Network Working Group
Internet-Draft
Intended status: Informational
Expires: August 29, 2007

D. Massey
Colorado State
L. Wang
U. Memphis
B. Zhang
U. Arizona
L. Zhang
UCLA

February 25, 2007

A Proposal for Scalable Internet Routing & Addressing
draft-wang-ietf-efit-00.txt

Status of this Memo

Copyright Notice

Abstract

Our measurement studies of the global Internet routing system show that prefix de-aggregation has led to the DFZ routing table size growing at a rate which is much faster than the Internet itself. The main causes of prefix de-aggregation include user site multihoming and traffic engineering. We propose to move Internet service providers to a separate address space as an effective solution to the routing scalability problem. We discuss different means to provide the mapping service between user and provider address spaces and the pros and cons of each approach, as well as why we believe such an architectural change is necessary to solve the routing scalability problem.

Table of Contents

1. Introduction

Analysis of the routing tables in the default free zone (DFZ) reveals important problems for Internet routing. The first and most noticeable aspect is the growth in the table size. Our analysis of data collected by RouteViews and RIPE BGP monitoring projects shows that the DFZ routing table size has been growing at a much higher rate than that of the routable address space in the Internet. At the same time, it is well known that IPv4 address allocations are constrained by the shortage of remaining IPv4 address space. An (unintended) side effect of this shortage may be limiting the speed of route table growth. When IPv6 deployment starts rolling out in a wide scale and removes the current address space limits, we envision that the DFZ routing table could substantially accelerate its growth.

By raising concerns over table size, we do not mean to imply that growth is inherently bad. When managed correctly, growth in the table size can reflect the increasing importance and increasing scale of the network and is a sign of a healthy network. But the current situation is an example of poorly managed growth. In today's Internet routing and addressing architecture, both user sites and Internet service providers share the same address and routing space. But user sites and service providers do not share the same goals and same challenges. From a provider perspective, the routing system needs topologically aggregatable address assignment and usage. The assignment and use of topologically aggregatable addresses could not only reduce the table size, but enable better network routing by creating more meaningful entries in the table. However enterprise customers desire Provider-Independent (PI) address blocks in order to maintain the freedom to switch between ISPs while avoiding renumbering, and to ease multihoming which has become a universal practice. From the perspective of a user site, multihoming is necessary and should be encouraged to help provide a more robust Internet. In fact, our data suggests that a main cause of the rapid routing table growth is site multihoming. Transit providers and user sites effectively seek opposite approaches to address allocation and routing announcements with both sides supported by valid technical and economic reasons. Overall, the global routing system is losing the ground of topologically aggregatable address usage, as evidenced by increasingly fragmented prefixes in BGP routing announcements. These issues have also been identified in the report of IAB Workshop on Routing and Addressing [RAWS].

Recognizing the direct conflicts between Internet users and ISPs with regard to the IP address allocation and usage, we propose to solve the global routing scalability problem by putting the ISPs in a separate IP address space, where the addresses can be allocated in a topologically aggregatable way to enable scalable routing. Data

traffic between Internet users can be tunneled over the transit core, in a way similar to VPNs which are widely used today for interconnecting multiple sites of the same enterprise network across the global Internet backbone. Our proposed solution is to tunnel all data traffic between all user sites across the backbone. The separate address spaces allows both providers and user sites to move forward in a way that meets their respective technical and economic objectives. It has further advantages for facilitating the roll-out of IPv6 addresses since data is now tunneled across the core, irrespective of whether the user site has deployed IPv4 or IPv6 addresses. Overall, we believe the approach has great promise for both provider and user site view address allocations, route announcements, and routing in general.

In the rest of this draft we describe the proposed solution in more detail and outline the open issues in its implementation. We also identify the pros and cons of different solutions to the open issues.

2. The eFIT Proposal

The Internet is comprised of two types of networks: user networks and transit provider networks. User networks correspond to business, universities, and organizations. To them, the Internet is a means to some end(s), i.e., they run applications to communicate with other user networks over the Internet. On the other hand, transit networks are ISPs, whose goal is to realize and sell end-to-end data delivery service. The number of user networks is much larger than that of transit networks. Moreover, user networks are growing at a faster rate than transit networks which actively adjust their connectivity in order to accommodate the growing user networks. These two types of networks have different purposes, different growth trends, and different operational goals. Putting them in the same address and routing space has been the root cause for many problems as the Internet grows substantially.

We propose to separate user networks and transit provider networks in terms of both addressing and routing. First, user networks and transit networks use separate address spaces, and their address formats can be different. Second, their routing spaces are separated. A provider routing protocol is operated among routers inside the transit core to maintain routes to all the provider networks only. The provider routing fundamentally differs from the current BGP as it is confined within the provider space. Each user network runs its own routing protocol to maintain routes to reach internal subnets and its immediate neighbors (its providers or other directly connected user networks). There is no routing protocol operating across the links between the user networks and the transit core.

A network becomes part of the provider space if it obtains a provider address prefix and runs the provider routing protocol with other transit providers. A user network can continue using its current provider independent address block if it has one or it can obtain one from a regional Internet registry. It can also use its current intra-domain routing protocol.

End-to-end data delivery is achieved by tunneling user packets across the provider space. When a user packet reaches the border of the transit core, the ingress edge router will map the destination user address to a provider address that directly connects to the destination user network. The ingress edge router encapsulates the user packet and sends it to the destination provider address. Once the packet arrives at the other end of the tunnel, the egress edge router will decapsulate it and send the original data packet into the destination user network. A mapping service is needed to map user addresses to corresponding provider addresses. It will be discussed

in Section 4.

On the surface the encapsulation step in crossing the transit core may bear a resemblance to NAT (Network Address Translation); however, our proposal differs from NAT in fundamental ways. We assign unique and provider-independent addresses to all hosts in user space, thus they can be reached in the face of individual provider failures. Any user host can directly talk to any other user host by simply putting the destination address in the packet. Therefore our proposal helps reinstall the end-to-end transparency model in the Internet.

Viewed from user networks, the provider space is a single logical hop connecting all user networks, very much like a single "wire" providing data transit service. Therefore, we call our proposal eFIT (enable Future Internet innovation through Transit wire). In this architecture, user networks and transit provider networks are decoupled in both addressing and routing. It helps scale global routing and also allow user networks and transit networks to evolve independently.

With this separation, the routing table size of global DFZ routers is under control, because it only needs to record routes to each provider network, whose number is small and grows slowly. One provider network may announce multiple prefixes. However, since the addressing is separated from that of user networks, we can design the format of provider addresses and allocate them in a way to encourage aggregation. More importantly, the routing table will not be inflated by user networks' multi-homing and traffic engineering practices or the fast increase of user networks. Another benefit is that the number of routing updates is under control for DFZ routers, because the instability from user networks is insulated from the transit core. Considering that there is a large and increasing number of user networks, and some of them are the sources of frequent routing instability, the separation will reduce the routing updates in the transit core significantly. In short, since the global routing is confined among provider networks only, its scalability will be much better than that of today's Internet, and will also be able to sustain Internet growth in the near future.

The separation is also beneficial to user networks. They will have truly provider-independent addresses, which enable them to change providers without renumbering. They may also have an address format that is designed to best facilitate their internal routing. They can explore different site multihoming and traffic engineering techniques without affecting the scalability of the global routing system.

3. How To Bridge The Two Spaces?

Separating the user networks and provider networks makes the global routing system scale. But at the same time, end-to-end data delivery also requires bridging the two spaces. Three mechanisms are essential to the bridging.

First, a mapping service is needed to map each user address to its provider attachment points, i.e., the egress transit router(s) through which data packets can be delivered to the user address. These egress transit routers are directly connected to the user network that owns the particular user address.

Second, end-to-end data delivery across the transit core is achieved by encapsulating user data packets in a provider space packet header, whose source address is the ingress transit router connected to the source user network, and whose destination address is the egress transit router connected to the destination user network.

Third, the failures of links between user space and provider space must be handled by a mechanism external to the routing system. In eFIT, provider network routing and user network routing are confined to their own space, while links between them are in neither space. Therefore, failures of such border links are not reflected in either provider routing or user routing. A mechanism is needed so that the ingress transit router can stop sending data packets destined to the failed link until it is up again.

In the next two section, we will discuss potential approaches to implement the mapping service and handle border link failures.

4. The Mapping Service

The basic functionality of the mapping service is that, given a destination user address, it should return one or more destination provider addresses so that the packet can be encapsulated and forwarded across the transit core to reach the destination user network.

The mapping service should be implemented by the transit core; otherwise it will introduce another dependency between the two spaces. Since the transit service providers are responsible for data delivery between user networks, they have the incentive to provide quick and secure mapping service.

In general, there are two types of approaches to implement this service.

Flooding: This approach simply floods the mapping data through all transit service providers; Each provider can then decide how it may further distribute the information to its own edge routers. An advantage of this approach is that the mapping information is readily available locally at the edge routers, therefore packet forwarding should not experience any significant delay from looking up the mapping information. The disadvantage is that any change in the mapping information must be proactively propagated to all providers, even when the change may not affect any data traffic. Given the number of user networks may grow at a rapid rate, the dissemination system can potentially face a scalability challenge.

Distributed Servers: this approach builds a system of distributed servers that make the mapping information available through query and reply. This system of servers can be implemented either in a hierarchy such as that of the DNS, or in a flat structure such as distributed hash tables (DHTs). One advantage is that a change in the mapping information only results in a change to some servers, rather than being proactively propagated globally. Another advantage is that individual responsible parties can selectively enhance their own mapping service through more replications or fast servers. The disadvantage of this approach is that the lookup process may add extra delay to packet forwarding. Caching and prefetching popular mapping entries can provide effective performance improvement, but with associated increase in system complexity.

Overall, there are interesting trade-offs in each approach and further research is needed.

The basic mapping service essentially provides the binding between a user address and its provider attachment points. This binding can be used to support the migration from IPv4 to IPv6. For example, a user network can move to IPv6 independently from whether its ISPs have deployed IPv6, as long as the mapping service supports the mapping between the IPv6 user address to the IPv4 provider addresses. The mapping service also opens the door to new services or functionality. For example, the mapping information could specify the address owner's preferred traffic distribution among its multiple provider attachment points, in order to load balance incoming traffic.

5. Handling Border Link Failures

In this section, we discuss how to handle the failures of border links between user networks and their providers effectively and efficiently. We note that these failures could be treated as mapping changes and thus handled by the mapping service directly, i.e., treating every link up and down as a change of the mapping information. However, the link failures could occur much more frequently than changes in business relationship such as a user network subscribing to a new ISP. If the link failures are propagated as mapping changes, a damping mechanism must be in place to prevent a flapping link from overloading the mapping service with its frequent status changes.

Below we discuss other options that do not necessarily involve the mapping service, i.e., the link failures are not treated as mapping changes.

1. Masking the failure via tunneling: This approach lets the egress transit router mask the link failure by redirecting the packet to another router that is also connected to the destination user network; The first router can use the mapping service to find the second router. This is a viable approach if both routers belong to the same provider, or otherwise the first router may not have the economic incentive to redirect the packet to a competitor.

2. Notifying ingress transit routers after the failure: this approach notifies ingress transit routers about the link failure so that they will stop sending data packets towards the failed link for a period of time specified in the notification messages. The propagation of such notification messages can be proactive or reactive.

   * In the proactive mode, the egress transit router sends the failure notification to all the ingress transit routers whenever the failure occurs, e.g., through a flooding mechanism. However, this mode is appropriate only if the link fails occasionally. Otherwise, a flapping link could lead to frequent waves of failure notification messages flooded everywhere. Note that if the mapping service is implemented using flooding, this function could be (but does not have to be) provided by the mapping service, except that the receiver needs to know that this is a temporary failure, not a long-term change. If the mapping service is implemented using a network of servers, the proactive propagation of the notification messages to the ingress transit routers needs to be handled by a separate mechanism.

* In the reactive mode, the router propagates the information only to the ingress transit routers that are currently communicating with the affected destination users. For example, whenever the router receives a packet destined to the other side of a failed link, it sends an ICMP packet back to the the ingress transit router in the source provider network. This approach has a lower control traffic overhead than the proactive mode because it limits the impact to only active sources. On the flip side, it incurs extra delay for the first few packets sent to the destination, because the packets get dropped and retransmissions will not succeed until the notification message is received. Moreover, it could lead to a lot of notification traffic if the egress router does not keep track of which ingress routers have received its notifications.

Overall, The tradeoff depends on the scale of the system and failure impact. Proactive notification is more suitable where the overhead of propagating updates everywhere is manageable and when the destination is very popular. Reactive notification is more suitable when the system is very large and the destination has only a few active sources. Another critical problem is security: How does the receiving router trust that the failure notification message indeed comes from the egress transit router? More research is needed to design an efficient and scalable mechanism to handle border link failures.

6. Related Efforts

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. The proposed solutions share major goals of scalable support for multihoming and avoiding user renumbering when switching providers, and their approaches fall into one of the two categories, 1) putting user and provider into separate address spaces and 2) encoding location information into IP addresses. Although these designs were not (or have yet to be) adopted for deployment, they offer important insights on the reasons why they have yet to materialize.

6.1. Putting users and providers into separate IP address spaces

Recognizing the fundamental conflict between providers' desire for prefix aggregation for routing scalability and user sites' desire for provider-independent addresses to ease multihoming and avoid renumbering, Hinden & Deering proposed ENCAPS in 1996 ([RFC1955], [ENCAPS]). The basic idea is to separate provider networks and user sites into two address spaces, and to use IP-in-IP tunnels to carry packets from source user networks over the provider space to reach destination user networks. Our eFIT proposal shares the same solution direction with ENCAPS, so is another more recent effort LISP [LISP] which sketched out an instantiation of ENCAPS implementation.

O'Dell made another new routing design, named GSE [GSE], in 1997. The basic idea is to divide IPv6's 16-byte address into two parts, with the lower N bytes being used for the End System Designator (ESD) and local routing, and the higher (16 - N) bytes used for routing between providers. [GSEOverview] provides a comprehensive analysis of GSE's pros and cons, as well as the open issues in its implementation.

In essence GSE uses the upper (16-N) bytes of IPv6 address to represent the address space in the provider domain, hence GSE shares the fundamental idea with ENCAPS in envisioning a network where customers and providers live on separate address spaces. As such it also shares with ENCAPS the need for a mapping service. ENCAPS needs this mapping service to map the destination user address to the address of the tunnel exit point which should be a router of the provider serving the destination user, while GSE needs this mapping service to map ESDs to the corresponding upper (16 - N) bytes of IPv6 addresses.

6.2. Location-Based Addressing

Another way to allocate addresses in an aggregatable manner is to base the allocation on locations, which was proposed by Deering in early 90's [MetroAddr]. This approach can avoid user renumbering when they change providers, as long as they stay in the same location.

More recently a similar proposal, Geo-based addressing [GeoAddr] has been made. Although this proposal has certain differences from [MetroAddr], for example encoding latitude and longitude information into the address instead of metro-area ID, the two proposals bear fundamental similarities. They are both proposed as one of the ways, but not necessarily the only way, to allocate IPv6 addresses, both envisioned coexistence of location-based and provider-based addresses, and which type to use would be based on the need of individual parties.

However location-based addressing imposes two infeasible conditions to the routing system. First, routing over location-based addresses requires that ISPs interconnect at each location. Second, location-based addresses do not reflect interconnectivity among providers to enable routing policies.

6.3. Summary of Previous and Ongoing Efforts

As time goes, multiple solution development efforts have pointed to the same direction of separating user sites and provider 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.

We also believe that encoding location information into IP address can serve very useful purposes. However solely location-based addressing is problematic as it is unable to support routing policies. We have an ongoing effort which proposes a new address structure for the provider address space and utilizes location information to facilitate scalable routing and traffic engineering.

7. Summary

In concluding this draft we would like to clarify three important points: (1) exactly what need to be separated, (2) the impact on existing implementations, and (3) why we believe it is necessary to separate ISPs from the user address space in order to solve the routing scalability problem.

As far as the global routing scalability is concerned, the root cause of the problem is due to user sites and ISPs live on the same address and routing space, while each has goals conflicting with that of the other. Thus our proposed solution calls for a separation of the provider and user address spaces. The IAB workshop report [RAWS] identified the same problem, although it phrased the problem as "the overloading of IP address semantics". We would like to clarify that the problem is not overloading address and identifier, but overloading providers and end users on the same address space.

There may exist a need for host identifiers. 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 [HIP]. However deploying HIP alone is not a solution to the routing scalability problem, even though it offers each host an identifier. Both provider networks and user sites need IP addresses to manage their networks and forward packets. One can envision a host identifier solution being deployed on top of, but not in place of, user IP addresses.

The second point is about the impact of our proposed solution on currently deployed systems and protocols. eFIT and similar solutions introduce a new component into the Internet architecture, the mapping service. If we design the mapping service right, then by and large all user sites and ISPs should be able to stay with their current operational practice regarding packet transmission and forwarding, with all user sites using provider-independent addresses and all ISPs using topologically aggregatable addresses. The edge routers connecting user sites to the transit core will need to be changed to use the mapping service and tunnel packets over the transit core.

The last point we would like to stress is the necessity for deploying our proposed solution (or a similar one in the same direction). Putting ISPs in a separate IP address space for a scalable global routing system requires a new mapping component to bridge the two address spaces. Research efforts are needed to develop a reliable and robust mapping service. This new mapping service will necessarily bring additional complexity into the Internet

architecture, thus a question naturally arises: why is it necessary to change the existing addressing and routing architecture?

We believe the answer lies in the fact that the Internet has grown by orders of magnitude. The existing address architecture of having all the IP boxes living in the same address space was designed at the birth of the Internet when it was very small in size. Today the Internet has become the largest cohesive system we have ever built, and perhaps the most important infrastructure for the society. Different parts of Internet have become specialized to serve different purposes, for example user sites are multi-homed for enhanced reliability and performance, while service provider networks are specialized for high performance, yet economical, packet delivery service. The different goals of different parties brought different and conflicting requirements to the shared address space. Thus the original address architecture can no longer meet the functional requirements of today's grown up Internet.

In a 1928 article by J. B. S. Haldane, "being the right size" [RIGHTSIZE], 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 hundredfold, 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."

We believe the same is true for Internet. As it grows large in user population size, it is no longer feasible for its transit core to deliver packets by maintaining the reachability information of end users. In addition, the transit core is also under competitive market forces to maintain a modest cost in carrying out packet delivery service. The growth makes it necessary for the transit core to operate in a separate address space than the edge users, so that each can evolve independently to fulfill its own role. We also believe that this separation opens the door for adding new functions and capabilities to the routing system; we will elaborate in more detail in our future documents.

We would like to solicit the community's input and comments regarding

moving the Internet routing and addressing architecture towards this proposed direction. Comments can be sent directly to the authors.

8. Acknowledgments

Our efforts on global routing system studies have been supported by research fundings from DARPA and NSF.

9. IANA Considerations

This document requires no actions by IANA.

10. Security Considerations

The security of the global routing system is of great concern. This document introduces a proposed solution to routing scalability problem. The proposed solution has a potential to enhance routing system security, although the specific design and evaluation are yet to be carried out. The document is informational and it proposes no new protocol or protocol usage, and as such presents no new security issues.

11. Informative References

[ENCAPS] Deering, S., "The Map & Encap Scheme for scalable IPv4 routing with portable site prefixes", http://irl.cs.ucla.edu/references/Deering-encap.pdf, March 1996.

[GSE] O'Dell, M., "GSE - An Alternate Addressing Architecture for IPv6", Internet Draft, http://www.watersprings.org/pub/id/draft-ietf-ipngwg-gseaddr-00.txt, 1997.

[GSEOverview] Zhang, L., "An Overview of Multihoming and Open Issues in GSE", IETF Journal http://www.isoc.org/tools/blogs/ietfjournal/?p=98#more-98, 2006.

[GeoAddr] Hain, T., "An IPv6 Provider-Independent Global Unicast Address Format", Internet Draft, http://www.ietf.org/internet-drafts/draft-hain-ipv6-pi-addr-10.txt, August 2006.

[HIP] Moskowitz et al., R., "Host Identity Protocol", http://www.ietf.org/internet-drafts/draft-ietf-hip-base-07.txt, 2007.

[LISP] Farinacci, D., Fuller, V., and D. Oran, "Locator/ID Separation Protocol (LISP)", Internet Draft, http://www.ietf.org/internet-drafts/draft-farinacci-lisp-00.txt, 2007.

[MetroAddr] Deering, S. and R. Hinden, "IPv6 Metro Addressings", http://irl.cs.ucla.edu/references/Deering-metro.txt, March 1996.

[RAWS] Meyer, D., Zhang, L., and K. Fall, "Report from the IAB Workshop on Routing and Addressing", Internet Draft, http://www.ietf.org/internet-drafts/draft-iab-raws-report-01.txt, 2007.

[RFC1955] Hinden, R., "New Scheme for Internet Routing and Addressing (ENCAPS) for IPNG", RFC 1955 , June 1996.

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

Authors' Addresses

Dan Massey
Colorado State

Email: massey@cs.colostate.edu

Lan Wang
U. Memphis

Email: lanwang@memphis.edu

Beichuan Zhang
U. Arizona

Email: bzhang@cs.arizona.edu

Lixia Zhang
UCLA

Email: lixia@cs.ucla.edu

Full Copyright Statement

Intellectual Property

Acknowledgment

Funding for the RFC Editor function is provided by the IETF Administrative Support Activity (IASA).