# The Case for Source Address Dependent Routing in Multihoming

Marcelo Bagnulo, Alberto García-Martínez, Juan Rodríguez, Arturo Azcorra.
Universidad Carlos III de Madrid

Av. Universidad, 30. Leganés. Madrid. España.
{marcelo,alberto,jrh,azcorra}@it.uc3m.es

**Abstract.** Multihoming is currently widely adopted to provide fault tolerance and traffic engineering capabilities. It is expected that, as telecommunication costs decrease, its adoption will become more and more prevailing. Current multihoming support is not designed to scale up to the expected number of multihomed sites, so alternative solutions are required, especially for IPv6. In order to preserve interdomain routing scalability, the new multihoming solution has to be compatible with Provider Aggregatable addressing. However, such addressing scheme imposes the configuration of multiple prefixes in multihomed sites, which in turn causes several operational difficulties within those sites that may even result in communication failures when all the ISPs are working properly. In this note we propose the adoption of *Source Address Dependent* routing within the multihomed site to overcome the identified difficulties.

## 1 Introduction[1]

Since the operations of a wide range of organizations rely on communications over the Internet, access links are a critical resource to them. As a result, sites are improving the fault tolerance and QoS capabilities of their Internet access through *multi-homing*, i.e. the achievement of global connectivity through several connections supplied by different Internet Service Providers (ISPs). However, the extended usage of the currently available IPv4 multi-homing solution is jeopardizing the future of the Internet, since this use has become a major contributor to the post-CIDR growth in the number of global routing table entries [1]. Therefore, in IPv6 the usage of *Provider Aggregatable (PA)* addressing is recommended for all sites, included multihomed ones, in order to preserve inter domain routing system scalability. While such addressing architecture reduces the amount of routing table entries in the *Default Free Zone* of the Internet, its adoption presents a fair amount of challenges for the end-sites, especially for those who multihome. Essentially, when PA addressing is adopted, a multihomed site will have to configure multiple addresses, one per ISP, in every node of the site, in order to be reachable through all its providers. Such

[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

configuration pose quite a few number of challenges for its adoption, since current hosts are not prepared to deal with multiple addresses per interface as it is required. In this note, we will present how *Source Address Dependent (SAD)* routing can be adopted to deal with some of the difficulties present in this configuration.

The rest of this paper is structured as follows: First we will present the rationale for adopting SAD routing within multihomed sites. Then, we will detail the different configurations of SAD routing that may be required in different sites, including some trials performed, and next we will present the capabilities of the resulting configuration. Finally we will present the conclusions of this work.

## 2 Rationale

### 2.1 Current IPv4 multihoming technique and capabilities

As mentioned above, a site is multi-homed when it obtains Internet connectivity through two or more service providers. Through multi-homing an end-site improves the fault tolerance of its connection to the global network and it can also perform *Traffic Engineering (*hereafter *TE)* techniques to select the path used to reach the different networks connected to the Internet.

In IPv4, the most widely deployed multi-homing 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 [2]. Then the multihomed site providers announce the prefix to its 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: The multihomed site can define which one of the available exit paths will be used to carry outgoing traffic to a given destination by proper configuration of the `LOCAL_PREFERENCE` attribute of BGP [3]. For incoming traffic, the multihomed site can influence the ISP through which it prefers to receive traffic by using the *AS prepending* technique, which consists in artificially making the route through one of the providers less attractive to external hosts by adding AS numbers in the `AS_PATH` attribute of BGP [3] (it should be noted that in this case, the ultimate decision of which ISP will be used to forward packets to the site belongs to the external site that is actually forwarding the traffic).

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 by the multihomed sites, each multi-homed 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 multi-homing solutions are being explored for IPv6 [4], in particular solutions that are compatible with the usage of PA addressing in multihomed sites, as it will be presented next.

### 2.2 Provider Aggregation and Multi-Homing

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 provider only announce the complete aggregate to their providers, and they do not announce prefixes belonging to other ISP aggregates, as presented in figure 1.

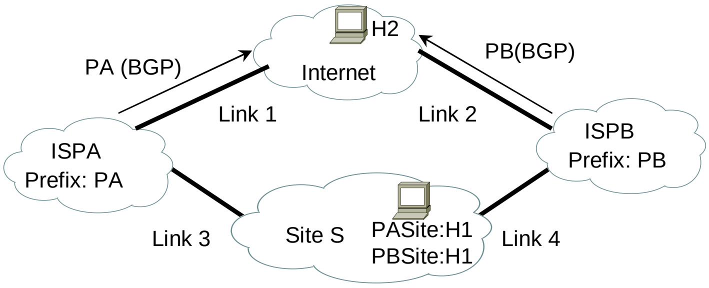

Figure 1: Provider aggregation of end-site prefixes

When provider aggregation of end-site prefixes is used, each end-site host interface obtains one IP address from each allocation, in order to be reachable through all the providers and benefit from multi-homing capabilities (note that ISPs will only forward traffic addressed to their own aggregates).

This configuration presents 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.
- Ingress filtering [5] is a widely used technique for preventing the usage of spoofed addresses. However, in the described configuration, its usage presents additional difficulties for the source address selection mechanism and 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 to allow the usage of PA addresses while still provide incumbent multi-homing solution equivalent benefits. In this note, we will explore the possibility of using Source Address Dependent routing as a tool to help to overcome the identified difficulties.

## 3 Source Address Dependent (SAD) Routing

Source Address Dependent (SAD) routing essentially means that routers maintain as many routing tables as source address prefixes involved, and packets are routed according to the routing table corresponding to the source address prefix that best matches the source address contained in the packet header.

SAD routing can be used to provide ingress filtering compatibility for routing packets flowing from the multihomed site to the Internet. In this case, the source address of the exiting packets has been determined by the host that initiated the communication (the host in the multihomed site, or the external host through the selection of the destination address of the initial packet) and then the routing system will forward the packet to the appropriate exit router in order to guarantee ingress filtering compatibility. The source address selection determines the ISP to be used for routing packets, since, because of address filtering, the source address determines the forward path from the multihomed site to the rest of the Internet, and it also determines the ISP to be used in the reverse path, since the source address used in the initial packets will become the destination address of the reply packets.

Since source address selection implies ISP selection, the adoption of SAD routing will also affect the mechanisms to be used in multihomed sites to define TE. In particular, it will shift TE capabilities from the routing system to the hosts themselves.

We will next evaluate the adoption of SAD routing in two typical multihomed configurations: sites running BGP but without redistributing the BGP information into an IGP, and sites running an IGP to select the exit path. There is an additional possible configuration using static routes in the multihomed site. However, this last configuration is fairly simple and several commercial routers already support it, so we won’t provide a full description of it. Nevertheless, it should be noted that when SAD routing is used, it is possible to obtain fault tolerance and TE capabilities without requiring dynamic routing, since those features are now supported by the hosts themselves and not by the routing system.

In order to enable SAD routing within a site, SAD routing support is not required in all the routers within the site, but it has to be adopted in a connected SAD routing domain that contains all the exit routers [6], as presented in the figure below.

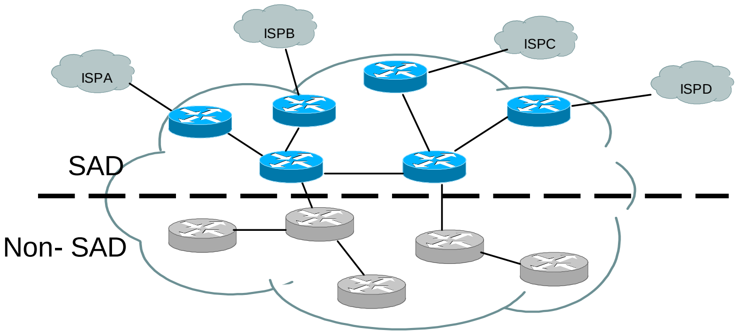

Figure 2: SAD routing domain

Note that it is not necessary for the generic routing domain to be connected, i.e. it can be formed by a set of disconnected domains, all connected to the SAD routing domain.

### 3.1 Sites running BGP but without redistribution of BGP information into IGP.

Current IPv4 multihomed sites usually run BGP with their providers. Through BGP, they obtain reachability information from each of their ISPs. However, because of operational issues, some sites do not redistribute the information obtained through BGP into the IGP [3]. So, in order to be able to properly select the intra site path towards an external destination, they include all the routers that are required to properly select the exit path in the IBGP mesh, including not only site exit routers, but also other internal routers that have access to multiple exit routers. This means that the IBGP cloud is wrapping the non-BGP aware routing domain, as presented in figure 3.

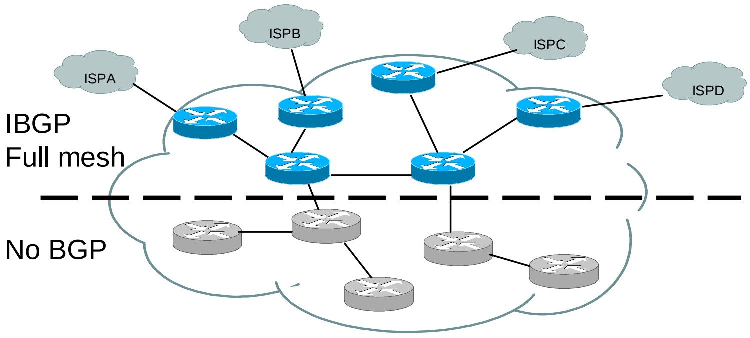

Figure 3: IBGP full mesh

It should be noted that only the IBGP mesh must be connected, and that the non-BGP aware region may be formed by multiple disconnected domains, only linked by the IBGP domain. It is clear that only the routers included in the IBGP mesh need to implement SAD routing in order to properly select the site exit path. So, since all these routers are running BGP, we can use BGP capabilities to provide SAD routing support.

In order to implement SAD routing, each exit router that is running EBGP has to attach a color tag to the routes received from the ISP, so that it is possible to identify the routes learned through each different ISP. Additionally, once the routing information is colored, it is necessary to map each of the colors to a source address prefix. Once that the information of both a given color and its correspondent prefix is available, it is possible to construct SAD routing tables, containing routing information per source prefix.

SAD routing can be implemented in this scenario using the BGP Communities [7] attribute to color the routing information. So, we assume a multihoming scenario where a multihomed site has n ISPs, each one of them has assigned *Pref_i* to the multihomed site, with *i=1,...,n*. In order to adopt SAD routing it is required that:

- First, a private community value is assigned to each different ISP. Therefore, *Com_i* value is assigned to the routes obtained from *ISPi*, being *i=1,..,n*
- Second, *n* routing tables are created in each of the routers involved, so that each router has one routing table per prefix in the site (i.e. per ISP). Additionally each router is configured to route packets containing a source address matching *Pref_i* using the routing table *i*.
- Third, BGP processing rules are configured in each router, so that routes containing a community attribute value equal to *Com_i* only affect routing table *i*.
- Finally, each exit router that is peering with an external router in *ISPi* is configured to attach the community value *Com_i* to all the routes received from *ISPi*, when announcing them through IBGP.

The resulting behavior is that each router within the IBGP mesh will have separate routing tables containing the information learned through each ISP. Packets containing a source address with the prefix of the *ISPi* will be routed using the corresponding routing table.

### 3.2 Sites using IGP

In this scenario, the multihomed site is using an IGP to inform about both internal and external destinations. The IGP learns about external destinations in one of the following three ways:

- Manually configured routes are imported into the IGP
- BGP redistribution into the IGP
- IGP exchange with the providers

As in the previous case, the whole multihomed site routing system is not required to support SAD routing but only a connected domain that has to contain all the exit routers. However, while BGP provides mechanisms to tag routing information so that the same protocol instance can be used to propagate information with different scopes, as presented in the previous section, current IGPs do not provide such capability.

In order to provide SAD routing support, different instances of the routing protocol run in parallel, each one of them associated with a source address prefix. In this way, different instances of the IGP will update different routing tables within the routers. The main difficulty with this approach is to differentiate messages corresponding to the different instances of the IGP. Normally, different instances of the IGP run in different interfaces, so that each instance only receives its own messages. But in this case we want to run multiple instances of the IGP in the same interfaces, so we need a way to separate messages according to the instance of the IGP they belong to.

A possibility would be to send IGP messages using global addresses as source addresses. Usually, IGP messages are sent using link local addresses. But, since each router can be configured with multiple IP addresses, one per prefix, the router includes different source addresses in the messages corresponding to different instances of the IGP. This ships-in-the-night strategy would allow each IGP instance to believe that they are running alone in the link

In particular OSPF for IPv6 [8] explicitly supports running multiple instances in the same link and packets belonging to different instances are identified using the `Instance_ID` field in the OSPF header.

### 3.3 Experimenting with SAD routing

We will next analyze the deployability of the approach by evaluating the available support for SAD routing in current implementations. In order to asses the deployment effort required to adopt the proposed solution, we have built a testbed with widely available commercial routers and we have performed some trials in the framework of the Optinet6 research project. The testbed evaluated the capabilities to support SAD routing of Cisco 2500 routers, Cisco 7500 routers and Juniper M10 routers.

All the tested routers support static SAD routing, i.e. routing based on the source address of the packets according to statically-defined routes. However, the implementation of the SAD routing support differs considerably between them. Cisco IOS supports static SAD routing through manually defined rules, called route-maps, that affect the processing of packets. In order to enable SAD routing, route-maps corresponding to each source address dependent route have to be defined. On the other hand, Juniper routers support multiple routing tables, so that it is possible to create as many routing tables as source address prefixes are involved, and then define the required rules so that the router will forward packets according to the routing table associated with the prefix contained in the source address. In the case of static SAD routing, the multiple routing tables are configured manually with the desired static routes.

Regarding dynamic SAD routing, the support provided by Cisco routers is very limited. Because SAD routing is supported as a manually defined route-map, and because route-map definition is mainly a manual process performed by the router operator, Cisco routers cannot update the routing information (i.e. route-maps) involved in the SAD routing. This means that neither the BGP nor the IGP case are supported by this router vendor.

Because Juniper routers support multiple parallel routing tables, the support for dynamic SAD routing is provided more naturally. In the case of BGP, it is needed that different routing tables are updated depending on the values of the community attribute contained in the BGP route. While this seems pretty straightforward, it is not currently supported by Juniper routers because of the existent constraint that imposes that a given instance of a routing protocol can only update a single routing table, making not viable that the BGP instance can update different routing tables based on the value of the community attribute. Such limitation does not apply for the IGP case, since the considered approach proposes the usage of multiple instances of the IGP running simultaneously, one per source prefix involved, and that each instance of the IGP updates its corresponding routing table. This configuration is currently supported in Juniper routers for OSPFv2 and also for BGP. It should be noted that this approach, i.e. running multiple instances of BGP in parallel, can be used as a temporary solution for the BGP case while the community based approach is not available.

# 4. Resulting Capabilities

## 4.1 Fault Tolerance Capabilities

Since the basic assumptions behind adopting SAD routing for multihoming support are that the source address is determined by the initiating host, and that each source address prefix determines an exit ISP, fault tolerance capabilities will be provided by the hosts themselves. As described in the Host Centric Approach [6], such mechanisms are based on a trial and error procedure. Considering that each source address available in a host is bound to an exit path, the host can try different exit paths by changing the source address. The main difference between the approaches is how fast the host can learn that a destination address is unreachable through the selected path.

When external routes are static, the intra site routing system has no external reachability information, so the packet will be forwarded outside the site and only when it reaches routers that have richer knowledge about the topology it will be possible to determine whether the requested destination is reachable through the selected path. In the worst case, the initiating host will timeout and will retry with a different path.

When the multihomed site runs BGP or an IGP with its providers, reachability information is available closer to the host, i.e. in the site's routers, so in some cases, unreachability will be discovered faster than in the general case, where unreachability information is learned through timeouts. So, the host will attempt to use one of its source addresses to reach a certain destination. The packet will be routed through the generic routing domain to the SAD routing domain. Once there, the routers will determine whether the selected destination is reachable with the selected source address. This means that a route to the selected destination exists in the routing table associated with the selected source address prefix. The possible resulting behaviors are:

- If the selected destination is reachable through the selected source address, then the packet is forwarded towards the site exit router that leads to the ISP corresponding to the source address prefix selected.
- If the selected destination is not reachable through the selected source address, but it is reachable through an alternative source address, then the packet is discarded and an *ICMP Destination Unreachable* with *Code 5* which means *Source Address Failed Ingress Policy* [9] is sent back to the host. The information about the proper source address prefix can be included in this message, for instance in the source address of the ICMP message. The host will then retry using the suggested source address.
- If the selected destination is unreachable, the packet is discarded and an *ICMP Destination Unreachable* is sent back to the host. In this case, the host may retry if an alternative destination address is available.

### 4.2 Traffic Engineering (TE) Capabilities

As a consequence of using multiple prefixes in multihomed sites in conjunction with SAD routing, 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 they will affect packets belonging to externally initiated communications and packets belonging to internally initiated communications differently. This is the first difference with the previous case.

#### 4.2.2 TE for externally initiated communications

When a host outside the multihomed hosts attempts to initiate a communication with a host within the multihomed site, it first obtains the set of destination addresses, then it selects one according to the Default Address Selection procedure [10]. It seems then that the only point 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 constraint.

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. In addition SRV [11] records can be used to provide enhanced capabilities by those applications that support them. When the host receives the list of addresses, it will process them according to RFC3484. If none of the rules described works, the list is unchanged and the first address received is tried first. Note that the list may be changed by the address selection algorithm because of the host policies.

#### 4.2.3 TE 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 [10] will determine the exit ISP. RFC 3484 defines a policy table that can be configured in order to express TE considerations. The policy table allows a fine grained policy definition where a source address can be matched with a destination address/prefix, allowing most of the required policy configurations.

## 4 Conclusions

In this note we have presented the case for the adoption of SAD routing in multihomed environments. The scalability limitations of the current multihoming solution based on the usage of Provider Independent addressing have been largely acknowledged by the Internet community, and there is a consensus that only a new multihoming solution compatible with PA addressing will preserve IPv6 inter domain routing system scalability. However, the adoption of PA addressing in multihomed environments implies that multihomed sites need to internally configure as many prefixes as providers they multihome to, causing several difficulties, such as incompatibilities with ingress filtering, incapability to preserve established connections through outages and so on and so forth. This is basically due to the fact that when multiple PA prefixes are present in the multihomed site, the source address selection process determines the ISP to be used in the communication. This is so because in order to preserve ingress filtering compatibility, the packet has to be forwarded through the ISP that is compatible with the selected source address. Current destination address based routing does not take into account the source address of the packet, making it unsuitable to provide ingress filtering compatibility, that is source address related. SAD routing is then the natural option to overcome the difficulties caused by ingress filtering. Moreover, once that SAD routing is available on the multihomed site, it is possible to obtain additional benefits such as fault tolerance and traffic engineering capabilities with a reduced complexity. SAD routing is not a new technology and it is available in some form in current router implementation, which facilitates its adoption and deployment. However, SAD routing is currently a special feature whose applicability was limited to very specific scenarios. But, if SAD routing is adopted as a fundamental part of the IPv6 multihoming solution as it proposed in this note, it would imply a massive adoption of SAD routing technology, based on the expected number of multihomed sites.

## References

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

[2] Y. Rekhter, T. Li, “A Border Gateway Protocol 4 (BGP-4)”, RFC 1771, 1995.

[3]. I. Van Beijnum, “BGP”, Oreilly, 2002.

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

[5] P. Ferguson, D. Senie, “Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing”, RFC 2827, 2000.

[6] C. Huitema, R. Draves, M. Bagnulo, “Host-Centric IPv6 Multihoming”, Internet-Draft (Work in proress), 2004.

[7] ] R. Chandra, P. Traina, T. Li, “BGP Communities Attribute ”, RFC 1997, 1996.

[8] R. Coltun, D. Ferguson, J. Moy, “OSPF for IPv6”, RFC 2740, 1999.

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

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

[11] A. Gulbrandsen, P. Vixie, L. Esibov, “A DNS RR for specifying the location of services (DNS SRV)”, RFC 2782, 2000.