Traffic Engineering in Multihomed Sites

Marcelo Bagnulo, Alberto García-Martínez, Carlos Jesús Bernardos,
Isaías Martínez, Arturo Azcorra
Universidad Carlos III de Madrid
Avda. Universidad, 30 Leganés 28911 – Madrid – España
Tel: +34 916248859 Fax: +34 916248749
{marcelo, alberto, cjbc, imyelmo, azcorra}@it.uc3m.es

Abstract

It is expected that IPv6 multihomed sites will obtain as many global prefixes as direct providers they have, so Traffic Engineering techniques currently used in IPv4 multihomed sites will no longer be suitable. However, Traffic Engineering is required for several reasons, and in particular, for being able to properly support multimedia communications. In this paper[1] we present a framework for Traffic Engineering in IPv6 multihomed sites with multiple global prefixes. Within this framework, we have included several tools such as DNS record manipulation and proper configuration of the Policy Table defined in RFC 3484. To provide automation in the management of Traffic Engineering, we will analyze the usage of two mechanisms to configure the Policy Table.

1. Introduction

As the number of organizations relying on the Internet to conduct their business continues to grow, more and more sites are protecting their global connectivity through multihoming, i.e., attaching to the Internet through several providers. In addition to enhanced reliability, multihoming can also be used to improve the communication through Traffic Engineering, i.e., the proper selection of the path used to forward packets towards different destinations. Support for multimedia communications would benefit from – and sometimes require – the ability to define and implement Traffic Engineering (hereafter TE) policies. In IPv4, multihoming is achieved by announcing the site's address block through all its providers using BGP, and the provision of fault tolerance capabilities and TE can be performed through proper BGP manipulation. However, this approach presents limited scalability, since each multihomed site contributes with one route to the global BGP routing tables. In order to preserve routing system scalability, the usage of Provider Aggregation of addresses [2] is proposed. Such approach implies that multihomed sites will no longer obtain a single prefix but that they will obtain one global prefix per each of their providers. In this scenario, each path to the multihomed site is bounded to the prefix delegated by the correspondent ISP, so the selection of the prefix used will determine the path used to reach the multihomed site. Consequently, TE will be heavily related to address selection mechanisms. In this paper we will present a framework composed of various tools that can be successfully combined for providing TE capabilities in IPv6 multihomed sites that have multiple global prefixes configured. We will also analyze the resulting capabilities, comparing them to the ones available in the current IPv4 multihoming solution.

The rest of this paper is structured as follows: in section 2 we present the IPv4 multihoming approach and its capabilities. In section 3 we discuss the challenges imposed by the adoption of Provider Aggregatable (PA) addressing in multihomed sites, and in section 4 we present a set of tools to provide TE in IPv6 multihomed sites. We finish with the conclusions.

2. Traffic Engineering in IPv4 multihomed sites

In IPv4, the most widely deployed multihoming solution is based on the announcement of the site prefix through all its providers. In this configuration, the site S obtains a Provider Independent (PI) prefix allocation directly from the Regional Internet Registry. Then, the site announces this prefix to its providers using BGP [3]. The multihomed site's providers

[1] This work has been partly supported by the European Union under the E-Next Project FP6-506869 and by the OPTINET6 project TIC-2003-09042-C03-01

announce the prefix to their own providers and so on, so that eventually the route is announced in the Default Free Zone. This mechanism provides fault tolerance capabilities, including preserving established connections throughout an outage. In addition, the following TE tools are available to the multihomed site:

TE mechanisms for outgoing traffic: For outgoing traffic, multihomed sites use BGP attributes to express TE considerations. Essentially, the Local Preference attribute is set accordingly to the site TE requirements, so that preferred routes are selected when they are available in order to reach selected destinations. In this scheme, TE is determined and enforced by the routers through manual configuration. Hosts are not involved in TE.

Traffic Engineering mechanisms for incoming traffic: For incoming traffic, multihomed sites can inject a combination of routes to the interdomain routing system that includes several more or less specific prefixes referring to their own addresses. Less specific prefixes provide fall back routes, in case that more specific routes are not available. More specific routes express TE policies, so that traffic for these more specific prefixes is routed through the desired path. In addition, a multihomed site can somehow influence part of the path through which packets will flow to the site using AS path prepending, so that one of the paths to the same prefix seems less attractive than the other ones. However, it must be noted that even if this procedure is used, the ultimate decision belongs to the sites that are forwarding the packet, since they can select a path with a longer AS path. In this case the TE capabilities also reside in the routers, and hosts cannot influence the path used.

While the presented IPv4 multihomed solution provides fairly good features regarding to fault tolerance and TE, it presents very limited scalability properties with respect to the interdomain routing system. Because of the usage of PI addressing, each multihomed site using this solution contributes with routes to the Default Free Zone routing table, imposing additional stress to already oversized routing tables. For this reason, more scalable multihoming solutions are being explored for IPv6, in particular solutions that are compatible with the usage of PA addressing in multihomed sites, as it will be presented next.

3. Provider Aggregation and IPv6 multihoming

3.1. Multihoming setup with PA addresses

In order to reduce the routing table size, the usage of PA addressing is required. This means that sites obtain prefixes which are part of their provider's allocation, so that its providers only announce the complete aggregate to their subsequent providers, and they do not announce prefixes belonging to other ISP aggregates, as it is presented in Figure 1.

When provider aggregation of end-site prefixes is used, each end-site host obtains at least one IP address from each allocation, in order to be reachable through all the providers, since ISPs will only forward traffic addressed to their own aggregates.

This configuration raises several concerns as it will be presented next.

- Difficulties in the communication in case of failure. When *Link1* or *Link3* becomes unavailable, addresses containing the *PASite* prefix are unreachable from the Internet.

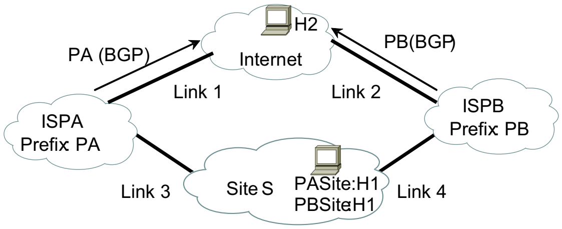

Figure 1. Provider aggregation of end-site prefixes

- Ingress filtering [4] is widely used for preventing the usage of spoofed addresses. However, in the described configuration, its usage presents additional difficulties for the source address selection mechanism and for intra-site routing systems, since the exit path and source address of the packet must be coherent with the path, in order to bypass ingress filtering mechanisms.
- Established connections will not be preserved in case of outage. If *Link1* or *Link3* fails, already established connections that use addresses containing *PASite* prefix will fail, since packets addressed to the *PASite* aggregate will be dropped because there is no route available for this destination. Note that an alternative path exists, but the routing system is not aware of it.

The presented difficulties show that additional mechanisms are needed in order allow the usage of PA addresses while still provide incumbent multihoming solution equivalent benefits. A solution for the first two points is proposed in [5] and multiple solutions for the third point are currently under study [6].

3.2. Multiaddressing and Traffic Engineering

With respect to TE, the multiaddressing configuration depicted above greatly modifies the situation currently available in IPv4.

Ingress traffic (from the multihomed site's perspective).

In the currently deployed IPv4 multihoming solution, each multihomed host normally has a single IP address, and there are multiple paths available in the interdomain routing system to that particular address. TE is then performed by proper selection of the multiple routes available in the routing system for that address. When multiaddressing is adopted, the multihomed site is reachable through a given route/ISP only through the proper prefix, so in order to reach the multihomed site through a given ISP, the correspondent prefix/address has to be used in the communication. This implies that the path used is determined by which address is used among the multiple addresses available for a multihomed host, and that TE capabilities for traffic flowing to the multihomed site will be heavily influenced by the address selection process.

Egress traffic (from the multihomed site's perspective).

In the current IPv4 solution, any of the outgoing paths will forward packets that contain a source address with the prefix assigned to the multihomed site. In the multiaddressing scenario, because of ingress filtering [5], each ISP will only forward packets that carry the appropriate prefix in the source address. So, in order to avoid being discarded by ingress filters, packets will have to flow through the ISP associated with the prefix included in the source address. This implies that the selection of the address of the multihomed host that is used for the communication will also determine the ISP used for outgoing packets.

3.3. Address Selection mechanisms

When multiple PA addresses are available in a multihomed site, the selection of the address of the multihomed host that is used for the communication determines both the ingress and egress path. It becomes relevant then to understand how the address selection is performed. Current Default Address Selection (DAS) algorithms are defined in RFC 3484 [6] and they specify a set of rules and data structures that allows the host to select among the multiple addresses available. We will next present the DAS Policy Table defined in the specification and then both the source address selection procedure and the destination address selection mechanism.

3.3.1. Policy Table. The DAS Policy Table provides the means to express policy considerations when selecting among multiple addresses. It is a longest prefix match table that takes an address (source or destination) as input, and returns two values: a label value and a precedence value. The label value is used to match destination addresses with source addresses. The precedence value is used to select destination address among a set of available destination addresses. The suggested default DAS Policy Table is included in Table 1 [6]

Table 1. Default DAS Policy Table

Prefix	Precedence	Label
::1/128	50	0
::/0	40	1
2002::/16	30	2
::/96	20	3
::ffff:0:0/96	10	4

3.3.2. Source address selection. When the communication is initiated by the multihomed host, the source address selection algorithm [6] will determine which one of the available addresses in the multihomed host will be used. The process is as follows: Once a packet is to be sent to a destination address, the host routing mechanisms will select the interface used for delivering the packet. Then the source address selection algorithm starts with a destination address (D) and the first two source addresses (SA and SB) from a proposed candidate source address set as inputs, and it returns the source address that fits best with the destination address. Successive pair-wise comparisons are performed throughout all addresses in the candidate set to obtain the best one. The algorithm is implemented as an ordered set of rules; if a rule selects one of the two addresses, no further rules are processed. Here we list the proposed rules (Sx refers to any SA and SB):

- Rule 1: If Sx=D then prefer Sx.
- Rule 2: Prefer appropriate scope.
- Rule 3: Avoid deprecated addresses.
- Rule 4: Prefer Home Address.
- Rule 5: Prefer source address of the selected outgoing interface.
- Rule 6: Prefer matching label. Obtain label for SA, SB and D from DAS Policy Table and prefer SA if label(SA)=label(D) and label(SB)<>label(D).
- Rule 7: Prefer public address.
- Rule 8: Use longest matched prefix of Sx with D.

3.3.3. Destination address selection. When the communication is initiated by the external host, this host will probably perform a DNS query to obtain the addresses available for the multihomed host. The destination address selection mechanism [6] of the external host will then select which one of the available addresses of the multihomed host will be used, determining the path used to reach the multihomed site.

When the initiating host receives a set of addresses available for a given destination, it will apply the following set of rules to obtain an ordered list of addresses, which will be delivered to the application. The application will then try with the first address of the list, and if the selected address is unreachable, it should try with the next one in the list.

As said above, when one of the rules succeeds then the remaining rules are not applied. The rules are:

- Rule 1: Avoid unusable destinations.
- Rule 2: Prefer matching scope.
- Rule 3: Avoid deprecated addresses.
- Rule 4: Prefer Home Address.
- Rule 5: Prefer matching label.
- Rule 6: Prefer higher precedence.
- Rule 7: Prefer native transport.
- Rule 8: Prefer smaller scope.
- Rule 9: Use longest matching prefix.
- Rule 10: Otherwise, leave the order unchanged.

4. Tools for Traffic Engineering in multiaddressed multihomed IPv6 sites

As described earlier, the ISP used by packets to ingress to the multihomed site is determined by the address of the multihomed host used as destination address for the communication. Additionally, because of ingress filtering, the ISP used by packets to egress from the multihomed site is determined by the source address set by the multihomed host. So, both ingress and egress ISP is determined by the address of the multihomed host used in the communication. This means that the party selecting the address of the multihomed host to be used during the communication is the party that determines the ISP to be used for the packets involved in this communication. So, TE mechanisms will have to influence such selection. It must be noted that the addresses used in a communication are determined by the party initiating the communication, so in this environment, policy mechanisms will not affect incoming and outgoing traffic separately as in the IPv4 case, but policy considerations will be applied differently to externally initiated communications and internally initiated communications.

4.1. Traffic Engineering for externally initiated communications

When a host outside the multihomed hosts attempts to initiate a communication with a host within the multihomed site, it obtains the set of destination addresses, and it selects one according to RFC 3484. It seems then that the only place where the multihomed site can express TE considerations is through the DNS server replies. The DNS server can be configured to modify the order of the addresses returned to express some form of TE. This mechanism can work fine to provide some form of load balancing and load sharing. The DNS server can be configured so that x% of the queries are replied with an address with prefix of ISPA first and the rest of the times (100-x)% are replied with an address with prefix of ISPB first. When the host receives the list of addresses, it will process them according to RFC 3484 as described above. If none of those rules applies, the list is unchanged and the first address received is tried. Note that the list may be changed by the address selection algorithm because of the host policies, which will affect the address used. In this scenario, it seems reasonable to expect that x% of the externally initiated communications with different destinations will carry their traffic using ISPA and the rest through ISPB. In addition, SRV records [7] can be used to provide more fine grained features, when they are supported by the applications.

4.2. Traffic Engineering for internally initiated communications

For internally initiated communications, the exit ISP is determined by the source address included in the initiating packet. This means that the source address selection mechanism defined in RFC 3484 will determine the exit ISP. RFC 3484 defines a DAS Policy Table that can be configured in order to express TE considerations. Current specification only defines a manual procedure to configure the DAS Policy Table, which is clearly unsuitable perhaps even for small sites. So, a mechanism to provide automatic DAS Policy Table configuration would be required; several possibilities are discussed in the next section.

4.3. DAS Policy Table distribution

In this section we will provide two different approaches to distribute DAS Policy Tables to final hosts to allow automatic configuration. We state the requirements, and then we will present and analyze two solutions, one based on Router Advertisement (RA) and another one based on DHCPv6.

4.3.1. Requirements. The initial general requirements identified for an automatic distribution of the DAS Policy Table are:

- The distribution should rely on standard configuration protocols to facilitate its adoption.
- The distribution of the DAS Policy Table should be atomic, i.e., the complete DAS table is distributed as a unit without partition, since the use of an incomplete DAS Policy Table may result in a behavior different than expected.
- The mechanism should be fully automatic, i.e., not requiring manual configuration on end systems.
- The mechanism should allow the final host to use exclusively DAS Policy Tables that have been locally configured. In this case, the distributed entries would be overridden.
- The mechanism must not interfere with other IPv6 mechanisms.

The currently defined standard IPv6 configuration protocols for end hosts can be classified into stateless [8] and stateful [9].

Stateless configuration is based on the exchange of Router Solicitation and Router Advertisement packets [10] between end hosts and routers within the same link, allowing the distribution of parameters such as IPv6 Prefixes, default routers, etc. This mechanism is fairly simple for both final hosts and routers, but presents some drawbacks. In particular, this approach lacks of capabilities for centralized management, imposing that every change has to be replicated in all the routers within a site through an alternative procedure. In addition, the stateless configuration mechanism does not allow per-host configurations either. Last but not least, this mechanism for distributing configuration data presents some security issues some of which are addressed by the IETF SEcurity Neighbour Discovery Working Group.

Stateful configuration through DHCPv6 [9] allows a host to configure several parameters in addition to the ones currently supported by the stateless configuration method. Additionally, stateful configuration enables the provision of per-host configuration (relevant for the provision of QoS), security mechanisms (e.g., authentication), and the centralized management of the configuration information to ease the administration of an entire site. However, this approach requires the configuration and management of the DHCP server, which may be too demanding for certain environments, for instance unmanaged networks.

Consequently, it is likely that both types of solutions are required in order to satisfy the requirements that can be found in different scenarios.

4.3.2. Router Advertisement Option. The distribution of DAS Policy Tables through Router Advertisements allows the administrator to configure the required parameters easily through a well-known protocol, with an implementation in the router and client.

With this approach, the complete DAS Policy Table is transmitted in a single RA with the format shown next.

<table>
<tr><td>Type (8 bits)</td><td>Len. (8 bits)</td><td>Reserved 0 (16 bits)</td></tr>
<tr><td colspan="3">Lifetime (32 bits)</td></tr>
<tr><td colspan="3">Prefix 1
(128 bits)</td></tr>
<tr><td colspan="2">Label 1 (16 bits)</td><td>Precedence 1 (16 bits)</td></tr>
<tr><td>Pref. Len. 1 (8 bits)</td><td colspan="2">Reserved 1 (24 bits)</td></tr>
</table>

…

Figure 2. RA DAS Policy Table option

The RA option is composed of the mandatory Type and Length fields, followed by a reserved field and a valid lifetime for the complete DAS Policy Table. Next, the entries of the table are specified including the IPv6 prefix and its length, the label and precedence fields.

The RA option must contain the complete DAS Policy Table in order to allow atomic configuration. Therefore it constraints the maximum number of entries that can be distributed by this mechanism to 50, for the minimum 1280-bytes MTU allowed for IPv6. As a consequence, this mechanism, without modification, could only be used to distribute relative small tables. Modifications to circumvent this constraint would violate either the atomicity of the operation or the RA semantic (since no configuration may depend on the results of the processing of other RA options). A preliminary proposal in this subject can be found in [11], which defines a less compact DAS Policy Table entry.

4.3.3. DHCPv6 Option. The format of the DHCPv6 option for distributing DAS Policy Tables is shown in Figure 3.

<table>
<tr><td colspan="2">Option-code (16 bits)</td><td>Option-len (16 bits)</td></tr>
<tr><td colspan="3">Lifetime (32 bits)</td></tr>
<tr><td colspan="3">Prefix 1
(128 bits)</td></tr>
<tr><td colspan="2">Label 1 (16 bits)</td><td>Precedence 1 (16 bits)</td></tr>
<tr><td>Pref. Len. 1
(8 bits)</td><td colspan="2">Reserved 1 (24 bits)</td></tr>
</table>

…

Figure 2. DHCPv6 DAS Policy Table option

The complete DAS Policy Table must be distributed in a single DHCPv6 message, with no size limitation in this case, resulting in an additional benefit.

5. Conclusions

In this paper we have analyzed Traffic Engineering (TE) in IPv6 multihomed environments. In order to preserve scalability of the routing system, the usage of PA addresses is considered necessary in IPv6 multihomed sites. This implies that multihomed sites will have multiple global prefixes, and that multihomed hosts will be then multiaddressed. Such configuration precludes the usage of current TE tools, so new means to provide TE capabilities are required for IPv6 multihomed sites.

In this paper we have presented a set of tools to enable TE in multihomed environments. For externally initiated communications, the usage of "smart" DNS replies is proposed and the resulting capabilities are deemed similar to those achieved with the current solution. For internally initiated communications, the usage of the DAS Policy Table defined in RFC 3484 is considered, and two mechanisms, one based on DHCP and another based on Router Advertisement, are proposed to provide automatic configuration of the tables. Implementations for FreeBSD KAME hosts, since this is the only OS fully supporting DAS configuration, and Linux routers have been developed to perform functional tests. The Router Advertisement implementation is a modification of radvd[2], involving access to the configuration file, new structures to hold the DAS Policy Table data and logic for building the new option. For DHCPv6 testing, a complete C++ implementation has been developed.

The proposed tools provide most of the TE features available in current IPv4 multihoming solution. Moreover, in some aspects, the new solution provides even improved capabilities. For instance, the new solution allows per host configurations (as opposed to IPv4 solution which is configured in a per site basis), and also the DAS Policy Table used in the proposed mechanism is likely to support more fine grained policy expressions.

6. References

[1] G. Huston, "RFC 3221. Commentary on Inter-Domain Routing in the Internet", December 2001.

[2] V. Fuller, T. Li, J. Yu and K. Varadhan, "RFC 1519. Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy", September 1993.

[3] I. Van Beijnum. *BGP*. O'Reilly, 2002.

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

[5] C. Huitema, R. Draves and M. Bagnulo, "Host-Centric IPv6 Multihoming", draft-huitema-multi6-hosts-03, work in progress, February 2004.

[6] J. Ylitalo, V. Torvinen and E. Nordmark, "Weak Identifier Multihoming Protocol (WIMP)", draft-ylitalo-multi6-wimp-00, work in progress, January 2004.

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

[7] A. Gulbrandsen and P. Vixie, "RFC 2052. A DNS RR for specifying the location of services (DNS SRV)", October 1996.

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

[9] R. Droms, et al, "RFC 3315. Dynamic Host Configuration Protocol for IPv6 (DHCPv6)", July 2003

[10] T. Narten, E. Nordmark and W. Simpson, "RFC 2461. Neighbour Discovery for IP version 6", December 1998

[11] C. Patel, "Automated config of address selection policy tables", draft-cpatel-ipv6-automated-policy-table-cfg-00, work in progress, August 2003.

[2] Router Advertisement Daemon for Linux. http://v6web.litech.org/radvd/