Network Working Group
Internet-Draft
Intended status: Informational
Expires: August 21, 2010

T. Narten
IBM

February 17, 2010

On the Scalability of Internet Routing
draft-narten-radir-problem-statement-05.txt

Abstract

There has been much discussion over the last years about the overall scalability of the Internet routing system. Some have argued that the resources required to maintain routing tables in the core of the Internet are growing faster than available technology will be able to keep up. Others disagree with that assessment. This document attempts to describe the factors that are placing pressure on the routing system and the growth trends behind those factors.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

This Internet-Draft will expire on August 21, 2010.

Table of Contents

1. Introduction

Prompted in part by the October, 2006 IAB workshop on Routing & Addressing [RFC4984], there has been a renewed focus on the topic of routing scalability within the Internet. The issue itself is not new, with discussions dating back at least 10-15 years [GSE, ROAD].

This document attempts to describe the "pain points" being placed on the routing system, with the aim of describing the essential aspects so that the community has a way of evaluating whether proposed changes to the routing system actually address or impact existing pain points in a significant manner.

2. Terms and Definitions

Control Plane: The routing setup protocols, their associated state and the activity needed to create and maintain the data structures used to forward packets from one network to another. The term is defined broadly to include all protocols and activities needed to construct and maintain the forwarding tables used to forward packets.

Control Plane Load: The actual load associated with operating the Control Plane. The higher the control plane load, the higher the cost of operating the control plane (in terms of hardware, bandwidth, power, etc.). The terms "routing load" and "control plane load" are used interchangeably throughout this document.

Control Plane Cost: The overall cost associated with operating the Control Plane. The cost consists of capital costs (for hardware), bandwidth costs (for the control plane signalling) and any other ongoing operational cost associated with operating and maintaining the control plane.

Default Free Zone (DFZ): That part of the Internet where routers maintain full routing tables. Many routers maintain only partial tables, having explicit routes for "local" destinations (i.e., prefixes) plus a "default" for everything else. For such routers, building and maintaining routing tables is relatively simple because the amount of information learned and maintained can be small. In contrast, routers in the DFZ maintain complete information about all reachable destinations, which at the time of this writing number in the hundreds of thousands of entries.

Routing Information Base (RIB): The data structures a router maintains that hold the information about destinations and paths to those destinations. The amount of state information maintained is dependent on a number of factors, including the number of individual prefixes learned from peers, the number of BGP peers, the number of distinct paths interconnecting destinations, etc. In addition to maintaining information about active paths used for forwarding, the RIB may also include information about unused ("backup") paths.

Forwarding Information Base (FIB): The actual table consulted while making forwarding decisions for individual packets. The FIB is a compact, optimized subset of the RIB, containing only the information needed to actually forward individual packets, i.e., mapping a packet's destination address to an outgoing interface and next-hop. The FIB only stores information about paths actually used for forwarding; it typically does not store

information about backup paths. The FIB is typically constructed from specialized hardware components, which have different (and higher) cost properties than the hardware typically used to maintain the RIB.

Traffic Engineering (TE): In this document, "traffic engineering" refers to the current practice of inbound, inter-AS traffic engineering. TE is accomplished by injecting additional, more-specific routes into the routing system and/or increasing the frequency of routing updates in order to arrange for inbound traffic at the boundary of an Autonomous system (AS) to travel over a different path than it otherwise would.

Provider Aggregatable (PA) address space: Address space that an end site obtains from an upstream ISP's address block. The main benefit of PA address space is that reachability to all of a provider's customers can be achieved by advertising a single "provider aggregate" address prefix into the DFZ, rather than needing to announce individual prefixes for each customer. An important disadvantage is that when a customer changes providers, the customer must renumber their site into addresses belonging to the new provider and return the previously used addresses to the former provider.

Provider Independent (PI) address space: Address space that an end site obtains directly from a Regional Internet Registry (RIR) for addressing its devices. The main advantage (for the end site) is that it does not need to renumber its site when changing providers, since it continues to use its PI block. However, PI address blocks are not aggregatable and thus each individual PI assignment results in an individual prefix being injected into the DFZ.

Site: Any topologically and administratively distinct entity that connects to the Internet. A site can range from a large enterprise or ISP to a small home site.

3. Background

Within the DFZ, both the size of the RIB and FIB and the overall update rate have historically increased at a greater than linear rate. Specifically:

- The number of individual prefixes that are being propagated into the DFZ over time has been and continues to increase at a faster-than-linear rate. The term "super-linear" has been used to characterize the growth. The exact nature of the growth is much debated (e.g., quadratic, polynomial, etc.), but growth is clearly faster than linear. The reasons behind the rate increase are varied and discussed below. Because each individual prefix requires resources to process, any increase in the number of prefixes produces a corresponding increase in control plane load of the routing system. Each individual prefix that appears in routing updates requires state in the RIB (and possibly the FIB) and consumes processing and other resources when updates related to the prefix are received.

- The overall rate of routing updates is increasing [1], requiring routers to process updates at an increased rate or converge more slowly if they cannot. The rate increase of the control plane load is driven by a number of factors (discussed below). Further study is needed to better understand the factors behind the increasing update rate. For example, it appears that a disproportionate increase in observed updates originates from a small percentage of the total number of advertised prefixes.

The super-linear growth in the routing load presents a scalability challenge for current and/or future routers. While there appears to be general agreement that we will be able to build routers (i.e., hardware & software) actually capable of handling the control plane load, both today and going forward, there is considerable debate about the cost. In particular, will it be possible for ISPs that currently (or would like to) maintain routes as part of the DFZ be able to afford to do so, or will only the largest (and a shrinking number) of top tier ISPs be able to afford the investment and cost of operating the control plane while being a part of the DFZ?

Finally, the scalability challenge is aggravated by the lack of any firm limiting architectural upper-bound on the growth rate of the routing load and a weakening of social constraints that historically have helped restrain the growth rate so far. Going forward, there is considerable uncertainty (some would say doubt) whether future growth rates will continue to be sufficiently constrained so that router development can keep up at an acceptable price point.

3.1. Technical Aspects

The technical challenge of building routers relates to the resources needed to process a larger and increasingly dynamic amount of routing information. More specifically, routers must maintain an increasing amount of associated state information in the RIB, they must be capable of populating a growing FIB, they must perform forwarding lookups at line rates (while accessing the FIB) and they must be able to initialize the RIB and FIB after system restart. All of these activities must take place within acceptable time frames (i.e., paths for individual destinations must converge and stabilize within an acceptable time period). Finally, the hardware needed to achieve this cannot have unreasonable power consumption or cooling demands.

3.2. Business Considerations

While the IETF does not (and cannot) concern itself with business models or the profitability of the ISP community, the cost of running the routing subsystem as a whole is directly influenced by the routing architecture of the Internet, which clearly is the IETF's business. Thus, it is useful to consider the overall business environment that underlies operation of the DFZ routing infrastructure. The DFZ is run entirely by the private sector with no overall governmental oversight or regulatory framework to oversee or even influence what routes are propagated where, who must carry them, etc. ISPs decide (on their own) which routing updates to accept and how (if at all) to process them. Thus, there is no overall authority that can limit the number of prefixes that are injected into the DFZ or that insure that any particular prefixes are accepted at all. Today, the system functions because the set of entities that comprise the DFZ are (generally) able to accept the prefixes that are being advertised and some loose best practices have emerged that are generally followed (e.g., minimum prefix sizes that are routed coupled with RIR policies that place limitations on who may obtain PI prefixes).

In general the Internet would benefit if the cost of the (routing) infrastructure did not grow too rapidly as the Internet grows, since a lower infrastructure cost makes it possible to provide Internet service at a lower cost to a larger number of users. That said, some types of Internet growth tie directly to revenue opportunities or cost savings for an ISP (e.g., adding more users/customers, increasing bandwidth, technological advances, providing new or additional services, etc.). Upgrading or changing infrastructure is most feasible (and expected) when supported by a workable cost recovery model. Hence limiting the cost of self-induced scaling is a nice-to-have benefit, but not a requirement.

On the other hand, it is problematic when the infrastructure cost for an ISP grows (rapidly) due to factors outside of its own control, e.g., resulting from overall Internet growth external to the ISP. If an ISP that does not add new customers, upgrade the bandwidth for their customers, or provide new services needs to upgrade or replace their infrastructure in unexpected ways, then they have no natural cost recovery mechanisms. This is in essence what is happening with the scaling of the global routing table. An ISP that is part of the DFZ may need to upgrade its routers to handle an increased routing load just to maintain the same level of service with respect to their current customers and services.

Even if it is technically possible to build routers capable of meeting the technical and operational requirements, it is also necessary that the overall cost to build, maintain and deploy such equipment meet reasonable business expectations. ISPs, after all, are run as businesses. As such, they must be able to plan, develop and construct viable business plans that provide an acceptable return on investment (i.e., one acceptable to investors).

3.3. Alignment of Incentives

Today's growth pattern is influenced by the scaling properties of the current routing system. If the routing system had better scaling properties, we would be able support and enable more widespread usage of such services as multihoming and traffic engineering. The current system simply would not be able to handle to the routing load if everyone were to choose to multihome. There are millions of potential end sites that would benefit from being able to multihome. This compares with a low few hundred thousand prefixes being carried today. Broader availability of multihoming is limited by barriers imposed by operational practices that try to strike a balance between the amount of multihoming and preservation of routing slots. It is desirable that the routing and addressing system exert the least possible back pressure on end user applications and deployment scenarios, to enable the broadest possible use of the Internet.

One aspect of the current architecture is a misalignment of cost and benefit. Injecting individual prefixes into the DFZ creates a small amount of "pain" for those routers that are part of the DFZ. Each individual prefix adds only a small cost to the routing load, but the aggregate sum of all prefixes is significant, and leads to the key issue at hand. Those that inject prefixes into the DFZ do not generally pay the cost associated with the individual prefix -- it is carried by the routers in the DFZ. But the originator of the prefix receives the benefit. Hence, there is misalignment of incentives between those receiving the benefit and those bearing the cost of providing the benefit. Consequently, incentives are not aligned

properly to produce a natural feedback loop to balance the cost and benefit of maintaining routing tables.

3.4. Table Growth Targets

A precise target for the rate of table size or routing update increase that should reasonably be supported going forward is difficult to state in quantitative terms. One target might simply be to keep the growth at a stable, but manageable growth rate so that the increased router functionality can roughly be covered by improvements in technology (e.g., increased processor speeds, reductions in component costs, etc.).

However, it is highly desirable to significantly bring down (or even reverse) the growth rate in order to meet user expectations for specific services. As discussed below, there are numerous pressures to deaggregate routes. These pressures come from users seeking specific, tangible service improvements that provide "business-critical" value. Today, some of those services simply cannot be supported to the degree that future demand can reasonably be expected because of the negative implications on DFZ table growth. Hence, valuable services are available to some, but not all potential customers. As the need for such services becomes increasingly important, it will be difficult to deny such services to large numbers of users, especially when some "lucky" sites are able to use the service and others are not.

4. Pressures on Routing Table Size

There are a number of factors behind the increase in the quantity of prefixes appearing in the DFZ. From a theoretical perspective, the number of prefixes in the DFZ can be minimized through aggressive aggregation [RFC4632]. In practice, strict adherence to the CIDR principles is difficult.

4.1. Traffic Engineering

Traffic engineering (TE) is the act of arranging for certain Internet traffic to use or avoid certain network paths (that is, TE attempts to place traffic where capacity exists, or where some set of parameters of the path is more favorable to the traffic being placed there).

Outbound TE is typically accomplished by using internal interial gateway protocol (IGP) metrics to choose the shortest exit for two equally good BGP paths. Adjustment of IGP metrics controls how much traffic flows over different internal paths to specific exit points for two equally good BGP paths. Additional traffic can be moved by applying some policy to depreference or filter certain routes from specific BGP peers. Because outbound TE is achieved via a site's own IGP, outbound TE does not impact routing outside of a site.

Inbound TE is performed by announcing a more-specific route along the preferred path that "catches" the desired traffic and channels it away from the path it would take otherwise (i.e., via a larger aggregate). At the BGP level, if the address range requiring TE is a portion of a larger address aggregate, network operators implementing TE are forced to de-aggregate otherwise aggregatable prefixes in order to steer the traffic of the particular address range to specific paths.

TE is performed by both ISPs and customer networks, for three primary reasons:

- to match traffic with network capacity, or to spread the traffic load across multiple links (frequently referred to as "load balancing")
- to reduce costs by shifting traffic to lower cost paths or by balancing the incoming and outgoing traffic volume to maintain appropriate peering relations
- to enforce certain forms of policy (e.g., to prevent government traffic from transiting through other countries)

TE impacts route scaling in two ways. First, inbound TE can result in additional prefixes being advertised into the DFZ. Second, Network operators usually achieve traffic engineering by "tweaking" the processing of routing protocols to achieve desired results, e.g., by sending updates at an increased rate. In addition, some devices attempt to automatically find better paths and then advertise those preferences through BGP, though the extent to which such tools are in use and contributing to the control plane load is unknown.

In today's highly competitive environment, providers require TE to maintain good performance and low cost in their networks.

4.2. Multihoming

Multihoming refers generically to the case in which a site is served by more than one ISP [RFC4116]. Multihoming is used to provide backup paths (i.e., to remove single points of failure), to achieve load-sharing, and to achieve policy or performance objectives (e.g., to use lower latency or higher bandwidth paths). Multihoming may also be a requirement due to contract or law.

Multihoming can be accomplished using either PI or PA address space. A multihomed site advertises its site prefix into the routing system of each of its providers. For PI space, the site's PI space is used, and the prefix is propagated throughout the DFZ. For PA space, the PA site prefix may (or may not) be propagated throughout the DFZ, with the details depending on what type of multihoming is sought.

If the site uses PA space, the PA site prefix allocated from one of its providers (whom we'll call the Primary Provider) is used. The PA site prefix will be aggregatable by the Primary Provider but not the others. To achieve multihoming with comparable properties to that when PI addresses are used as described above, the PA site prefix will need to be injected into the routing system of all of its ISPs, and throughout the DFZ. In addition, because of the longest-match forwarding rule, the Primary Provider must advertise both its aggregate and the individual PA site prefix; otherwise, the path via the primary provider (as advertised via the aggregate) will never be selected due to the longest match rule. For the type of multihoming described here, where the PA site prefix is propagated throughout the DFZ, the use of PI vs. PA space has no impact on the control plane load. The increased load is due entirely to the need to propagate the site's individual prefix throughout the DFZ.

The demand for multihoming is increasing [2]. The increase in multihoming demand is due to the increased reliance on the Internet for mission and business-critical applications (where businesses require 7x24 availability for their services) and the general

decrease in cost of Internet connectivity.

4.3. End Site Renumbering

It is generally considered painful and costly to renumber a site, with the cost proportional to the size and complexity of the network and most importantly, to the degree that addresses are stored in places that are difficult in practice to update. When using PA space, a site must renumber when changing providers. Larger sites object to this cost and view the requirement to renumber akin to being held "hostage" to the provider from which PA space was obtained. Consequently, many sites desire PI space. Having PI space provides independence from any one provider and makes it easier to switch providers (for whatever reason). However, each individual PI prefix must be propagated throughout the DFZ and adds to the control plane load.

It should be noted that while larger sites may also want to multihome, the cost of renumbering drives some sites to seek PI space, even though they do not multihome.

4.4. Acquisitions and Mergers

Acquisitions and mergers take place for business reasons, which usually have little to do with the network topologies of the impacted organizations. When a business sells off part of itself, the assets may include networks, attached devices, etc. A company that purchases or merges with other organizations may quickly find that its network assets are numbered out of many different and unaggragatable address blocks. Consequently, an individual organization may find itself unable to announce a single prefix for all of their networks without renumbering a significant portion of its network.

Likewise, selling off part of a business may involve selling part of a network as well, resulting in the fragmentation of one address block into two (or more) smaller blocks. Because the resultant blocks belong to different organizations, they can no longer be advertised by a single aggregate and the resultant fragments may need to be advertised individually into the DFZ.

4.5. RIR Address Allocation Policies

ISPs and multihoming end sites obtain address space from RIRs. As an entity grows, it needs additional address space and requests more from its RIR. In order to be able to obtain additional address space that can be aggregated with the previously-allocated address space, the RIR must keep a reserve of space that the requester can grow into

in the future. But any reserved address space cannot be used for any other purpose (i.e., assigned to another organization). Hence, there is an inherent conflict between holding address space in reserve to allow for the future growth of an existing allocation holder and using address space efficiently. In IPv4, there has been a heavy emphasis on conserving address space and obtaining efficient utilization. Consequently, insufficient space has been held in reserve to allow for the growth of all sites and some allocations have had to be made from discontiguous address blocks. That is, some sites have received discontiguous address blocks because their growth needs exceeded the amount of space held in reserve for them.

In IPv6, its vast address space allows for a much a greater emphasis to be placed placed on preserving future aggregation than was possible in IPv4.

4.6. Dual Stack Pressure on the Routing Table

The recommended IPv6 deployment model is dual-stack, where IPv4 and IPv6 are run in parallel across the same links. This has two implications for routing. First, although alternative scenarios are possible, it seems likely that many routers will be supporting both IPv4 and IPv6 simultaneously and will thus be managing both IPv4 and IPv6 routing tables within a single router. Second, for sites connected via both IPv4 and IPv6, both IPv4 and IPv6 prefixes will need to be propagated into the routing system. Consequently, dual-stack routers will maintain both an IPv4 and IPv6 route to reach the same destination.

It is possible to make some simple estimates on the approximate size of the IPv6 tables that would be needed if all sites reachable via IPv4 today were also reachable via IPv6. In theory, each autonomous system (AS) needs only a single aggregate route. This provides a lower bound on the size of the fully-realized IPv6 routing table. (As of Feb 2010, [3] states there are 33,548 active ASes in the routing system.)

A single IPv6 aggregate will not allow for inbound traffic engineering. End sites will need to advertise a number of smaller prefixes into the DFZ if they desire to gain finer grained control over their IPv6 inbound traffic. This will increase the size of the IPv6 routing table beyond the lower bound discussed above. There is reason to expect the IPv6 routing table will be smaller than the current IPv4 table, however, because the larger initial assignments to end sites will minimize the de-aggregation that occurs when a site must go back to its upstream address provider or RIR and receive a second, non-contiguous assignment.

It is possible to extrapolate what the size of the IPv6 Internet routing table would be if widespread IPv6 adoption occurred, from the current IPv4 Internet routing table. Each active AS (33,548) would require at least one aggregate. In addition, the IPv6 Internet table would also carry more-specific prefixes for traffic engineering. Assume that the IPv6 Internet table will carry the same number of more specifics as the IPv4 Internet table. In this case one can take the number of IPv4 Internet routes and subtract the number of CIDR aggregates that they could easily be aggregated down to. As of Feb 2010, the 313,626 routes can be easily aggregated down to 193,844 CIDR aggregates [3]. That difference yields 119,782 extra more-specific prefixes. Thus if each active AS (33,548) required one aggregate, and an additional 119,782 more specifics were required, then the IPv6 Internet table would be 153,330 prefixes.

4.7. Internal Customer Routes

In addition to the Internet routing table, networks must also support their internal routing table. Internal routes are defined as more-specific routes that are not advertised to the DFZ. This primarily consists of prefixes that are a more-specific of a provider aggregate (PA) and are assigned to a single-homed customer. The DFZ need only carry the PA aggregate in order to deliver traffic to the provider. However, the provider's routers require the more-specific route to deliver traffic to the end site.

Internal routes could also come from more-specific prefixes advertised by multihomed customers with the "no-export" BGP community. This is useful when the fine grained control of traffic to be influenced can be contained to the neighboring network.

For a large ISP, the internal IPv4 table can be between 50,000 and 150,000 routes. During the dot com boom some ISPs had more internal prefixes than there were in the Internet table. Thus the size of the internal routing table can have significant impact on the scalability and should not be discounted.

4.8. IPv4 Address Exhaustion

The IANA and RIR free pool of IPv4 addresses will be exhausted within a few years. As the free pool shrinks, the size of the remaining unused blocks will also shrink and unused blocks previously held in reserve for expansion of existing allocations or otherwise not used due to their smaller size will be allocated for use. Consequently, as the community looks to use every piece of available address space (no matter how small) there will be an increasing pressure to advertise additional prefixes in the DFZ.

5. Pressures on Control Plane Load

This section describes a number of trends and pressures that are contributing to the overall routing load. The previous section described pressures that are increasing the size of the routing table. Even if the size could be bounded, the amount of work needed to maintain paths for a given set of prefixes appears to be increasing.

5.1. Interconnection Richness

The degree of interconnectedness between ASes has increased in recent years. That is, the Internet as a whole is becoming "flatter" with an increasing number of possible paths interconnecting sites [4]. As the number of possible paths increase, the amount of computation needed to find a best path also increases. This computation comes into effect whenever a change in path characteristics occurs, whether from a new path becoming available, an existing path failing, or a change in the attributes associated with a potential path. Thus, even if the total number of prefixes were to stay constant, an increase in the interconnection richness implies an increase in the resources needed to maintain routing tables.

5.2. Multihoming

Multihoming places pressure on the routing system in two ways. First, an individual prefix for a multihomed site (whether PI or PA) must be propagated into the routing system, so that other sites can find a good path to the site. Even if the site's prefix comes out of a PA block, an individual prefix for the site needs to be advertised so that the most desirable path to the site can be chosen when the path through the aggregate is sub-optimal. Second, a multihomed site will be connected to the Internet in more than one place, increasing the overall level of interconnection richness. If an outage occurs on any of the circuits connecting the site to the Internet, those changes will be propagated into the routing system. In contrast, a singly-homed site numbered out of a Provider Aggregate places no additional control plane load in the DFZ as the details of the connectivity status to the site are kept internal to the provider to which it connects.

5.3. Traffic Engineering

The mechanisms used to achieve multihoming and inbound Traffic Engineering are the same. In both cases, a specific prefix is advertised into the routing system to "catch" traffic and route it over a different path than it would otherwise be carried. When multihoming, the specific prefix is one that differs from that of its

ISP or is a more-specific of the ISP's PA. Traffic Engineering is achieved by taking one prefix and dividing it into a number of smaller and more-specific ones, and advertising them in order to gain finer-grained control over the paths used to carry traffic covered by those prefixes.

Traffic Engineering increases the number of prefixes carried in the routing system. In addition, when a circuit fails (or the routing attributes associated with the circuit change), additional load is placed on the routing system by having multiple prefixes potentially impacted by the change, as opposed to just one.

5.4. Questionable Operational Practices?

Some operators are believed to engage in operational practices that increase the load on the routing system.

5.4.1. Rapid shuffling of prefixes

Some networks try to assert fine-grained control of inbound traffic by modifying route announcements frequently in order to migrate traffic to less loaded links quickly. The goal of this is to achieve higher utilization of multiple links. In addition, some route selection devices actively measure link or path utilization and attempt to optimize inbound traffic by withholding or depreferencing certain prefixes in their advertisements. In short, any system that actively measures load and modifies route advertisements in real time increases the load on the routing system, as any change in what is advertised must ripple through the entire routing system.

5.4.2. Anti-Route Hijacking

In order to reduce the threat of accidental (or intentional) hijacking of its address space by an unauthorized third party, some sites advertise their space as a set of smaller prefixes rather than as one aggregate. That way, if someone else advertised a path for the larger aggregate (or a small piece of the aggregate), it will be ignored in favor of the more-specific announcements. This increases both the number of prefixes advertised, and the number of updates.

5.4.3. Operational Ignorance

It is believed that some undesirable practices result from operator ignorance, where the operator is unaware of what they are doing and the impact that has on the DFZ.

The default behavior of most BGP configurations is to automatically propagate all learned routes. That is, one must take explicit

configuration steps to prevent the automatic propagation of learned routes. In addition, it is often significant work to figure out how to (safely) aggregate routes (and which ones to aggregate) in order to reduce the number of advertisements propagated elsewhere. While vendors could provide additional configuration "knobs" to reduce leakage, the implementation of additional features increases complexity and some operators may fear that the new configuration will break their existing routing setup. Finally, leaking routes unnecessarily does not generally harm those responsible for the misconfiguration, hence, there may be little incentive to change such behavior.

5.5. RIR Policy

RIR address policy has direct impact on the control plane load because address policy determines who is eligible for a PI assignment (which impacts how many are given out in practice) and the size of the assignment (which impacts how much address space can be aggregated within a single assignment). If PI assignments for end sites did not exist, then those end sites would not advertise their own prefix directly into the global routing system; instead their address block would be covered by their provider's aggregate. That said, RIRs have adopted PI policies in response to community demand, for reasons described elsewhere (e.g., to support multihoming and to avoid the need to renumber). In short, RIR policy can be seen as a symptom rather than a root cause.

6. Summary

As discussed in previous sections, in the current operating environment, an ISP may experience an overall increase in the routing load due entirely to external factors outside of its control. These external pressures can make it increasingly difficult for ISPs to recover control plane related costs associated with the growth of the Internet. Moreover, real business and user needs are creating increasing pressure to use techniques that increase the control plane load for ISPs operating within the DFZ. While the system largely works today, there is a real risk that the current cost and incentive structures will be unable to keep control plane costs manageable (within the context of then-available routing hardware) over the next decades. The Internet would strongly benefit from a routing and addressing model designed with this in mind. Thus, in the absence of a business model that better supports such cost recovery, there is a need for an approach to routing and addressing that fulfils the following criteria:

1. Provides sufficient benefits to the party bearing the costs of deploying and maintaining the technology to recover the cost for doing so.

2. Reduces the growth rate of the DFZ control plane load. In the current architecture, this is dominated by the routing, which is dependent on:

   A. The number of individual prefixes in the DFZ

   B. The update rate associated with those prefixes.

   Any change to the control plane architecture must result in a reduction in the overall control plane load, and shouldn't simply shift the load from one place in the system to another, without reducing the overall load as a whole.

3. Allows any end site wishing to multihome to do so

4. Supports ISP and enterprise TE needs

5. Allows end sites to switch providers while minimizing configuration changes to internal end site devices.

6. Provides end-to-end convergence/restoration of service at least comparable to that provided by the current architecture

This document has purposefully been scoped to focus on the growth of the routing control plane load of operating the DFZ. Other problems

that may seem related, but do not directly impact on route scaling are not considered to be "in scope" at this time. For example, Mobile IP [RFC3344] [RFC3775] and NEMO [RFC3963] place no pressures on the routing system. They are layered on top of existing IP, using tunneling to forward packets via a care-of addresses. Hence, "improving" these technologies (e.g., by having them leverage a solution to the multihoming problem), while a laudable goal, is not considered a necessary goal.

7. Security Considerations

This document does not introduce any security considerations.

8. IANA Considerations

This document contains no IANA actions.

9. Acknowledgments

The initial version of this document was produced by the Routing and Addressing Directorate (http://www.ietf.org/IESG/content/radir.html). The membership of the directorate at that time included Marla Azinger, Vince Fuller, Vijay Gill, Thomas Narten, Erik Nordmark, Jason Schiller, Peter Schoenmaker, and John Scudder.

Comments should be sent to rrg@iab.org or to radir@ietf.org.

10. Informative References

[RFC3344] Perkins, C., "IP Mobility Support for IPv4", RFC 3344, August 2002.

[RFC3775] Johnson, D., Perkins, C., and J. Arkko, "Mobility Support in IPv6", RFC 3775, June 2004.

[RFC3963] Devarapalli, V., Wakikawa, R., Petrescu, A., and P. Thubert, "Network Mobility (NEMO) Basic Support Protocol", RFC 3963, January 2005.

[RFC4116] Abley, J., Lindqvist, K., Davies, E., Black, B., and V. Gill, "IPv4 Multihoming Practices and Limitations", RFC 4116, July 2005.

[RFC4632] Fuller, V. and T. Li, "Classless Inter-domain Routing (CIDR): The Internet Address Assignment and Aggregation Plan", BCP 122, RFC 4632, August 2006.

[RFC4984] Meyer, D., Zhang, L., and K. Fall, "Report from the IAB Workshop on Routing and Addressing", RFC 4984, September 2007.

[1] <http://www3.ietf.org/proceedings/06mar/slides/grow-3.pdf>

[2] <http://www.cidr-report.org/as2.0/, http://www.cidr-report.org/cgi-bin/plota?file=%2fvar%2fdata%2fbgp%2fas2.0%2fbgp%2das%2dcount%2etxt&descr=Unique%20ASes&ylabel=Unique%20ASes&with=step,http://www.potaroo.net/tools/asn32/>

[3] <http://www.cidr-report.org/as2.0/>

[4] <http://www.potaroo.net/bgprpts/bgp-average-aspath-length.png>

Author's Address

Thomas Narten
IBM

Email: narten@us.ibm.com