An Identifier/Locator Split Architecture for Exploring Path Diversity through Site Multi-homing - A Hybrid Host-Network Cooperative Approach

Subharthi Paul, Raj Jain, Jianli Pan
Department of Computer Science and Engineering
Washington University in Saint Louis
{spaul, jain, jp10}@cse.wustl.edu

Abstract
In this paper, we take a fresh look at stub-site multihoming within the paradigms of an identifier/locator split architecture. More specifically, we investigate the possibility of enabling multi-homed stub network sites to improve the performance of their end-to-end TCP flows by leveraging the path diversity of the underlying network. We design a host-network co-operative mechanism for end-to-end flow path switching based on reliable transport layer protocol "hints" indicating probable path problems. Our evaluations of actual Internet routing data strongly suggest significant degree of path diversity across path switches available to multihomed stub networks, even within the restricted precincts of inter-domain policy routing. Additionally, we also address the problems of global routing scalability and inbound traffic engineering control as pertaining to stub-site multi-homing.

Keywords: Multihoming, ID/locator Split, Routing Scalability, Future Networks, Path Diversity, Traffic Engineering

1. Introduction

It is now common knowledge that IP addresses are contextually overloaded. They serve the dual context of identifiers (IDs) as well as locators. Separation of IDs and locators add enormous flexibility to the basic underlying architecture of the Internet. To put it more generically, IDs and locators, with their different *"persistence"* properties optimize the context of their specific functional use. While, ID/locator split architectures represent a broad area of research, in this paper we address the specific problem of stub site-multihoming and investigate a host-network co-operative mechanism for site-multihoming within the paradigms of an ID/locator split architecture.

1.1 Problem Statement

A stub network site, typically represented by enterprise networks, does not provide *"transit paths"* to packets belonging to other networks. Stub sites multihome primarily for: 1) link redundancy – to serve as backup paths for Internet connectivity, and 2) traffic engineering- for load balancing, optimizing price-performance ratios etc. However, lack of inherent support from the underlying Internet design significantly limits the effectiveness of present multihoming solutions and also creates scalability issues. The key issues that we address in this paper are:

A. Issues with "Global routing" scalability: Multihoming, in the present Internet is done through: 1). assigning the stub-site with a Provider Independent (PI) prefix which is then advertized by all its upstream providers into the global routing system, or 2) assigning the stub-site with a Provider Aggregatable (PA) prefix from the address space of one of its providers and then advertising this more specific prefix into the global routing system. Both these mechanisms add a distinct entry into the global routing tables, thus creating scalability issues for the global routing system.

B. Loss of Path Redundancy Information: Even though multiple upstream providers advertise reachability to the PI prefix or more specific prefix of a multihomed stub-site, BGP, the *de-facto* inter-domain routing protocol of the Internet is designed such that each BGP router installs a single route for a particular destination prefix into its FIB (Forwarding Information Base). This route selection is dictated by a set of selection rules enforced by the AS routing policy. Thus, there is no way to leverage the fact that a multi-homed stub site is reachable though multiple paths leading to any of its upstream provider networks.

C. Inbound Traffic Engineering: Multihomed stub-sites may easily engineer outbound traffic by directing them through any of the egress paths. However, even for flows starting inside the stub-site, the stub site has no control over the path on which it receives the inbound traffic. This is an extremely relevant problem, especially because most stub sites, except for content provider sites, generally have more inbound traffic than outbound traffic. In such scenarios, lack of control over inbound traffic jeopardizes the traffic engineering goals of the site. The existing solutions for inbound traffic engineering is predominantly through NAT based solutions or through advertizing longer and hence more specific prefixes across different upstream providers to control incoming traffic on these specific prefixes. While NAT based solutions suffer from usual NAT problems [6][7], the latter solution of advertising longer prefixes hurt the scalability of global routing as discussed in Section 1.1A above.

D. End-to-End Path Diversity: This is not so much of a problem, as much as it is a weakness. The actual physical connectivity of the underlying network provides much more diversity than offered through BGP based inter-domain policy routing protocols. However, our evaluations of actual BGP connectivity data

Proceedings of IEEE International Conference on Communications (ICC 2010),
Cape Town, South Africa, May 23-27, 2010

of the Internet reveal that there still exists significant path diversity available for multi-homed sites to take advantage of. This diversity stems from the fact that an egress path choice by a multi-homed stub-site may result in more than one distinct end-to-end path to the destination.

2. Related Work

ID/locator split architectures [2][4][5][8][11][12] are being actively discussed and pursued at the IETF/IRTF communities to solve the problems of mobility, multihoming, security, etc in the current Internet. However, the discussions in this paper pertain to a specific "Policy Oriented Network Architecture" (PONA)[9] that realizes the explicit separation of ownerships of infrastructure, hosts, data and users through a "three-tier object model" (Figure 1). The bottom tier infrastructure is owned by multiple infrastructure owners. Individual users or different organizations such as DoE, DARPA, Amazon, etc. own the second tier consisting of of hosts. The third tier consists of users and data that may belong to specific organizations. "Objects" belong to "realms." Each realm has a "Realm Manager (RM)." "Realms" overlay objects with a discrete ownership framework. Each object is assigned an ID of the form <Realm ID, Object ID>. The "Object ID" is local to each realm. An identifier/locator split architecture represents a specific instantiation of this abstract model wherein the user/data tier is conflated with the host tier. "Locators" represent "Infrastructure object IDs" while "Identifiers" represent "host object IDs".

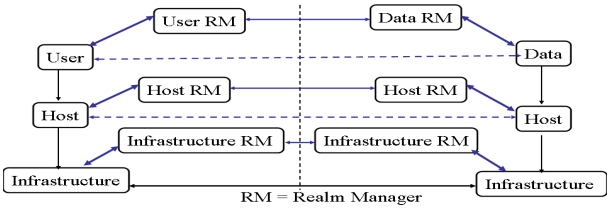

Figure 1. Three-Tier Object model

The host-network co-operative mechanism for multihoming discussed in this paper bears close resemblance to the multihoming mechanism discussed in Six/One [13]. Both, our mechanism and Six/One borrow the concept of stub-site multihoming using provider aggregatable prefixes from each of its providers from Shim6 [5] and extend a purely host-based multi-homing solution in Shim6 to allow a co-operative host-network approach where the network is allowed to override a host selection of an egress path to fulfill the stub-site traffic engineering goals. However, we extend Six/One to include an explicit "passive monitoring module" at the host that monitors end-to-end reliable transport protocol flows for probable upstream path problems and assists a "decision module" to switch source locators to improve end-to-end performance of these flows. Also, through extensive analysis of real Internet routing data we establish the existence of significant diversity in the underlying connectivity, as exposed through inter-domain routing protocols, across "path switches" of end-to-end flows initiated simply by switching the egress path.

3. Solution Approach

3.1 Allocation of Provider Aggregatable Locators: The term "locator", for purposes of the present discussion, refers to 32-bit IPv4 or 128-bit IPv6 addresses that are relieved of their overloaded context as identifiers and, hence, are better designed to optimize the "packet forwarding" function. The 128-bit IPv6 address space is actually shared equally between IDs and locators and hence can support only 127-bit locators. Similar to Shim6 [5], the multihomed stub site is assigned multiple locator prefixes from the locator space of each of its upstream providers. Each upstream provider along with the stub-site network represents a hierarchical infrastructure realm assigning separate "infrastructure realm ID" or locators to each network point-of-attachment inside the stub-site. A single point-of-attachment is thus represented by multiple "service access points" for multiple infrastructure realms and is the basis of our "path diversity" exploration.

3.2 Allocation of IDs: Each host is assigned an ID by each of its host realms. IDs are 128-bit long with the first bit distinguishing it from the locator space. IDs represent the logical organizational hierarchy of the host realm to which the host belongs. Similar to locators, IDs represent "service access point" for host realms. It is beyond the scope of this paper to explore the diversity gains for a host that is assigned multiple IDs from multiple host realms in terms of security, authorization, policy enforcements, etc.

3.3 ID-Locator Mapping Plane: Each host realm is responsible for maintaining an ID-locator(s) mapping and participating in an ID-locator mapping plane.

3.4 End-to-end TCP flows: End-to-end TCP flows bind to 128-bit identifiers. For present applications, only applications that are IPv6 compliant shall be able to bind with IDs. As commonly is the case with all ID/Locator split architectures, this binding remains valid for the entire duration of the end-to-end flow. ID-locator bindings on the contrary are more dynamic. However, changes in ID-locator bindings are transparent to upper layer protocols.

3.5 Network Controlled Traffic Engineering: Traffic engineering is a network function. In this case of multihomed stub-site, a traffic engineering proxy (TE-proxy) is responsible for realizing the traffic engineering goals of the network. As shown in Figure 2, every packet needs to pass through the TE-proxy, which might re-write the source locator of the packet and re-direct it to the proper egress router. The TE-proxy needs to be "*flow-aware*" to avoid per-packet re-direction causing problems related to out-of-order packet delivery. Also, the TE-proxy need not be explicitly implemented as a special router, but it could be a special packet-processing module

co-located in the border routers or built into the IGP logic of the network site.

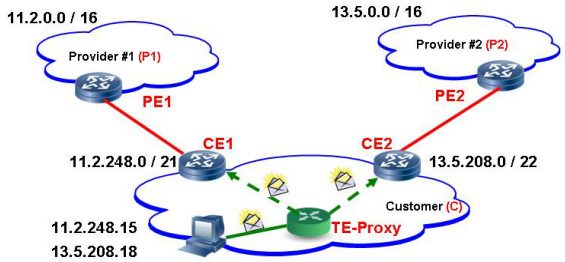

Figure 2. TE-Proxy

3.6 Host-driven Passive Monitoring: As shown in Figure 3, a "shim" layer called the "HID layer" is implemented between the transport layer and the IP layer. This HID layer manages host IDs. It is responsible for maintaining complete isolation between each host realm context within the host, registering ID-locator mappings, update mappings, remapping functions along with modules for specific functions such as mobility, security, policy enforcements etc. For multihoming, the HID layer implements two modules- 1) "*The passive monitoring module*", and 2) "*The decision module*". The passive module maintains state for each end-to-end TCP flow, "*snoops*" on TCP packets and their corresponding "ACKS", and tries to guess the "*path health*" of the upstream path for the flow. Apart from TCP-flows, the passive monitoring module also exposes an interface that enables next generation applications to specify and register their "*flow characteristics*" with the passive monitoring module. In such scenarios, unlike TCP flows, a co-operative flow-monitoring protocol between the source and the destination HID layers may be defined. Thus, the HID layer allows future implementations of transport protocols to outsource their end-to-end flow monitoring to the HID layer which exposes configurable parameters to satisfy specific monitoring requirements of specific applications.

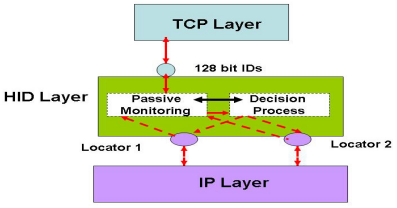

Figure 3. HID "Passive Monitoring Module"

4. Putting it all Together

A. Issues with "*Global Routing Scalability*": Realm hierarchies define aggregation boundaries. As is true for any aggregation, higher degrees of aggregation results in loss of specific contextual information leading to sub-optimal global optimizations. Specifically, for the "forwarding function" deep hierarchies in the infrastructure realms leading to higher degrees of aggregations lead to sub-optimal forwarding, adversely affecting "routing quality". However, for stub-network sites aggregations at the first-hop upstream provider do not effect "routing quality". On the contrary, issues in global scalability owing to multihoming are solved by making the multihomed site participate in the hierarchy of multiple infrastructure realms to which it connects rather than define a separate hierarchy of its own though a provider independent locator prefix or participating in one infrastructure realm hierarchy and advertising itself as a separate hierarchy through use of more specific locator prefixes.

B. Loss of Path Redundancy Information: The stub network-site now explicitly participates in multiple different infrastructure realm hierarchies. Global routing, even through BGP, needs to maintain at-least one separate entry per infrastructure realm hierarchy.

C. Inbound Traffic Engineering: The adverse effect of "aggregation" leading to loss of specific contextual information as discussed in Section 4.A above, actually enables the stub-site to have better control over inbound traffic. Loss of specific information ensures that global routing can forward packets only till the boundary of the hierarchy that is visible to it. Beyond that, packets have to be forwarded along the hierarchy. Since for a stub-network site, aggregations at the first-hop upstream provider do not affect "routing quality", this adversity of aggregation is actually beneficial to the cause of inbound traffic engineering of multi-homed stub sites.

D. End-to-End Path Diversity: To leverage the end-to-end path diversity available through multihoming, an effective "*path switching"* mechanism needs to be developed. The stub-network site does not have end-to-end path information except for the first hop to the upstream provider. For reliable transport protocols, such as TCP flows, end hosts do have limited information about probable path problems through their congestion control mechanisms. These "*path problem hints*" are monitored by the "*passive monitoring module*" at the end host HID layer to aid the "*source locator switch*" decision at the "*decision module"*. This method exploits path diversity to a limited extent only to improve the end-to-end performance of reliable transport protocol flows in the case of path failure/congestions. More aggressive mechanisms of selecting the best end-to-end path for the flow or for improving end-to-end performance of non-reliable transport protocols, require active monitoring through polling. However, if every host within the stub-site starts polling the network for each end-to-end flow, they introduce the same scalability issues as applies to "overlay network routing" [1].

E. Co-operative Host-Network Mechanism: As already discussed, the TE-proxy is allowed to overwrite the source locator in packets for traffic engineering goals of the network site. The *"decision module"* at the host selects the source locator based on "*probable path problem hints*" from the "*passive monitoring module.*" This distribution of decision points clearly suggests the need for an explicit signaling mechanism between the

host and the network. In our mechanism, we resort to per-packet marking by the host to convey its *mode of operation* to the TE-proxy. The different *modes of operation* could be: 1) *Normal*- indicating that the host does not care about which source locator is being used, 2) *Flow performance* – indicating that the host is operating to satisfy some flow requirement, 3) *Congestion*-indicating that the source is operating under probable upstream end-to-end path problems, and so on. These modes of operation are prioritized at the TE-proxy through the policies of the stub-network site and configured to treat each host operational mode differently when taking decisions on traffic engineering.

5. Communication between source and destination multi-homed sites: When, both the source and the destination of the communication session are multihomed, the source "should not" switch destination locators under any circumstances, except: 1) "*remapping*" the destination ID to a new locator through the ID/locator mapping plane, 2) If the destination changes its source locator in return or ACK packets, or 3) explicit signal from the destination allowing the source to change to a specified destination locator. The reason for these principles are: 1) The source can never monitor destination-source end-to-end path through passive monitoring owing to the inherent asymmetry in Internet paths, and 2) The source could jeopardize the traffic engineering policies of the destination.

6. Evaluation: In this section we present the results from our "*feasibility evaluation*" on real Internet routing data collected by "*Route Views*"[10] from several vantage Internet locations.

6.1 Annotate AS relationships: The raw routing data was annotated using Gao's algorithm [3] with AS level relationships ascertaining provider-customer and peering relationships. From this annotated AS data, we could parse for "*probable*" stub network ASs and also their degree of multihoming. The results, as shown in Table 1, suggest that around 90% of the ASs in the Internet are stub networks and more than 50% of them are multi-homed. This suggests that multihoming is an extremely common mechanism in stub networks. Further, 80% of multi-homed stub sites are 2-multihomed, making them candidates for our evaluations.

Table 1. General AS Multihoming Statistics

Total Number of ASs	31868
Total Number of Stub-sites	27035
Total number of Multihomed Stub-sites	15492
Total number of 2-homed Stub-sites	12052

6.2 Global Routing Scalability: Table 2 presents results on address aggregation. These results justify our claim that multihoming causes global routing scalability problems. Around 92% of 2-multihomed stub-sites use locator prefixes that are not aggregatable across any of their upstream providers. Only around 8% use aggregatable locator prefixes from each of their 2 upstream providers. However, we assume that sites that use aggregatable locator prefixes from each of their upstream providers, use it to partition their internal network and assign each partition with a locator prefix from one of its upstream providers.

Table 2. Address Aggregation

Total 2-Multihomed Stubs	12052*
Provider Independent (PI) Address Use	7841
Specific Prefix Advertisement	3222
Use Prefix from Both Providers	989

* Numbers in terms of "Number of ASs"

6.3 "End to end" Path Diversity: End-to-end path diversity, is evaluated across "*source locator switching.*" The three parameters used are: 1) *Path dissimilarity*, 2) *Path length ratio*, and 3) *Path connectivity ratio*. These are defined later in Section 6.3.2.

6.3.1 Data Set: To compute the data-set for evaluating end-to-end path diversity across source locator switches, we introduce two parameters: 1) *Provider connectivity*, and 2) *Provider balance*.

1. *Provider Connectivity: "*Provider connectivity" is the measure of the out-degrees of each provider of a 2-multihomed stub-site. It is calculated as the sum of the provider's providers, peers and non-stub customers. It is a measure of the diversity at the level of a single AS and is indicative of the capability of the AS to bypass local failures and route around them.
2. *Provider Balance:* This metric is associated with the stub-site and is defined as:

$$\text{Provider Balance} = \frac{\min(\text{Connectivity of Provider 1}, \text{Connectivity of Provider 2})}{\max(\text{Connectivity of Provider 1}, \text{Connectivity of Provider 2})}$$

As evident, the ratio is always <1. We conjecture that a low provider balance indicates that the stub-site uses one of its upstream connectivity simply to act as a backup path for link redundancy. A high provider balance indicates that the stub-site actively uses both its up-stream providers for actual data traffic. Figure 4 shows the distribution of stub-site provider balance, indicating that a large percentage of stub-sites do use multihoming simply for link redundancy (low provider balance). For our purposes, we select stub-sites that have medium-high provider balance (0.4< Provider balance <=1) making them the likely candidates that shall use multihoming more actively to improve their end-to-end performance.

6.3.2 Evaluation of end-to-end paths across source locator switching: Suppose that a stub-site is 2-multihomed with two distinct egress paths, S0 and S1. Also suppose that the destination is 2-multi-homed with two distinct ingress paths, D0 and D1. We evaluate the path diversity parameters across the path switches: 1) S0→D0 to S1→D0, and 2) S0→D1 to S1→D1.

We evaluate path diversity of on three parameters:

1. *Path Dissimilarity:* "*Path dissimilarity*" is a measure of the distinctness of the two end-to-end paths across "*source locator switching.*" We selected the shortest AS-length between a source and destination and computed their similarity. Table 3 presents fraction of cases with complete dissimilarity wherein the two paths did not have a single AS in common, thus completely isolating one path from failures in the other path. This is the *most* important parameter and the key motivation behind this paper.

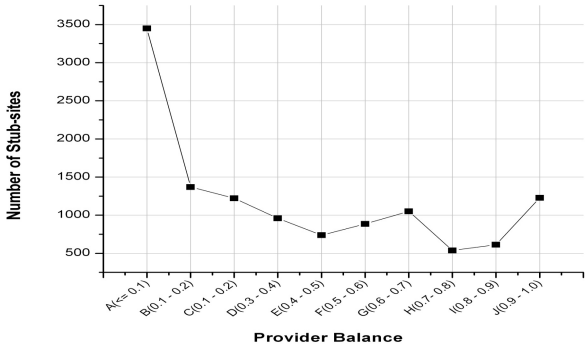

Figure 4. Provider Balance

Table 3. Path Dissimilarity

Path Dissimilarity [S0,S1] → [D0]	0.78
Path Dissimilarity [S0,S1] → [D1]	0.69

2. *Path Length Ratio:* The metric "*Path length*" is predominantly used in the BGP path selection process. "*Path length ratio*" is the ratio of the path lengths of the two paths across "*source locator switching.*" The results in Table 4 indicate that the change in path lengths across "*source locator switching.*" is within acceptable limits to consider path switching in cases failures without severely jeopardizing the flow parameters of end-to-end flows.

Table 4. Path Length Ratio

	Mean	**Std. Dev.**	**Conf. Interval**
[S0,S1] → [D0]	0.8416	0.1551	± 0.03
[S0,S1] → [D1]	0.8580	0.1532	± 0.03

3.Path Connectivity ratio: "Path connectivity" is the sum of the connectivity of all the distinct ASs lying in the path between a source and a destination. "*Path connectivity ratio*" is the ratio of the connectivity of the two paths across "*source locator switching.*" and is a very coarse grained measure of one paths fault tolerance versus the other. The reason for evaluating this parameter is because we do not want to switch paths to avoid failures and choose a new path which is even more prone to failures. This measure would help in determining the strategy in the "*decision process*" to avoid unnecessary path switches and instability.

Table 5. Provider Connectivity Ratio

	Mean	**Std. Dev.**	**Conf. Interval**
[S0,S1] → [D0]	0.8011	0.1585	± 0.03
[S0,S1] → [D1]	0.8234	0.1298	± 0.03

Table 5 shows encouraging results of significant equivalence of Internet paths in terms of *"Path connectivity"* across "*source locator switching.*"

7. Summary

Multihoming has been traditionally used for link redundancy and traffic engineering. In this paper, we leveraged the flexibility offered by a ID/locator split architecture to take a fresh look at stub-site multihoming and explored the possibility of improving the end-to-end performance of reliable transport layer flows through passive monitoring and "source locator switching." Our preliminary evaluations show that even the limited connectivity exposed by BGP-based inter-domain policy routing protocols provide significant diversity in end-to-end paths across switching flows through a different egress path for multi-homed stub sites.

References

[1] David G. Andersen, Hari Balakrishnan, M. Frans Kaashoek, Robert Morris, "Resilient Overlay Networks", Proc. 18th ACM SOSP, Banff, Canada, October 2001.

[2] D. Farinacci, V. Fuller, D. Meyer, and D. Lewis, "Locator/ID Separation Protocol (LISP)," Internet Draft, draft-farinacci-LISP-12.txt, March 2, 2009.

[3]L. Gao. On Inferring Automonous System Relationships in the Internet, IEEE Globe Internet, November 2000.

[4] R. Moskowitz, P. Nikander, "Host Identity Protocol (HIP) Architecture," RFC 4423, May 2006.

[5] E. Nordmark, M. Bagnulo, "Shim6: level 3 multihoming Shim protocol for IPv6," Internet Draft, draft-ietf-shim6-proto-12.txt, February 6, 2009.

[6] T. Hain, "Architectural Implications of NAT," RFC 2993, November 2000.

[7] M. Holdrege, P. Srisuresh, "Protocol Complications with the IP Network Address Translator," RFC 3027, January 2000.

[8] Jianli Pan, Subharthi Paul, Raj Jain, and Mic Bowman, "MILSA: A Mobility and Multihoming Supporting Identifier-Locator Split Architecture for Naming in the Next Generation Internet," Globecom 2008, November 2008.

[9] Subharthi Paul, Raj Jain, Jianli Pan, and Mic Bowman, "A Vision of the Next Generation Internet: A Policy Oriented View," British Computer Society conference on Visions of Computer Science, September 2008.

[10]The RouteViews project, http://www.routeviews.org/

[11] Xiaohu Xu, Dayong Guo, Raj Jain, Jianli Pan, Subharthi Paul, "RANGI: Routing Architecture for Next

Generation Internet," Presentation to Routing Research Group (RRG), Internet Research Task Force meeting, Minneapolis, MN, November 21, 2008, http://www.cse.wustl.edu/~jain/ietf/rangi.htm

[12] Xiaohu Xu, Dayong Guo, "Hierarchical Routing Architecture (HRA)", Proc. 4th Euro-NGI Conference on Next Generation Internetworks, Krakow, Poland, 28-30 April 2008, pp 92-99.

[13] C. Vogt, "Six/One: A Solution for Routing and Addressing in IPv6," Internet Draft, draft-vogt-rrg-six-one-01.txt, November, 2007