normal IPv6 would work normally in the multihomed site as if it were in a single homed site and it will also obtain some multihomed benefits.

9.1.4.12 Is your solution backward compatible with IPv4?

No. The proposed solution only works with IPv6.

9.1.4.13 Can IPv4 devices take advantage of this solution?

No.

9.1.4.14 What is the impact of your solution on different types of sites?

How are single homed sites impacted?

No impact.

How are small multihomed sites impacted?

The proposed solution for small sites is customized for their special needs. It doesn't requires complex management (like BGP) nor lot of resources. It is basically host based and does not requires much configuration.

How does it scale for large multihomed sites?

For large sites a different solution is presented that requires additional expertise and resources but enables a higher degree of centralized control.

What about ad-hoc sites such as an IETF event?

If the hosts are upgraded to support the mechanism used in the multihomed site, they would obtain the multihomed benefits. If non-upgraded hosts are connected, they will obtain a service slightly better than the one offered in a single homed site.

9.1.4.15 How will your solution interact with other middleboxes?

Just as regular IPv6 does.

9.1.4.16 Are there any implications for scoped addressing?

No changes are introduce to the address architecture, so it is expected that the proposed architecture will interact with scoped addressing just as regular IPv6.

9.1.4.17 Are there any layer 2 implications to your proposal?

No changes to the interaction with layer two are required. However, depending on how easily outages are detected, the performance of the solution may vary. For instance if direct link outages are rapidly detected, the correspondent prefix will be sooner deprecated and the performance of the solution will increase.

9.1.4.18 Referrals

Referrals can be handled just as in regular IPv6. However, if

multihoming benefits are expected, the referral should include all of the IP addresses assigned to the host within the multihomed site, so that the receiver of the referral can try with the different addresses in case of failure.

9.1.4.19 What new information should applications be aware of?

None

9.1.4.20 Legal Stuff

None

10. Security Considerations

The use of a site exit redirection ICMP message could potentially be used to redirect and intercept traffic; secure hosts should only accept such messages if they are properly authenticated.

11. IANA Considerations

This document requests allocation by IANA of 2 new ICMPv6 message types.

12. Acknowledgements

This memo incorporates text from a previous draft submitted by Richard Draves.

We acknowledge Alberto Garcia Martinez, Cedric de Launois, Brian Carpenter, Dave Crocker and Xiaowei Yang for their reviews and comments.

References

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

[2] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1998.

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

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

[5] Gilligan, R., Thomson, S., Bound, J. and W. Stevens, "Basic Socket Interface Extensions for IPv6", RFC 2553, March 1999.

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

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

[8] Crawford, M. and C. Huitema, "DNS Extensions to Support IPv6 Address Aggregation and Renumbering", RFC 2874, July 2000.

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

[10] Thomson, S. and C. Huitema, "DNS Extensions to support IP version 6", RFC 1886, December 1995.

[11] Johnson, D., "Mobility support in IPv6", Internet Draft , June 2003.

[12] van Beijnum, I., "BGP", OReilly , 2002.

[13] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[14] Conta, A. and S. Deering, "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", RFC 2463, December 1998.

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

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

[17] Lear, E., "Things MULTI6 Developers should think about", ID draft-lear-multi6-things-to-think-about-01, December 2003.

[18] Gupta, M., "Message about new ICMP code points", IPv6 list message http://www1.ietf.org/mail-archive/working-groups/ipv6/current/msg01431.html, February 2004.

Authors' Addresses

Christian Huitema
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052-6399
USA

Phone:
EMail: huitema@microsoft.com
URI:

Richard Draves
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052-6399
USA

Phone:
EMail: richdr@microsoft.com
URI:

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

Full Copyright Statement

Acknowledgment

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