A Scalable Routing System Design for Future Internet

Daniel Massey
Colorado State University

Lan Wang
University of Memphis

Beichuan Zhang
University of Arizona

Lixia Zhang
UCLA

ABSTRACT

Internet routing is at an important crossroad. The current global routing table, which is largely based on IPv4 addresses, has been growing at an alarming rate over the last few years, despite the constraints by the shortage of IPv4 addresses. IPv6 removes the address shortage problem, however its deployment may potentially further exacerbate the routing scalability challenges facing us today.

In this paper, we first examine and describe the root causes of the routing scalability problem and then discuss a promising direction towards an effective solution. The explosive growth of the Internet over the last decade made it no longer feasible to perform global routing based on all end user IP address prefixes. Yet at the same time, we must preserve the end-to-end model of the Internet architecture. We sketch out a basic approach to an effective solution which is to separate *globally routable addresses* (GRA) from *globally deliverable addresses* (GDA). This separation of address space can simultaneously achieve the goals of improved routing scalability, ease of site-multihoming without using multiple addresses, and elimination of the need for user renumbering when changing providers. An interesting aspect of this approach is that it both facilitates the deployment of IPv6 at edge sites and also does not require immediate changes at large IPv4 deployed bases.

Categories and Subject Descriptors

C.2.1 [**Computer Communication Networks**]: Network Architecture and Design

General Terms

Design

Keywords

Internet routing, scalability

1. INTRODUCTION

Several efforts, including our own, argue that the routing scaling problem necessitates architectural changes and now is an important crossroad when changes can and must be made. In this paper, we first provide a clear description of the root cause of the routing scalability problem. We then point out the direction to solve the problem and articulate the necessity of making the architectural changes as required by the proposed solution.

A recent workshop report by the Internet Architecture Board (IAB) [9] argues that Internet routing is facing a scaling challenge. The current global routing table, which is largely based on IPv4 addresses, has been growing at an alarming rate over the recent years, despite the constraints by the shortage of IPv4 addresses. IPv6 removes the address shortage problem, but one of the major concerns is that a wide IPv6 deployment could potentially cause the routing table size in the default free zone (DFZ) to grow dramatically. This routing scalability concern is exacerbated by an increase in users' requests for provider-independent addresses. Users' desire for more flexibility in changing providers is driving edge networks to seek out provider-independent address alloccations that come directly from the Regional Internet Registries. Our analysis in Section 4 shows that only 11% of the prefixes in a router's routing table belong to the provider network, and the provider-owned prefixes account for only 15% of the total routing updates received by RouteViews' Oregon collector over a recent month. At the same time, the need for effective traffic engineering at both edge networks and ISPs also add potential scaling challenges to the global routing infrastructure by techniques such as announcing sub-prefixes. Overall, the Internet community is presented with a challenge to keep the global routing system scalable in face of an expected growth in the address space, an increased allocation of provider-independent address prefixes, and a demand for effective traffic traffic engineering.

We would like to note that several recent efforts, including our own, all point to the same direction for solutions. However, one fundamental problem is terminology. Without good terminology, similar efforts may appear distinct and other efforts that appear compatible may in fact be orthogonal. We introduce two basic terms for addresses. First, *globally routable addresses* are addresses that appear in routing tables at the DFZ and are only reachable within the DFZ. These are distinct from *globally deliverable addresses*, which must be unique across the network and reachable from anywhere, but does not appear in the DFZ tables. We argue that it is essential to distinguish and separate *globally routable* and *globally deliverable* addresses. This paper illustrates the benefits of such a separation and sketches an approach to achieving the separation, which includes a mapping service to bridge the gap between the two address spaces. An interesting aspect of this approach is that it not only facilitates the deployment of IPv6 at edge sites but

IPv6'07, August 31, 2007, Kyoto, Japan.
Copyright 2007 ACM 978-1-59593-790-2/07/0008 ...$5.00.

also does not require immediate changes at large IPv4 deployed bases. Thus we believe a balance can be achieved that changes the basic architecture while preserving many aspects of the current practice at edge systems. Changing architecture is necessary, but changing practice must be avoided.

Section 2 discusses how we got to this crossroad in routing architecture. Based on this, Section 3 proposes to separate network addresses into two classes and Section 4 shows the benefits achieved by such a separation. While this approach can achieve great benefits, it also raises some new challenges and Section 5 discusses the challenges. The remainder of the paper offers discussion and conclusions.

2. HOW WE GOT WHERE WE ARE

The fundamental goal of the original IP design was to interconnect all packet switched networks so that packets could be delivered from any IP box to any other IP boxes [2]. At the time the basic Internet architecture was sketched out [1], developing an effective technique for multiplexed utilization of all existing networks was the primary goal and continued operation despite partial (physical component) failures was a close second goal. The underlying networks might be based on different communication technologies, but *IP Gateways* were invented to interconnect networks of different communication technologies. All the gateways ran in a single routing domain and they were expected to forward packets for all their neighbors. There was no contractual relationships between neighboring networks.

As the Internet expanded rapidly to a large number of institutions, it was soon realized that a flat routing architecture could no longer keep up with the increasing network scale and management complexity. In EGP [14], the concept of *Autonomous System (AS)* was developed – an AS consists of a group of routers under a single administrative control and runs its own routing protocol internally; a standard inter-domain routing protocol runs between an edge AS and the backbone AS. Later, BGP (as defined in [13]) replaced EGP to accommodate routing policies and more complex peering structure. This has been the basic routing architecture since the late 1980's.

The commercialization of Internet changed the landscape completely. Most importantly, it created a market for global data delivery service. Naturally ASes began differentiation driven by economic forces. The most prominent distinction is the one between *customer networks* (CN) and *provider networks* (PN). Customer networks serve end users directly and are consumers of the global data delivery service. Provider networks have the sole purpose of delivering packets *for a charge*. In return, they hold a *contractual obligation* to the customer networks for providing packet delivery service.

The now pervasive practice of multihoming has also made a profound impact on the scalability of the current routing and address architecture. A multihomed customer can be reached through more than one provider network and business users buy Internet service from multiple providers for improved Internet availability. In the presence of network failures, the customer remains reachable as long as one of its providers remains functioning. In the absence of failures, the customer can use multiple provider connectivity to maximize locally defined goals such as performance, throughput, or cost. However, *multihoming is impacting routing scalability because it essentially destroys topology-based prefix aggregation*. Being reachable through any of its providers implies that the customer must be visible in the global routing table. The customer may even split its address prefix into multiple longer ones to do load balancing on incoming traffic. Regardless of how customers received their prefix allocations (*i.e.*, whether the address prefix is provider allocated or provider independent), the customers' desire to multihome directly conflicts with that of the providers who strive to keep the global routing table size moderate and stable.

The fundamental problem in scaling the current routing architecture is that the architecture still treats every AS equally (*i.e.*, the routing at the Inter-domain level is flat), even though customer networks (*e.g.*, university campuses) and provider networks (*e.g.*, AT&T) have different business models, different growth trends, and different goals in network operations. At the same time, the Internet depends topologically aggregatable address assignments to scale its routing system. However, when users desire multihoming, they inject more address prefixes into the routing system that cannot be aggregated with existing prefixes. Worse yet, flat routing means that a routing flap to any destination triggers routing updates to be propagated to the entire Internet, even when no one communicates with the destination before its connectivity recovers. Both our own measurements and that of others have shown that the overwhelming majority of BGP updates are generated by a very small number of sources, most of them being small edge networks [6, 12]. Failing to accommodate the distinction between customer networks and provider networks is the root cause of the scalability problem facing today's global routing architecture.

The Internet user population has been growing rapidly. This growth will continue into the future (that is why IPv6 adopted such a large address space). It is time to evolve the routing architecture once again to keep up with the growth of the Internet. We believe one effective solution is to separate customers from the backbone routing system. As we explain later in the paper, this separation can bring fundamental benefits to the Internet – routing scalability, ease of site multihoming, security, and added functionality.

2.1 Regarding Locator/ID Separation

As we mentioned above, customer networks' site multi-homing practice has exacerbated the current routing scalability problem. A related concept is *host multihoming*. It used to be the case that most hosts had only one network interface. Today most hosts are shipped with multiple interfaces installed, and if connected, each interface will have its own IP address(es), and none of them is suited as the host identifier, as an address may change at any time, especially given the increasing number of mobile hosts. Site multihoming can lead to host multihoming, if the site gets a different address block from each of its providers. However, the reverse is not necessarily true. A host can have a wireless interface and Ethernet interface with different subnet addresses, but within the same address block assigned to the customer network. Therefore, host multihoming is not itself a contributor to the routing scalability problem, but the multiple addresses make it hard to identify an end system. If there is only one address for a host, the address can be an identifier. Otherwise, we may need a separate host identifier. For example a multi-connected host may have multiple IP addresses, one for each of its interfaces, and it may desire to move a running TCP connection from one interface to another. This would require a host identifier that is independent from IP addresses, such as the one defined by HIP [10]. However, *separating addresses (or locators) from identifiers is not a solution to the routing scalability problem*.

3. THE SEPARATION OF TWO ADDRESS CLASSES

The previous section identified the major causes of today's uncontrolled global routing table growth. In this section we sketch out a promising direction towards an effective long-term solution.

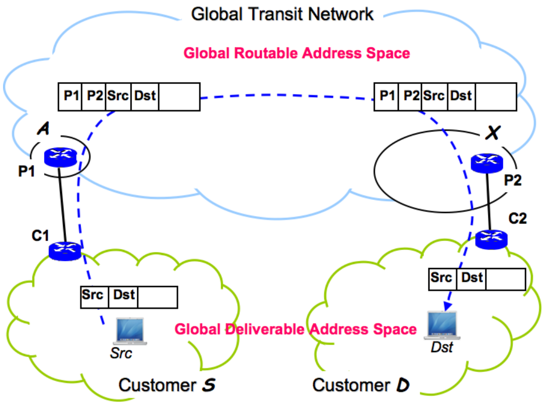

Figure 1: End-to-End Packet Delivery with Address Separation

As the Internet user population continues to grow with pervasive multihoming, the current routing practice of announcing all the end-site prefixes into the global routing system simply does not scale. As eloquently stated by Yakov Rekhter (referred to as ”Rekhter’s Law”), one of the fundamental assumptions underlying the scalability of a routing system can be stated as “Addressing can follow topology or topology can follow addressing. Choose one.” That is, the address prefixes in the routing system should be topologically aggregatable, and aggregated when necessary to keep the routing table size under control. Unfortunately, this desire of prefix aggregation runs into direct conflict with supporting end-site multihoming in the current routing system architecture.

To resolve this conflict, we propose to divide the IP address space into two parts, one part being *globally routable addresses* (GRA), and the other part being *globally deliverable addresses* (GDA). The global routing system runs on the GRA address space. That is, the Internet service providers (ISPs) will all be assigned address blocks from the GRA address space, hence any ISP can reach the prefixes of any other ISPs. Our measurement shows that the number of ISPs in the Internet is relative small compared to the number of end-sites; furthermore, the number of ISPs has been relatively stable over the years, in contrast to the number of end-sites which has been growing rapidly with time. Thus the routing table in the GRA address space is expected to be small and it is unlikely to grow fast. In addition, given the IPv6 address structure has the ISP ID coded in, even in case of excessive prefix de-aggregation to meet traffic engineering needs, an ISP that is one or more AS hops away from the de-aggregation point can potentially use the ISP ID field to re-aggregate the prefix announcements.

All the end-sites will be assigned addresses from *globally deliverable addresses* (GDA). These addresses are globally unique, so that each end-site can be uniquely identified by its GDA prefixes, and each end host can be reached by its GDA address(es). However, because the routing system runs in the GRA space, GDA prefixes will not be globally routable. Instead, one needs a mapping function that matches an end-site’s GDA prefix to the address(es) of its providers’ edge routers, which are in the GRA address space. Figure 1 illustrates how packets between two end hosts are forwarded over the GRA space. When a packet from Src to Dst reaches the edge router $P1$ of the source ISP, the router $P1$ will consult the mapping function and find the GRA address corresponding to the destination ISP’s router $P2$ that is connected to the destination end-site. $P1$ then encapsulates the packet, with its own address as the source and $P2$’s address as the destination in the tunnel header, and send the packet out. When the packet reaches $P2$, it is de-capsulated, forwarded to the end-site, and further forwarded to the destination host based on the GDA destination address of the packet.

We believe that this proposed solution brings several fundamental benefits to the overall architecture. It eliminates the concern of multihoming’s impact on routing system, and removes RIR’s concern regarding the size of site address allocation (i.e. smaller allocation leads to more prefixes in the routing table). In addition to resolving the global routing scalability problem, the GDA address space gives end-site provider-independent addresses. As a result, end-sites can freely change providers without renumbering their networks. The only thing that need to be changed is the mapping information, mapping the site’s GDA prefix to the new provider’s edge routers.

On the surface the encapsulation step in crossing the global backbone, as described above, may bear a remote resemblance to NAT (Network Address Translation). In reality, our proposed solution differs from NAT in two fundamental ways. First, we assign globally unique GDA addresses to all end-sites, thus any end host can *directly* talk to any other end host by simply putting the destination address in the packet. This is in direct contrast to NAT’s limitation where a host behind a NAT is not reachable. Second, this reachability is achieved in the face of individual provider failures as long as any alternative path exists. This is again in direct contrast to NAT’s vulnerability where the failure of a single NAT box will interrupt all the data flows going through that NAT box. Therefore we believe our design helps restore the end-to-end transparency model with robust delivery in the Internet.

As the saying goes, every coin has two sides – our proposed solution brings both benefits and associated costs. In the next two sections, we will first demonstrate the benefits from this separation of GRA and GDA address space, and then briefly mention the new challenges.

4. BENEFITS FROM THE SEPARATION

By separating the IP address space into GRA and GDA, we can significantly improve the scalability and stability of the global routing system, and at the same time remove the site renumbering pain and give end-sites the flexibility to switch providers and engineer their own data traffic. Moreover, the separation raises the bar for malicious users to attack the core routing system.

4.1 Routing Scalability and Stability

In our design, routing in the Internet global transit network (GTN) is only concerned with reachability among transit networks. The number of transit networks is much smaller than that of end-sites. More importantly, its growth trend is expected to be much slower. Although each transit network may announce multiple GRA prefixes depending on its size and traffic engineering practice, the GRA prefixes are *topologically aggregatable*. Moreover, the routing table size in the GTN should be independent from the number of end-sites, or how end-sites may be multihomed, or how end-sites perform load balancing.

Because of the separation of GDA from GRA, routing dynamics occurring inside end-sites or at the border (between end-sites and PNs) will no longer have an impact on the routing stability inside GTN. Also, since the number of prefixes in the GTN is expected to be much smaller than the number of the prefixes in the routing system today, routing convergence would be substantially faster than that of today’s BGP.

To verify the above expectations, we use RouteViews’ August 2006 data (BGP routing table dumps and update logs) to estimate, in the current Internet, how many prefixes belong to provider networks, how many prefixes belong to end-sites, and how many updates the latter generate, respectively.

During our measurement interval, there are a total of 23,021 ASes and 209,549 prefixes in the global routing table. ASes can be classified to either an *end-site*, which only appear at the end of an AS path, or *transit network*, which may appear in the middle of an AS path. Our measurement shows that the number of transit networks is less than 20% of the total ASes in today's Internet, and that the number of transit networks grows at 1/5 of the rate of all the ASes.

All the prefixes originated by end-sites will be allocated from the GDA address space in our design. Prefixes originated from a transit network may belong either to the network itself, or its customers. For example, if a customer network does not have an AS number or run BGP, it prefixes will be announced to the global routing table by its provider. Therefore, for prefixes originated by transit ASes, we need to separate them into transit network prefixes or end-site prefixes. Given a transit AS and one of its prefixes, we compare their descriptive names registered in the WHOIS database. If the two names match, we classify this prefix as a transit network prefix, otherwise a end-site prefix. Of course this approach may not be very accurate due to potential out-of-date records in the WHOIS data. We therefore performed verifications using AT&T's data. We first manually did the name match for all the prefixes originated from AS 7018 (AT&T). We found that only 39 prefixes, out of the 1,501 prefixes announced by AS 7018, actually belong to AT&T itself. The result has been confirmed by AT&T to be very close to the reality, which gives us confidence in our manual approach.

We then implemented a simple heuristic to automate the matching process. The results for some major ISPs are listed in Table-1, along with the results of manual matching. As one can see from the table, the difference between the manual match and automated match is very small in all but one cases. Thus, we applied the automated matching to all the transit networks and their prefixes. The results show that, out of 209,549 prefixes in the global routing table, only 22,733, about 11%, belong to transit networks (although more were originated by them). Next we count the number of updates for transit network prefixes and end-site prefixes respectively during the month of August 2006. Out of 367 million updates from all the RouteViews monitors, only 57 million updates (15%) are for transit network prefixes.

These results show that, if the GRA-GDA separation is applied to today's Internet, both the size of the routing table and the amount of routing churns can be reduced to one order of magnitude smaller than what we have today. Furthermore, the growth of the GRA routing table in the future is expected to be slow due to the slow increase of transit networks compared to the rapid growth of end-sites.

4.2 Site Multihoming and Traffic Engineering

Once we separate end-sites to a separate address space (GDA), naturally the entire GDA address space becomes provider-independent. Customers can change providers freely without renumbering their networks, and can subscribe to as many providers as they want with no negative impact on the global routing table. As a result, our design removes roadblocks for customers to adopt multi-homing, which improves the reliability of their Internet connectivity.

In addition to enhanced network reliability, customers may also want to fully utilize the parallel connectivities provided by multihoming. Since the address space separation between GDA and GRA introduces the need for a mapping function, we can utilize this mapping function for effective traffic engineering support. In addition to the basic goal of mapping a customer address to that of its providers, customers can inject into the mapping record additional *policy* information to facilitate the selection of provider address among multiple alternatives. For example, a customer network C may want to split its incoming traffic between its two providers X and Y. It can specify in its mapping record a preference of receiving 60% traffic via provider X and 40% traffic via provider Y. A sender S learns C's preference through the mapping service. Now S can make an informed decision based on both C's and its own preferences, taking full advantage of multihoming. In the current Internet, on the other hand, there is no effective way for a receiver to influence its incoming traffic paths except by announcing longer prefixes and prefix splitting, which contribute to the routing scalability problem.

AS Number (ISP name)	Total Prefix	Transit Net. Pref (manual)	Transit Net. Pref (automated)
7018 (ATT)	1501	39	35
174 (Cogent)	930	21	19
1668 (AOL)	202	115	100
1239 (Sprint)	852	133	131
701 (Verizon)	4989	537	570
3549 (GBLX)	342	133	81
3561 (Savvis)	521	231	263
3356 (Level3)	514	50	99
209 (Qwest)	691	59	63

Table 1: Prefixes of some major ISPs

4.3 Security Enhancement

Because our design puts all end hosts in an address space separate from that of backbone routers, all user data packets are encapsulated when they cross the backbone. As a result, compromised hosts in the customer space no longer have direct access to the provider infrastructure. Attackers can still use compromised hosts within an end-site to DDoS the local GTN *border* routers, but such attacks only make a local impact and are relatively easy to deal with. Attackers may also use compromised hosts from multiple end-sites (e.g., a botnet) to DDoS the routing infrastructure by flooding packets to some remote end-sites. However, given the GTN topology is opaque to end users, attempting to DDoS any specific component in the provider topology becomes more difficult. Although our design does not eliminate any specific security threat, it raises the barrier against malicious attacks targeted at the global routing infrastructure.

The encapsulation of end-user packets also makes it easy to trace attack packets back to the GTN ingress router even if they have spoofed source addresses, since the encapsulation header records the addresses of the GTN entry and exit routers. In today's Internet, some providers follow the recommended practice and configure border routers to check the source address of packets coming from their customers. However, not all providers implement such ingress filtering, as they do not perceive a direct benefit for the deployment cost.

Although our design makes it difficult to gain unauthorized access to GTN routers, we do not expect GTN to be free of malice. Routers in GTN may still get compromised, and providers in GTN may belong to unknown parties of different interests. Thus, it is necessary to develop effective mechanisms to detect compromised routers and misbehaving transit networks *within GTN*. However, we believe that compromised end hosts are a major source of attacks, and our design raises the barrier against such attacks.

Furthermore, the GTN is expected to be substantially smaller compared to today's Internet, making it much easier to detect and diagnose problems in the GTN.

5. CHALLENGES

With all the benefits from separating the two address spaces, the separation also raises a few challenges in the overall system design and deployment. The essential ones include how to design scalable, secure and efficient mapping function, how to handle the failures between GRA and GDA, and how to conduct network measurement on the Internet backbone after the GRA and GDA separation.

5.1 The Mapping Function

The basic functionality of the mapping service is that, given a destination customer address, it should return a destination provider address so that the packet can be encapsulated and forwarded across the Internet. The mapping service can also be augmented to include traffic engineering information of the receiving network. The mapping service must provide:

- Fast lookup: packets cannot be forwarded until the mapping is completed, so a fast lookup service is essential for good performance.
- Fast failure recovery: mapping entries should adapt quickly with changes.
- Resilience to abuses and attacks: mapping service can be a potential target for attacks. Updates to the mapping service or query replies from mapping service must be authenticated.

We are experimenting with three basic designs for the mapping service. A brute force solution floods the mapping data between GDA and GRA throughout the *providers*. In another approach, one could include mapping data, e.g. the corresponding GRA address for the GDA address queried, in the existing DNS. Finally, a hybrid approach using distributed hash tables also has promise. Overall, there are interesting trade-offs in each approach, but this is primarily engineering problem that can be addressed by modeling and simulations.

The first approach of disseminating the mapping data to every entry router of the backbone can also be done in multiple ways. One possibility is to attach the information to existing BGP routing updates. Another possibility is to run a separate dissemination protocol among provider routers to propagate mapping information. Yet another possibility is to build an overlay network to broadcast or multicast the information. The common advantage of these schemes is that lookup can be done locally at entry routers, therefore the mapping does not incur significant delay in packet forwarding. The disadvantage is that any change of GDA-GRA mapping must be *proactively* propagated *globally*, even if the change may not affect any data traffic (i.e. no one is sending to the affected destination site). Given that the number of end-sites grows at a rapid rate, the dissemination system itself faces a scalability challenge.

The second approach is to provide the mapping service by distributed servers in a way similar to DNS system. The information can be directly included in the DNS system, or otherwise a separate DNS system can be deployed solely for the mappping function, so that entry routers can query the servers for the information needed to forward each packets. The advantages of this approach are that changes are only propagate *locally*, i.e. only the servers responsible for the changes need to be updated, rather than being proactively propagated globally, and that individual responsible parties can selectively enhance *their own* mapping performance through faster servers or more replications. The disadvantage is that the query will add extra delay to packet forwarding. Caching and prefetching popular entries may provide effective performance improvement.

It should be emphasized that the mapping service is a necessary cost for the gains from the separated GRA and GDA address spaces. Up to now, packet delivery relies only on the routing to work correctly. Our design introduces a new dependency, the mapping service, which represents a system cost in providing the mapping function, a performance cost in look up delays, and a target for potential attacks. We would like to point out that the introduction of DNS 20 years ago could be considered remotely analogous: DNS introduced a new dependency in data delivery (DNS name to IP address translation), the cost in providing DNS servers and lookup delays, and a target for potential attacks, which have occurred frequently in recent years. Yet the gain from DNS is essential that its cost is considered necessary tradeoffs. We believe the same arguments can be made for the new mapping service.

Securing the mapping service is essential for our design to succeed. Two types of attacks are of particular concern: denial of service attacks and response modification attacks. By disabling access to the mapping service for a given customer network, one can deny packet delivery to that network. To make the mapping service resilient to DoS attacks, data can be widely replicated and cached so that knocking down one or a few servers does not disrupt packet delivery for a customer, as DNS has demonstrated. The modifications to the mapping replies is also a shared problem with DNS, which may require cryptographic authentication protection. Non-crypto solutions include (1) querying the information from multiple servers for mutual checking, assuming man-in-middle attackers cannot hijack all the queries or replies, and (2) periodic queries to monitor the mapping service in an effort to detect false data.

5.2 Handling Border Link Failures

Our proposed solution separate GRA and GDA address space, so that only topological changes in the GRA space, i.e. inside the global backbone, are handled by the global routing protocols. However, a link between an end-site D and its provider P is not part of the GRA routing space. Thus when this link, or D's router at the other side of the link, fails, no routing update would be generated in the global routing system. This can be viewed as an advantage as it provides the insulation of edge dynamics from the global routing system. At the same time this also introduces a challenge in assuring packet delivery, if the mapping function only reflects which providers D connects to, but not whether the connectivity is up on a real time basis.

Consider the problem of sending from an end-site S to another site D. Assume the packet is forwarded through S' provider router R_1. When the mapping function shows that R_2 is the egress router for the destination site D, R_1 forwards the packet to R_2. Assume that the link between R_2 and D fails. Due to the separation of the two address spaces, router R_1 is not informed of the failure. However when R_2 receives the packet, it cannot forward the packet onto the destination. At this point R_2 may look for alternate route to D (*e.g.*, R_2 could re-encapsulate the packet and forward to another provider for D'), or drop the packet and send an ICMP "Destination Network Unreachable" message to R_1. In the first option, the provider may not be willing to perform the added work of finding and using the alternate paths. In the second option, the sender will not learn of the failure until the ICMP message is received, any packets in transit during this time period may also have to be dropped. It is also possible to have a combination of both approaches: by default R_2 sends a notification message if the link to the destination site has failed; however for a premium price, R_2

may also forward packets along alternate routes as a value-added service. Detailed evaluation needs to be done to fully understand the performance impact and other tradeoffs.

5.3 Network Diagnosis

There has been a growing interest in diagnostic tools that measure path characteristics, detect path changes, and diagnosis network errors; some of our own previous work addresses problems of diagnosing the cause of large scale routing changes [8] and detecting route hijacking events[7]. However the separation of GRA and GDA address space effectively presents end users a black box, which connects up all user networks but does not offer user networks any visibility or influence over the internal paths being used inside the transit backbone.

However end users can still measure the external behavior of this black box, detect any problems that affect their data delivery, and move traffic between different access ISPs. It remains an open research question as whether the tunneling mechanism used to cross the transit backbone should hide all the information about the backbone, or should reveal limited information, such as router hop count (by deducting the appropriate value from the user packet header when it is de-capsulated at the exit of GRA space), to aim the end-to-end network measurement and diagnosis.

6. RELATED WORK

The scalability problem of the existing addressing and routing architecture has long been recognized. Over the years a number of alternate routing designs have been proposed. Recognizing the fundamental conflict between providers' desire for prefix aggregation and user sites' desire for provider-independent addresses, Hinden and Deering proposed ENCAPS [5] in 1996. The basic idea is to separate transit networks and ends sites into two address spaces, and to use IP-in-IP tunnels to carry packets across the global transit networks. Our proposal shares the same solution direction with ENCAPS, so is another more recent effort LISP [3] which sketches out an instantiation of ENCAPS implementation. Our work aims at a long-term architectural design for the Internet, instead of a short-term solution. We try to build support for important goals such as traffic engineering, security and diagnosis in addition to routing scalability in the new architecture, and we plan to explore and compare different designs for various architecture components. We would also like to emphasize that the key to solve the routing scalability problem is to separate the two types of networks (address spaces), which is not necessarily the same as the separation of locators and identifiers depending on their definitions.

In [11], O'Dell proposed a new routing design for IPv6, commonly known as GSE. The basic idea is to divide IPv6's 16-byte address into two parts, with the lower $16 - N$ bytes being the End System Designator (ESD), and the higher N bytes (called Route Goop, or RG) being used for inter-AS routing. The GSE hides a customer site's RG from internal hosts, which is filled in when packets exit the customer site. This late binding can offer several benefits similar to those provided by our approach. HLP [15] compartmentalizes the lower tiers in Internet's topological hierarchy into separate regions, thereby improving the scalability and stability in today's inter-domain routing system. HLP shares a common goal with our approach, in that it isolates edge instability from the backbone core. However, it does not address the scaling issue caused by today's pervasive multihoming practice.

7. DISCUSSIONS

As time goes, multiple solution development efforts have pointed to the same direction of separating end sites and transit networks into distinct address spaces in order to solve routing scalability problem. We believe that this is not coincidental, but rather showing a convincing sign that the separation is a right way forward.

Stepping up a level, we would like to re-emphasize the necessity to evolve the Internet routing architecture once again to keep up with the growing nature of the Internet. To that end we would like to cite a 1928 article by J. B. S. Haldane, "being the right size" [4], where the author illustrated the relation between the size and complexity of biological entities through a vivid example. As stated in the article, "a typical small animal, say a microscopic worm or rotifer, has a smooth skin through which all the oxygen it requires can soak in." However, "increase its dimensions tenfold in every direction, and its weight is increased a thousand times, so that if it is to use its muscles as efficiently as its miniature counterpart, it will need a thousand times as much food and oxygen per day. Now if its shape is unaltered its surface will be increased only a hundred-fold, and ten times as much oxygen must enter per minute through each square millimeter of skin." That is why all large size animals have lung, an organ specialized for soaking oxygen. The author concludes 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." It would be unimaginable for small insects to have a lung, but it is also impossible for big animals to live without a lung.

We believe the same is true for Internet. It would not have made any sense to have the original addressing architecture splitting into two parts with the added complexity of a mapping service in between. However today the Internet customer base has grown to be big enough which makes it both technically and economically infeasible to have all the IP boxes live on the same address space. As Internet grows large in user population size, it is no longer feasible for its transit core to deliver packets by maintaining the reachability information to the billions of end users.

8. REFERENCES

[1] V. Cerf and R. Kahn. A Protocol for Packet Network Intercommunication. *IEEE Trans. Comm.*, 22(5):637–48, May 1974.

[2] D. Clark. The Deisgn Philosophy of the DARPA Internet Protocols. In *ACM SIGCOMM*, pages 106–114, 1988.

[3] D. Farinacci, V. Fuller, and D. Oran. Locator/ID Separation Protocol (LISP). draft-farinacci-lisp-00.txt, 2007.

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

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

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

[7] M. Lad, D. Massey, D. Pei, Y. Wu, B. Zhang, and L. Zhang. Phas: A prefix hijack alert system. In *Proc. of the 15th USENIX Security Symposium*, 2006.

[8] M. Lad, D. Massey, and L. Zhang. A graphical tool for capturing BGP routing dynamics. In *NOMS*, April 2004.

[9] D. Meyer, L. Zhang, and K. Fall. Report from the IAB Workshop on Routing and Addressing. draft-iab-raws-report-01.txt, 2007.

[10] R. Moskowitz, P. Nikander, P. Jokela, and T. Henderson. Host Identity Protocol. draft-ietf-hip-base-07.txt, 2007.

[11] M. O'Dell. GSE - An Alternate Addressing Architecture for IPv6. February 1997.

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

[13] Y. Rekhter and T. Li. Border Gateway Protocol 4. RFC 1771, Internet Engineering Task Force, July 1995.

[14] E. C. Rosen. Exterior Gateway Protocol (EGP). RFC 827, 1982.

[15] L. Subramanian, M. Caesar, C. T. Ee, M. Handley, Z. M. Mao, S. Shenker, and I. Stoica. HLP: A Next Generation Inter–domain Routing Protocol. In *ACM SIGCOMM*, 2005.