Towards A New Internet Routing Architecture: Arguments for Separating Edges from Transit Core

Dan Jen, Michael Meisel
UCLA
{jenster,meisel}@cs.ucla.edu

He Yan, Dan Massey
Colorado State University
{yanhe,massey}@cs.colostate.edu

Lan Wang
University of Memphis
lanwang@memphis.edu

Beichuan Zhang
University of Arizona
bzhang@arizona.edu

Lixia Zhang
UCLA
lixia@cs.ucla.edu

ABSTRACT

In recent years, the size and dynamics of the global routing table have increased rapidly along with an increase in the number of edge networks. The relation between edge network quantity and routing table size/dynamics reveals a major limitation in the current architecture; there is a conflict between provider-based address aggregation and edge networks' need for multihoming. Two basic directions towards resolving this conflict have surfaced in the networking community. The first direction, which we dub *separation*, calls for separating edge networks from the transit core, and engineering a control and management layer in between. The other direction, which we dub *elimination*, calls for edge networks to adopt multiple provider-assigned addresses to enable provider-based address aggregation. In this paper, we argue that separation is a more promising approach to scaling the global routing system than elimination, and can potentially be leveraged to bring other architectural improvements to today's Internet that an elimination approach cannot.

1. INTRODUCTION

A recent workshop report by the Internet Architecture Board (IAB) [16] revealed that Internet routing is facing a serious scalability problem. The current global routing table size in the default-free zone (DFZ) has been growing at an alarming rate over recent years, despite the existence of various constraints such as a shortage of IPv4 addresses and strict address allocation and routing announcement policies. Though the deployment of IPv6 will remove the address shortage, there is an increasing concern that wide-scale IPv6 deployment could result in a dramatic increase of the routing table size, which may exceed our ability to engineer the operational routing system.

A major contributor to the growth of the routing table is site multihoming, where individual edge networks connect to multiple service providers for improved availability and performance [25]. In the presence of network failures, a multihomed edge network remains reachable as long as any one of its providers remains functioning. In the absence of failures, the edge network can utilize multiple-provider connectivity to maximize some locally defined goals such as higher aggregate throughput, better performance, and less overall cost. However, for an edge network to be reachable through any of its providers, the edge network's address prefix(es) must be visible in the global routing table. In other words, no service provider can aggregate a multihomed edge network's prefix into its own address prefix, even if the edge network may be using a provider-assigned (PA) address block. In addition, more and more edge networks are getting provider-independent (PI) address allocations that come directly from the Regional Internet Registries to avoid renumbering when changing providers. In short, multihoming destroys topology-based prefix aggregation by providers and leads to fast global routing table growth.

Routing table size is not the only scalability concern. Equally important is the amount of updates the system must process. Under the current, flat inter-domain routing system, a connectivity flap to *any* destination network may trigger routing updates to propagate throughout the entire Internet, even when no one is communicating with the unstable destination network at the time. Several measurement studies have shown that the overwhelming majority of BGP updates are generated by a small number of edge networks [12, 20]. Unfortunately, a large-scale, decentralized system such as the Internet will surely contain a small number of poorly managed or even suspicious components.

A number of solutions to the routing scalability problem have been proposed, most recently in the IRTF Routing Research Group [1]. Though all the proposals share a common goal of bringing routing scalability under control by removing PI prefixes and de-aggregated PA prefixes from the global routing system, they differ in how to achieve this goal. We observe that all the proposals fall into one of two categories: *separation* or *elimination*. Solutions in the *separation* category insert a control and management layer between edge networks and

today's DFZ, which we refer to as the Internet's *transit core*; edge networks would no longer participate in transit core routing nor announce their prefixes into it. Solutions in the *elimination* category require that edge networks take address assignments from their providers; as a result a multihomed edge network will use multiple PA addresses internally and must modify end hosts to support multihoming.

The purpose of this paper is to compare the two approaches described above and articulate our arguments for supporting the *separation* direction towards routing scalability. Note that, *if fully deployed*, each of the two approaches can be effective in achieving routing scalability in a pervasively multihomed environment. Therefore, our comparison is based on the following high-level criteria to determine the actual impact of a proposed solution: (a) the difficulty in realizing the solutions in the Internet; not only does this involve design issues, but also deployment issues such as the ability to accommodate heterogeneity in the uncertain future, alignment of costs and benefits, and effectiveness in partial deployment; (b) architectural benefits other than scalability – we believe that IP routing and addressing play an essential role in the overall architecture, and that the right kind of changes could help rectify other problems that stem from the same architectural deficiencies.

2. SEPARATION

The root cause of the routing scalability problem facing us today is the fact that all the networks operate in the same routing and addressing space. As a result, edge growth is directly reflected in the core routing table size, and unstable edge networks can flood the entire Internet with frequent updates. The separation approach addresses this root cause by separating edge networks from the transit core in the routing architecture. Generally speaking, Internet service providers (ISPs) fall into the category of *transit networks* who operate in the transit core. The business of transit networks is to provide packet transport services for other networks. End-user sites are generally *edge networks*, which only function as sources and sinks of IP packets. After these two types of networks are separated, edge network prefixes are eliminated from the DFZ routing table. Thus, the DFZ routing table will grow with the number of ISPs, which is much smaller and grows slower compared to that of edge networks. More importantly, the separation enables aggregation of routing announcements on a per-ISP basis. Since most routing dynamics are generated by edge networks, separation will also greatly reduce routing churn in the core. A previous study estimates that removing edge networks from the core routing system can reduce the routing table size and routing dynamics by an order of magnitude [15]. However, due to the absence of edge-

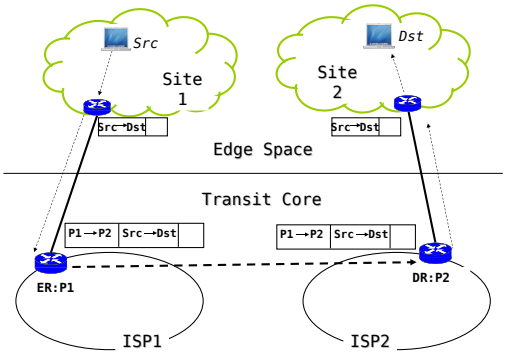


Figure 1: Separation via Map & Encap

prefixes from the DFZ, end-to-end data delivery requires mapping a destination edge prefix to one or more transit addresses that correspond to that edge network's attachment points to the transit core.

One realization of separation is Map & Encap [4, 11], which uses IP-in-IP encapsulation to carry packets across the transit core. As shown in Figure 1, each ISP has border routers that perform encapsulation (*Encapsulation Router or ER*) and ones that perform decapsulation (*Decapsulation Router or DR*). When an ER receives a data packet, it must discover the mapping from the packet's destination address to the corresponding DR address. It then encapsulates the packet and forwards it directly to the DR, who decapsulates and delivers the packet to the final destination. Internal ISP routers or routers connecting two ISPs do not need to understand the encapsulation/decapsulation mechanism; they function the same way as they do today, only with a much smaller routing table.

A number of Map & Encap schemes are under active development and discussion in the IRTF Routing Research Group community, including APT [13], LISP [6], and TRRP [10]. There are also other types of separation solutions besides Map & Encap. For example, Six-One Router [23] and GSE [19] use address rewriting, which rewrites the packet header to include information about the destination's attachment point to the transit core. A common requirement of all the separation solutions is a mapping system that associate an edge prefix with the corresponding transit addresses.

Designing a mapping system is a challenging problem. Because failures of the mapping system can disrupt packet delivery, it is vitally important to make the mapping system robust against failures and attacks. Other issues include the difficulty of handling a large mapping database and the potential overhead and delay introduced by the mapping and encapsulation process. Note, however, that compared with routing data, mapping data has several characteristics that make it easier to scale and secure. First, a piece of mapping data reflects a long-term

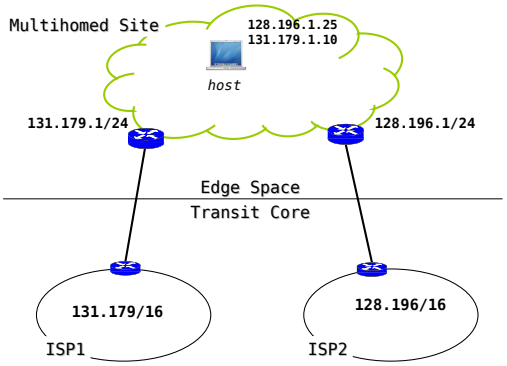


Figure 2: Elimination: Pushing multiple PA addresses to hosts

business relationship, so its changes should occur over a relatively longer time scale (*e.g.*, on a monthly basis). Second, the change of one edge network's provider only affects that edge network's mapping data, whereas link failures in the routing system may affect many prefixes.

Several mapping systems designs have been proposed. APT [13] propogates the full mapping table to each ISP. ERs in each ISP use caching internal mapping queries to deliver data. TRRP [10] proposes to set up another DNS to serve mapping information. On the other hand, LISP has proposed a number of different mapping system designs, including LISP-CONS [2], LISP-ALT [5], and LISP-NERD [14]. LISP-NERD distributes the full mapping table to every ER, while LISP-CONS and LISP-ALT build a DNS-like hierarchical overlay to retrieve mapping data when needed. Each design has its own pros and cons in terms of scalability, controllability, cost, performance and security.

3. ELIMINATION

In order to achieve routing scalability, the elimination approach enforces provider-based address aggregation by eliminating all PI prefixes and de-aggregated PA prefixes. Each multihomed edge network will receive from each of its providers an address block out of a larger, aggregated block announced by the provider. The multihomed site does not inject PI prefixes or more specific PA prefixes into the routing system. Instead, each host in a multihomed site is given multiple PA addresses. For example, as shown in Figure 2, the host obtains two addresses, one from each of its network's ISPs.

In the elimination approach, each host in a multihomed site must be upgraded to understand how to utilize multiple addresses for packet delivery. Each host must also be able to detect and handle potential failures of its upstream connections to its providers. Otherwise, the benefits of multihoming are lost. One elimination scheme, Shim6 [18], proposes to augment the IP layer for this purpose. Shim6 defines a shim sublayer, placed in the IP layer, which ensures that the transport layers at both ends of a given communication sees the same IP identifiers, even though different IP addresses can be used to forward packets along different paths. Prompt failure detection at the IP layer, however, has proven to be difficult and involves a complex tradeoff between overhead, recovery delay, and impact on transport layers [3].

Elimination can also be achieved through multipath transport [9] [21] which can overcome the above-mentioned issues associated with Shim6. Multipath transport works as follows. To communicate with a destination in a multihomed site, a source first uses DNS to find at least one address for the destination. During the initial three-way TCP handshake, the sender and the receiver exchange all of their addresses. The transport layer then creates multiple subflows from all sender addresses to all receiver addresses. Each subflow performs its own congestion control, and subflows may cooperate with each other. That is, if a packet gets dropped due to one subflow being congested, it can be resent on another uncongested subflow. Assuming transport protocols provide reliable delivery, their closed-loop data exchange provides automatic failure detection. At the same time, the use of multiple paths simultaneously reduces the dependancy on any specific paths. By choosing different (source,destination) address pairs, hosts can utilize the end-to-end paths to achieve higher throughput, better performance and faster failure handling.

Multipath transport realization also faces a number of challenges. Being able to effectively gauge the status of multiple paths requires transmitting a large quantity of data and sophisticated subflow control; not all applications can continuously send large quantities of data (*e.g.*, VoIP connections), and not all end points are suited to perform complex control (*e.g.*, small sensors). It also remains an open question whether all multihomed edge sites are willing to handle multiple PA addresses internally and perform renumbering when changing providers. Moreover, since providers announce aggregated prefixes, failures of links to individual edge networks will no longer be reflected in the routing system; thus even long after a link has failed, new nodes may still attempt to use the failed link because individual transport connections detect failures individually. A single failure inside the core may also affect a large number of transport connections, potentially triggering synchronized recovery attempts by all of them. How to make effective use of multiple addresses and how to detect and recover from failures are open challenges when designing an elimination scheme.

4. WHY SEPARATION?

If fully deployed, both the separation approach and the elimination approach can achieve the same goal of routing scalability. However, there are important differences that reveal separation to be a better direction than elimi-

nation towards routing scalability.

4.1 Aligning Cost with Benefits

For any significant change to happen on the Internet, the cost of deployment must align with the benefits of the deployment. Since it is the transit networks that are facing the routing scalability problem, naturally they would have incentive to deploy a solution once it is available. With the separation approach, transit networks can deploy the solution directly and receive the benefits mentioned in the previous section. In other words, the parties responsible for fixing the problem are also the parties who suffer the negative effects if the problem goes unaddressed.

The elimination approach does not change the routing architecture per se; it requires changes of network operations in edge networks and software upgrade at end hosts. At first glance it may appear simpler than the separation approach because it does not need the development of a mapping system. However to remove any of the PI prefixes from the global routing table, the edge networks using PI prefixes must agree to relinquish them and accept PA addresses from their providers instead. The amount of routing table size reduction depends on the number of edge networks that choose to give up their PI prefixes. Under Elimination, transit networks can do nothing but wait for a unanimous action by all the edge networks before the routing table begins to scale. Unfortunately, the routing system has no control over edge site deployment of new solutions. By the time a significant portion of edge sites deploy the new Elimination-based solution(assuming that time ever comes), the routing table may have already grown beyond critical mass.

4.2 Accommodating Heterogeneity

The separation approach has the ability to accommodate heterogeneity in network operations. Different networks have different operational practices and considerations. The elimination approach requires all edge networks to use PA addresses, but some networks may not want to do so – it may cause them trouble in renumbering when they switch providers, or they may not want to give end hosts the ability to affect traffic engineering within their network. Since the elimination approach pushes multiple PA addresses all the way to end hosts, what an edge site does within its network can impact the deployment and effectiveness of the elimination approach. On the contrary, the separation approach is flexible in that it does not enforce any particular operational practices within edge networks. Some may choose to give hosts multiple addresses to improve user experience, while others may choose not to in order to tighten traffic control. Both can be accommodated by the separation approach because what an edge site does within its network will not affect the transit core. The Internet is inherently heterogeneous. A main reason for the success of the original Internet design is its ability to accommodate heterogeneity at many different levels, and we believe we must continue to accommodate heterogeneity in any new architecture.

4.3 Other Architectural Benefits

Separating edges from the transit core provides additional features that are sorely missing in today's Internet. With separation, an end host can send packets through the transit core, but can no longer address a packet to any specific device inside the transit core. Although the separation does not eliminate any specific security threat, it raises the bar against malicious attacks targeted at the global routing infrastructure. In addition, the mapping layer between edge and core networks can serve as a mounting point for badly-needed control and protection mechanisms, and can also act as a cushion layer between the edge and core, allowing each side to deploy innovations without any involvement of the other side. We now elaborate on each of these benefits.

Rolling out new protocols. Internet user innovations don't just happen at the application layer; they also occur at transport and network layers. Intserv/Diffserv, IPv6, and IP multicast, are just a few examples of this. Currently, those innovations require changes to the transit core. In other words, users cannot roll out their new transport and network layer protocols without actions from ISPs which may not have financial incentive to support them.

Separation allows edge networks to develop and deploy new innovative address structures and new protocols. For example, suppose two edge networks $Site_1$ and $Site_2$ could develop a new IPvX address structure. The process of sending an IPvX packet from $Site_1$ to $Site_2$ works as follows. First, the $Site_1$ network routes the IPvX packet to one of its border routers. The router then encapsulates the packet with one of the transit core addresses associated with $Site_2$ (selected by the mapping service). It is essential to note that global agreement on IPvX is not required. Only the mapping service needs to know how to translate an IPvX address to one or a set of transit core addresses.

DDoS mitigation. DDoS attacks abuse the open nature of the Internet architecture by sending attack traffic from multiple compromised hosts to a single, overwhelmed target. In the last few years, a number of efforts have been devoted to developing DDoS mitigation solutions [24].

As described in [17], the DDoS mitigation solution space has become increasingly complex over time. One

critical question is where to install the various traffic identification, filtering and blocking functions proposed by the solutions. Various proposals place the needed functions at the victim, the victim network entry point, some intermediate point along the path, the source network, and/or the source. We believe that the fundamental reason for this diversity is due to the lack of a common architectural framework for solution development. The existing Internet architecture has no convenient hinges or plug-in points where a defense layer could be easily mounted when needed.

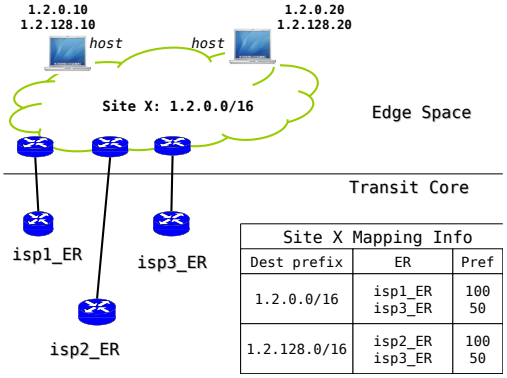
The mapping layer provides such a mounting point. CIRL[7] is one example of approach that leverages the mapping layer. The encapsulation of end-user packets makes it easy to trace attack packets back to the ER, even if they have spoofed source addresses, since the encapsulation header records the addresses of the ER and DR. CIRL lets ERs perform rate-limiting on the traffic going to each specific DR in a way adopted from TVA [24], but without requiring symmetric routing or host changes. Feedback can be provided from DR to ER to adapt the control parameters used for rate limiting.

Ingress traffic engineering. Today, multihomed edge sites already have the ability to forward outgoing traffic to whichever of their providers they prefer. However, edge sites may also want control over their inbound traffic flow for load balancing or using a particular provider only as a backup. Today, edge sites' options are limited – they must resort to prefix splitting and BGP trickery.

Under separation, with the help of the mapping service, an edge site can explicitly express its ingress traffic engineering preferences in its mapping information. For example, say edge site $Site_1$ wants to communicate with multihomed edge site $Site_2$. When packets from $Site_1$ to $Site_2$ enter the transit core, the mapping system will need to select one of $Site_2$'s connections to the transit core as the exit. The mapping system has $Site_2$'s explicit preferences for this selection, and can therefore make this decision based on some combination of $Site_1$ and $Site_2$'s preferences. Though these preferences may be in conflict, this tussle between $Site_1$, $Site_2$, and their respective providers plays out only in the mapping service's selection mechanism. That is to say, this decision takes place at the edges of the network and remains distinct from the specialized problem of transporting packets across the core in the most efficient manner.

5. Separation is Compatible with Multipath Transport

Multipath transport can actually be a great feature for the transport layer. As multihoming (both host multihoming and site multihoming) becomes more and more prevalent, there is an increasing need for TCP to explicitly select among multiple end-to-end paths. For example, TCP may use multiple paths simultaneously to improve throughput or switch from one path to another to avoid congestion or decrease latency. If hosts have multiple addresses, each of which corresponds to a network attachment point, then they can use different (source,destination) address pairs to utilize all available paths.



Site X Mapping Info

Dest prefix	ER	Pref
1.2.0.0/16	isp1_ER isp3_ER	100 50
1.2.128.0/16	isp2_ER isp3_ER	100 50

Figure 3: Adding Multipath Transport to Separation

One misconception is that multipath transport is inseparably tied to the elimination approach. On the contrary, multipath transport is orthogonal to elimination, and can be used with PI addresses under separation as well. Each edge network can split its provider-independent (PI) prefix into multiple, longer subprefixes, mapping each subprefix to different network attachment points (*e.g.*, a provider's router or an Internet exchange point). Those hosts that desire multipath transport are assigned multiple addresses, one from each subprefix. In this way, hosts get multiple source-destination address pairs providing multiple end-to-end transport paths.

Additionally, the use of PI prefixes for multipath transport provides an opportunity for edge site operators to constrain an end user's path selection. Figure 3 illustrates how this can be done. In the figure, $Site_X$ has a PI prefix 1.2.0.0/16 and is multihomed with three providers, $isp1$, $isp2$, and $isp3$. $Site_X$ only intends to use $isp3$ as a backup – that is, $isp3$ should be used only if the link to $isp1$ or $isp2$ fails. However, $Site_X$ would still like to offer its users some degree of path selection. Thus, $Site_X$ simply splits its prefix into two subprefixes, $1.2.0.0/17$ and $1.2.128.0/17$, and assigns each end host two addresses. In the mapping table, $Site_X$ explicitly maps $1.2.0.0/17$ to $isp1$ with $isp3$ as a backup, and maps $1.2.128.0/17$ to $isp2$ with $isp3$ as a backup.

6. Summary

In the last few years a number of research efforts have independently reached, or rediscovered, the same basic idea: add a new *layer of indirection* in routing and addressing [15, 22, 26]. In addition to solving the routing scalability problem, this separation solution offers a number of other advantages explained earlier in the

paper: enabling end-path selection and multipath routing, raising the barrier against malicious attacks to the routing infrastructure, allowing the edges and the core to freely evolve independently from each other, and providing a boundary around the transit core in the form of a mapping service, where various new security and control functions can be easily implemented.

Host-based solutions, such as Shim6 [18] and multipath transport [9], can be used to realize the elimination approach to the routing scalability problem. An ongoing debate in the IRTF Routing Research Group involves whether separation is still necessary, if and once the multipath transport solution is deployed. In this paper, we point out that the current proposal is actually a combination of two pieces: multipath transport for better transport performance as the primary goal, and elimination of PI prefixes for better routing scalability as a consequence. We explained why separation is preferable over elimination to solve the scalability problem, and sketched out how multipath transport can be incorporated into separation solutions.

In his 1928 article, "Being the Right Size" [8], J.B.S. Haldane illustrated the relationship between the size and complexity of biological entities and concluded that, "for every type of animal there is a most convenient size, and a large change in size inevitably carries with it a change of form." We believe that the same holds true for the Internet. It would not have made any sense to have the original routing system design split the network into two parts, core and edges, with the added complexity of a mapping service in the middle. However, the Internet has grown so large over time that it is now technically and economically infeasible to have all IP devices continue to live in the same address and routing space. Hence, a separation, along with a new mapping service, is both necessary and justified.

7. REFERENCES

[1] IRTF Routing Research Group. http://www.irtf.org/charter?gtype=rg\&group=rrg.

[2] S. Brim, N. Chiappa, D. Farinacci, V. Fuller, D. Lewis, and D. Meyer. LISP-CONS: A Content distribution Overlay Network Service for LISP. draft-meyer-lisp-cons-04, April 2008.

[3] A. de la Oliva1, M. Bagnulo, A. Garca-Martnez, and I. Soto1. Performance analysis of the reachability protocol for ipv6 multihoming. *Conference on Next Generation Teletraffic and Wired/Wireless Advanced Networking (NEW2AN 2007)*.

[4] S. Deering. The Map & Encap Scheme for Scalable IPv4 Routing with Portable Site Prefixes. Presentation, Xerox PARC, March 1996.

[5] D. Farinacci, V. Fuller, and D. Meyer. LISP Alternative Topology (LISP-ALT). draft-fuller-lisp-alt-02, April 2008.

[6] D. Farinacci, V. Fuller, D. Oran, D. Meyer, and S. Brim. Locator/ID Separation Protocol (LISP). draft-farinacci-lisp-08, July 2008.

[7] C. Frost and M. Mammarella. CIRL: DDoS mitigation in eFIT. Work in progress, 2007.

[8] J. B. S. Haldane. Being the Right Size. http://irl.cs.ucla.edu/papers/right-size.html, 1928.

[9] M. Handley, D. Wischik, and M. B. Braun. Multipath Transport, Resource Pooling, and implications for Routing. Presentation at IETF-71, http://www.cs.ucl.ac.uk/staff/M.Handley/slides/rpool-rrg.pdf, July 2008.

[10] W. Herrin. Tunneling Route Reduction Protocol (TRRP). http://bill.herrin.us/network/trrp.html.

[11] R. Hinden. New Scheme for Internet Routing and Addressing (ENCAPS) for IPNG. *RFC 1955*, 1996.

[12] G. Huston. 2005 – A BGP Year in Review. APNIC 21, March 2006.

[13] D. Jen, M. Meisel, D. Massey, L. Wang, B. Zhang, and L. Zhang. APT: A Practical Tunneling Architecture for Routing Scalability. Technical Report 080004, UCLA, 2008.

[14] E. Lear. NERD: A Not-so-novel EID to RLOC Database. draft-lear-lisp-nerd-04, April 2008.

[15] D. Massey, L. Wang, B. Zhang, and L. Zhang. A Scalable Routing System Design for Future Internet. In *Proc. of ACM SIGCOMM Workshop on IPv6*, 2007.

[16] D. Meyer, L. Zhang, and K. Fall. Report from the IAB Workshop on Routing and Addressing. *RFC 4984*, 2007.

[17] J. Mirkovic and P. Reiher. A Taxonomy of DDoS Attacks and Defense Mechanisms. *SIGCOMM CCR*, 34(2):38–47, 2004.

[18] E. Nordmark and M. Bagnulo. Shim6: Level 3 Multihoming Shim Protocol for IPv6. draft-ietf-shim6-proto-09, October 2007.

[19] M. O'Dell. GSE – An Alternate Addressing Architecture for IPv6. draft-ietf-ipngwg-gseaddr-00, February 1997.

[20] R. Oliveira, R. Izhak-Ratzin, B. Zhang, and L. Zhang. Measurement of Highly Active Prefixes in BGP. In *IEEE GLOBECOM*, 2005.

[21] P. F. Tsuchiya. Efficient and robust policy routing using multiple hierarchical addresses. *SIGCOMM Comput. Commun. Rev.*, 21(4):53–65, 1991.

[22] P. Verkaik, A. Broido, kc claffy, R. Gao, Y. Hyun, and R. van der Pol. Beyond CIDR Aggregation. Technical Report TR-2004-1, CAIDA, 2004.

[23] C. Vogt. Six/One Router: A Scalable and Backwards-Compatible Solution for Provider-Independent Addressing. In *ACM SIGCOMM MobiArch Workshop*, 2008.

[24] X. Yang, D. Wetherall, and T. Anderson. A DoS-limiting Network Architecture. *SIGCOMM 2005*.

[25] L. Zhang. An Overview of Multihoming and Open Issues in GSE. *IETF Journal*, 2006.

[26] X. Zhang, P. Francis, J. Wang, and K. Yoshida. Scaling IP Routing with the Core Router-Integrated Overlay. *ICNP 2006*.