Network Working Group
Request for Comments: 4984
Category: Informational

D. Meyer, Ed.
L. Zhang, Ed.
K. Fall, Ed.

September 2007

Report from the IAB Workshop on Routing and Addressing

Status of This Memo

Abstract

This document reports the outcome of the Routing and Addressing Workshop that was held by the Internet Architecture Board (IAB) on October 18-19, 2006, in Amsterdam, Netherlands. The primary goal of the workshop was to develop a shared understanding of the problems that the large backbone operators are facing regarding the scalability of today's Internet routing system. The key workshop findings include an analysis of the major factors that are driving routing table growth, constraints in router technology, and the limitations of today's Internet addressing architecture. It is hoped that these findings will serve as input to the IETF community and help identify next steps towards effective solutions.

Note that this document is a report on the proceedings of the workshop. The views and positions documented in this report are those of the workshop participants and not of the IAB. Furthermore, note that work on issues related to this workshop report is continuing, and this document does not intend to reflect the increased understanding of issues nor to discuss the range of potential solutions that may be the outcome of this ongoing work.

Table of Contents

1. Introduction

It is commonly recognized that today's Internet routing and addressing system is facing serious scaling problems. The ever-increasing user population, as well as multiple other factors including multi-homing, traffic engineering, and policy routing, have been driving the growth of the Default Free Zone (DFZ) routing table size at an increasing and potentially alarming rate [DFZ][BGT04]. While it has been long recognized that the existing routing architecture may have serious scalability problems, effective solutions have yet to be identified, developed, and deployed.

As a first step towards tackling these long-standing concerns, the IAB held a "Routing and Addressing Workshop" in Amsterdam, Netherlands on October 18-19, 2006. The main objectives of the workshop were to identify existing and potential factors that have major impacts on routing scalability, and to develop a concise problem statement that may serve as input to a set of follow-on activities. This document reports on the outcome from that workshop.

The remainder of the document is organized as follows: Section 2 provides an executive summary of the workshop findings. Section 3 describes the sources of stress in the current global routing and addressing system. Section 4 discusses the relationship between Moore's law and our ability to build large routers. Section 5 describes a few foreseeable factors that may exacerbate the current problems outlined in Section 2. Section 6 describes previous work in this area. Section 7 describes the problem statements in more detail, and Section 8 discusses the criteria that constrain the solution space. Finally, Section 9 summarizes the recommendations made by the workshop participants.

The workshop participant list is attached in Appendix B. The agenda can be found in Appendix C, and Appendix D provides pointers to the presentations from the workshop.

Finally, note that this document is a report on the outcome of the workshop, not an official document of the IAB. Any opinions expressed are those of the workshop participants and not of the IAB.

2. Key Findings from the Workshop

This section provides a concise summary of the key findings from the workshop. While many other aspects of a routing and addressing system were discussed, the first two problems described in this section were deemed the most important ones by the workshop participants.

The clear, highest-priority takeaway from the workshop is the need to devise a scalable routing and addressing system, one that is scalable in the face of multihoming, and that facilitates a wide spectrum of traffic engineering (TE) requirements. Several scalability problems of the current routing and addressing systems were discussed, most related to the size of the DFZ routing table (frequently referred to as the Routing Information Base, or RIB) and its implications. Those implications included (but were not limited to) the sizes of the DFZ RIB and FIB (the Forwarding Information Base), the cost of recomputing the FIB, concerns about the BGP convergence times in the presence of growing RIB and FIB sizes, and the costs and power (and hence heat dissipation) properties of the hardware needed to route traffic in the core of the Internet.

2.1. Problem #1: The Scalability of the Routing System

The shape of the growth curve of the DFZ RIB has been the topic of much research and discussion since the early days of the Internet [H03]. There have been various hypotheses regarding the sources of this growth. The workshop identified the following factors as the main driving forces behind the rapid growth of the DFZ RIB:

- Multihoming,
- Traffic engineering,
- Non-aggregatable address allocations (a big portion of which is inherited from historical allocations), and
- Business events, such as mergers and acquisitions.

All of the above factors can lead to prefix de-aggregation and/or the injection of unaggregatable prefixes into the DFZ RIB. Prefix de-aggregation leads to an uncontrolled DFZ RIB growth because, absent some non-topologically based routing technology (for example, Routing On Flat Labels [ROFL] or any name-independent compact routing algorithm, e.g., [CNIR]), topological aggregation is the only known practical approach to control the growth of the DFZ RIB. The following section reviews the workshop discussion of the implications of the growth of the DFZ RIB.

2.1.1. Implications of DFZ RIB Growth

Presentations made at the workshop showed that the DFZ RIB has been growing at greater than linear rates for several years [DFZ]. While this has the obvious effects on the requirements for RIB and FIB memory sizes, the growth driven by prefix de-aggregation also exposes the core of the network to the dynamic nature of the edges, i.e., the de-aggregation leads to an increased number of BGP UPDATE messages injected into the DFZ (frequently referred to as "UPDATE churn"). Consequently, additional processing is required to maintain state for the longer prefixes and to update the FIB. Note that, although the size of the RIB is bounded by the given address space size and the number of reachable hosts (i.e., O(m*2^32) for IPv4, where <m> is the average number of peers each BGP router may have), the amount of protocol activity required to distribute dynamic topological changes is not. That is, the amount of BGP UPDATE churn that the network can experience is essentially unbounded. It was also noted that the UPDATE churn, as currently measured, is heavy-tailed [ATNAC2006]. That is, a relatively small number of Autonomous Systems (ASs) or prefixes are responsible for a disproportionately large fraction of the UPDATE churn that we observe today. Furthermore, much of the churn may turn out to be unnecessary information, possibly due to instability of edge ASs being injected into the global routing system [DynPrefix], or arbitrage of some bandwidth pricing model (see [GIH], for example, or the discussion of the behavior of AS 9121 in [BGP2005]).

Finally, it was noted by the workshop participants that the UPDATE churn situation may be exacerbated by the current Regional Internet Registry (RIR) policy in which end sites are allocated Provider-Independent (PI) addresses. These addresses are not topologically aggregatable, and as such, bring the churn problem described above into the core routing system. Of course, as noted by several participants, the RIRs have no real choice in this matter, as many enterprises demand PI addresses that allow them to multihome without the "provider lock" that Provider-Allocated (PA) [PIPA] address space creates. Some enterprises also find the renumbering cost associated with PA address assignments unacceptable.

2.1.2. Implications of DFZ FIB Growth

One surprising outcome of the workshop was the observation made by Tony Li about the relationship between "Moore's Law" [ML] and our ability to build cost-effective, high-performance routers (see Appendix D). "Moore's Law" is the empirical observation that the transistor density of integrated circuits, with respect to minimum component cost, doubles roughly every 24 months. A commonly held wisdom is that Moore's law would save the day by ensuring that technology will continue to scale at historical rates that surpass the growth rate of routing information handled by core router hardware. However, Li pointed out that Moore's Law does not apply to building high-end routers as far as the cost is concerned.

Moore's Law applies specifically to the high-volume portion of the semiconductor industry, while the low-volume, customized silicon used in core routing is well off Moore's Law's cost curve. In particular, off-chip SRAM is commonly used for storing FIB data, and the driver for low-latency, high-capacity SRAM used to be PC cache memory. However, recently cache memory has been migrating directly onto the processor die, and cell phones are now the primary driver for off-chip SRAM. Given cell phones require low-power, small-capacity parts that are not applicable to high-end routers, the SRAMs that are favored for router design are not volume parts and do not track with Moore's law.

2.2. Problem #2: The Overloading of IP Address Semantics

One of the fundamental assumptions underlying the scalability of routing systems was eloquently stated by Yakov Rekhter (and is sometimes referred to as "Rekhter's Law"), namely:

> "Addressing can follow topology or topology can follow addressing. Choose one."

The same idea was expressed by Mike O'Dell's design of an alternate address architecture for ipv6 [GSE], where the address structure was designed specifically to enable "aggressive topological aggregation" to scale the routing system. Noel Chiappa has also written extensively on this topic (see, e.g., [EID]).

There is, however, a difficulty in creating (and maintaining) the kind of congruence envisioned by Rekhter's Law in today's Internet. The difficulty arises from the overloading of addressing with the semantics of both "who" (endpoint identifier, as used by transport layer) and "where" (locators for the routing system); some might also add that IP addresses are also overloaded with "how" [GIH]. In any

event, this kind of overloading is felt to have had deep implications for the scalability of the global routing system.

A refinement to Rekhter's Law, then, is that for the Internet routing system to scale, an IP address must be assigned in such a way that it is congruent with the Internet's topology. However, identifiers are typically assigned based upon organizational (not topological) structure and have stability as a desirable property, a "natural incongruence" arises. As a result, it is difficult (if not impossible) to make a single number space serve both purposes efficiently.

Following the logic of the previous paragraphs, workshop participants concluded that the so-called "locator/identifier overload" of the IP address semantics is one of the causes of the routing scalability problem as we see today. Thus, a "split" seems necessary to scale the routing system, although how to actually architect and implement such a split was not explored in detail.

2.3. Other Concerns

In addition to the issues described in Section 2.1 and Section 2.2, the workshop participants also identified the following three pressing, but "second tier", issues.

The first one is a general concern with IPv6 deployment. It is commonly believed that the IPv4 address space has put an effective constraint on the IPv4 RIB growth. Once this constraint is lifted by the deployment of IPv6, and in the absence of a scalable routing strategy, the rapid DFZ RIB size growth problem today can potentially be exacerbated by IPv6's much larger address space. The only routing paradigm available today for IPv6 is a combination of Classless Inter-Domain Routing (CIDR) [RFC4632] and Provider-Independent (PI) address allocation strategies [PIPA] (and possibly SHIM6 [SHIM6] when that technology is developed and deployed). Thus, the opportunity exists to create a "swamp" (unaggregatable address space) that can be many orders of magnitude larger than what we faced with IPv4. In short, the advent of IPv6 and its larger address space further underscores both the concerns raised in Section 2.1, and the importance of resolving the architectural issue raised in Section 2.2.

The second issue is slow routing convergence. In particular, the concern was that growth in the number of routes that service providers must carry will cause routing convergence to become a significant problem.

The third issue is the misalignment of costs and benefits in today's routing system. While the IETF does not typically consider the "business model" impacts of various technology choices, many participants felt that perhaps the time has come to review that philosophy.

2.4. How Urgent Are These Problems?

There was a fairly universal agreement among the workshop participants that the problems outlined in Section 2.1 and Section 2.2 need immediate attention. This need was not because the participants perceived a looming, well-defined "hit the wall" date, but rather because these are difficult problems that to date have resisted solution, are likely to get more unwieldy as IPv6 deployment proceeds, and the development and deployment of an effective solution will necessarily take at least a few years.

3. Current Stresses on the Routing and Addressing System

The primary concern voiced by the workshop participants regarding the state of the current Internet routing system was the rapid growth of the DFZ RIB. The number of entries in 2005 ranged from about 150,000 entries to 175,000 entries [BGP2005]; this number has reached 200,000 as of October 2006 [CIDRRPT], and is projected to increase to 370,000 or more within 5 years [Fuller]. Some workshop participants projected that the DFZ could reach 2 million entries within 15 years, and there might be as many as 10 million multihomed sites by 2050.

Another related concern was the number of prefixes changed, added, and withdrawn as a function of time (i.e., BGP UPDATE churn). This has a detrimental impact on routing convergence, since UPDATEs frequently necessitate a re-computation and download of the FIB. For example, a BGP router may observe up to 500,000 BGP updates in a single day [DynPrefix], with the peak arrival rates over 1000 updates per second. Such UPDATE churn problems are not limited to DFZ routes; indeed, the number of internal routes carried by large ISPs also threatens convergence times, given that such internal routes include more specifics, Virtual Private Network (VPN) routes, and other routes that do not appear in the DFZ [ATNAC2006].

3.1. Major Factors Driving Routing Table Growth

The growth of the DFZ RIB results from the addition of more prefixes to the table. Although some of this growth is organic (i.e., results simply from growth of the Internet), a large portion of the growth results from de-aggregation of address prefixes (i.e., more specific

prefixes). In this section, we discuss in more detail why this trend is accelerating and may be cause for concern.

An increasing fraction of the more-specific prefixes found in the DFZ are due to deliberate action on the part of operators [ATNAC2006]. Motivations to advertise these more-specifics include:

- Traffic Engineering, where load is balanced across multiple links through selective advertisement of more-specific routes on different links to adjust the amount of traffic received on each; and

- Attempts to prevent prefix-hijacking by other operators who might advertise more-specifics to steer traffic toward them; there are several known instances of this behavior today [BHB06].

3.1.1. Avoiding Renumbering

The workshop participants noted that customers generally prefer to have PI address space. Doing so gives them additional agility in selecting ISPs and helps them avoid the need to renumber. Many end-systems use DHCP to assign addresses, so a cursory analysis might suggest renumbering might involve modification of a modest number of routers and servers (perhaps rather than end hosts) at a site that was forced to renumber.

In reality, however, renumbering can be more cumbersome because IP addresses are often used for other purposes such as access control lists. They are also sometimes hard-coded into applications used in environments where failure of the DNS would be catastrophic (e.g., some remote monitoring applications). Although renumbering may be a mild inconvenience for some sites and guidelines have been developed for renumbering a network without a flag day [RFC4192], for others, the necessary changes are sufficiently difficult so as to make renumbering effectively impossible.

For these reasons, PI address space is sought by a growing number of customers. Current RIR policy reflects this trend, and their policy is to allocate PI prefixes to all customers who claim a need. Routing PI prefixes requires additional entries in the DFZ routing and forwarding tables. At present, ISPs do not typically charge to route PI prefixes. Therefore, the "costs" of the additional prefixes, in terms of routing table entries and processing overhead, is born by the global routing system as a whole, rather than directly by the users of PI space. The workshop participants observed that no strong disincentive exists to discourage the increasing use of PI address space.

3.1.2. Multihoming

Multihoming refers generically to the case in which a site is served by more than one ISP [RFC4116]. There are several reasons for the observed increase in multihoming, including the increased reliance on the Internet for mission- and business-critical applications and the general decrease in cost to obtain Internet connectivity. Multihoming provides backup routing -- Internet connection redundancy; in some circumstances, multihoming is mandatory due to contract or law. Multihoming can be accomplished using either PI or PA address space, and multihomed sites generally have their own AS numbers (although some do not; this generally occurs when such customers are statically routed).

A multihomed site using PI address space has its prefixes present in the forwarding and routing tables of each of its providers. For PA space, each prefix allocated from one provider's address allocation will be aggregatable for that provider but not the others. If the addresses are allocated from a 'primary' ISP (i.e., one that the site uses for routing unless a failure occurs), then the additional routing table entries only appear during path failures to that primary ISP. A problem with multihoming arises when a customer's PA IP prefixes are advertised by AS(es) other than their 'primary' ISP's. Because of the longest-matching prefix forwarding rule, in this case, the customer's traffic will be directed through the non-primary AS(s). In response, the primary ISP is forced to de-aggregate the customer's prefix in order to keep the customer's traffic flowing through it instead of the non-primary AS(s).

3.1.3. Traffic Engineering

Traffic engineering (TE) is the act of arranging for certain Internet traffic to use or avoid certain network paths (that is, TE puts traffic where capacity exists, or where some set of parameters of the path is more favorable to the traffic being placed there). TE is performed by both ISPs and customer networks, for three primary reasons:

- First, as mentioned above, to match traffic with network capacity, or to spread the traffic load across multiple links (frequently referred to as "load balancing").

- Second, to reduce costs by shifting traffic to lower cost paths or by balancing the incoming and outgoing traffic volume to maintain appropriate peering relations.

- Finally, TE is sometimes deployed to enforce certain forms of policy (e.g., Canadian government traffic may not be permitted to transit through the United States).

Few tools exist for inter-domain traffic engineering today. Network operators usually achieve traffic engineering by "tweaking" the processing of routing protocols to achieve desired results. At the BGP level, if the address range requiring TE is a portion of a larger PA 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.

In today's highly competitive environment, providers require TE to maintain good performance and low cost in their networks. However, the current practice of TE deployment results in an increase of the DFZ RIB; although individual operators may have a certain gain from doing TE, it leads to an overall increased cost for the Internet routing infrastructure as a whole.

3.2. IPv6 and Its Potential Impact on Routing Table Size

Due to the increased IPv6 address size over IPv4, a full immediate transition to IPv6 is estimated to lead to the RIB and FIB sizes increasing by a factor of about four. The size of the routing table based on a more realistic assumption, that of parallel IPv4 and IPv6 routing for many years, is less clear. An increasing amount of allocated IPv6 address prefixes is in PI space. ARIN [ARIN] has relaxed its policy for allocation of such space and has been allocating /48 prefixes when customers request PI prefixes. Thus, the same pressures affecting IPv4 address allocations also affect IPv6 allocations.

4. Implications of Moore's Law on the Scaling Problem

[Editor's note: The information in this section is gathered from presentations given at the workshop. The presentation slides can be retrieved from the pointer provided in Appendix D. It is worth noting that this information has generated quite a bit of discussion since the workshop, and as such requires further community input.]

The workshop heard from Tony Li about the relationship between Moore's law and the ability to build cost-effective, high-performance routers. The scalability of the current routing subsystem manifests itself in the forwarding table (FIB) and routing table (RIB) of the routers in the core of the Internet. The implementation choices for FIB storage are on-chip SRAM, off-chip SRAM, or DRAM. DRAM is commonly used in lower end devices. RIB storage is done via DRAM.

[Editor's note: The exact implementation of a high-performance router's RIB and FIB memories is the subject of much debate; it is also possible that alternative designs may appear in the future.]

The scalability question then becomes whether these memory technologies can scale faster than the size of the full routing table. Intrinsic in this statement is the assumption that core routers will be continually and indefinitely upgraded on a periodic basis to keep up with the technology curve and that the costs of those upgrades will be passed along to the general Internet community.

4.1. Moore's Law

In 1965, Gordon Moore projected that the density of transistors in integrated circuits could double every two years, with respect to minimum component cost. The period was subsequently adjusted to be between 18-24 months and this conjecture became known as Moore's Law [ML]. The semiconductor industry has been following this density trend for the last 40 or so years.

The commonly held wisdom is that Moore's law will save the day by ensuring that technology will continue to scale at the historical rate that will surpass the growth rate of routing information. However, it is vital to understand that Moore's law comes out of the high-volume portion of the semiconductor industry, where the costs of silicon are dominated by the actual fabrication costs. The customized silicon used in core routers is produced in far lower volume, typically in the 1,000-10,000 parts per year, whereas microprocessors are running in the tens of millions per year. This places the router silicon well off the cost curve, where the economies of scale are not directly inherited, and yield improvements are not directly inherited from the best current practices. Thus, router silicon benefits from the technological advances made in semiconductors, but does not follow Moore's law from a cost perspective.

To date, this cost difference has not shown clearly. However, the growth in bandwidth of the Internet and the steady climb of the speed of individual links has forced router manufacturers to apply more sophisticated silicon technology continuously. There has been a new generation of router hardware that has grown at about 4x the bandwidth every three years, and increases in routing table size have been absorbed by the new generations of hardware. Now that router hardware is nearing the practical limits of per-lambda bandwidth, it is possible that upgrades solely for meeting the forwarding table scaling will become more visible.

4.1.1. DRAM

In routers, DRAM is used for storing the RIB and, in lower-end routers, is also used for storing the FIB. Historically, DRAM capacity grows at about 4x every 3.3 years. This translates to 2.4x every 2 years, so DRAM capacity actually grows faster than Moore's law would suggest. DRAM speed, however, only grows about 10% per year, or 1.2x every 2 years [DRAM] [Molinero]. This is an issue because BGP convergence time is limited by DRAM access speeds. In processing a BGP update, a BGP speaker receives a path and must compare it to all of the other paths it has stored for the prefix. It then iterates over all of the prefixes in the update stream. This results in a memory access pattern that has proven to limit the effectiveness of processor caching. As a result, BGP convergence time degrades at the routing table growth rate, divided by the speed improvement rate of DRAM. In the long run, this is likely to become a significant issue.

4.1.2. Off-chip SRAM

Storing the FIB in off-chip SRAM is a popular design decision. For high-speed interfaces, this requires low-latency, high-capacity parts. The driver for this type of SRAM was formerly PC cache memory. However, this cache memory has recently been migrating directly onto the processor die, so that the volumes of cache memory have fallen off. Today, the primary driver for off-chip SRAM is cell phones, which require low-power, small-capacity parts that are not applicable to high-end router design. As a result, the SRAMs that are favored for router design are not volume parts. They have fallen off the cost curve and do not track with Moore's law.

4.2. Forwarding Engines

For many years, router companies have been building special-purpose silicon to provide high-speed packet-forwarding capabilities. This has been necessary because the architectural limitations of general purpose CPUs make them incapable of providing the high-bandwidth, low latency, low-jitter I/O interface for making high speed forwarding decisions.

As a result, the forwarding engines being built for high-end routers are some of the most sophisticated Application-specific Integrated Circuits (ASICs) being built, and are currently only one technological step behind general-purpose CPUs. This has been largely driven by the growth in bandwidth and has already pushed the technology well beyond the knee in the price/performance curve. Given that this level of technology is already a requirement to meet the performance goals, using on-chip SRAM is an interesting design

alternative. If this choice is selected, then growth in the available FIB is tightly coupled to process technology improvements, which are driven by the general-purpose CPU market. While this growth rate should suffice, in general, the forwarding engine market is decidedly off the high-volume price curve, resulting in spiraling costs to support basic forwarding.

Moreover, if there is any change in Moore's law or decrease in the rate of processor technology evolution, the forwarding engine could quickly become the technological leader of silicon technology. This would rapidly result in forwarding technology becoming prohibitively expensive.

4.3. Chip Costs

Each process technology step in chip development has come at increasing cost. The milestone of sending a completed chip design to a fabricator for manufacturing is known as 'tapeout', and is the point where the designer pays for the fixed overhead of putting the chip into production. The costs of taping out a chip have been rising about 1.5x every 2 years, driven by new process technology. The actual design and development costs have been rising similarly, because each new generation of technology increases the device count by roughly a factor of 2. This allows new features and chip architectures, which inevitably lead to an increase in complexity and labor costs. If new chip development was driven solely by the need to scale up memory, and if memory structures scaled, then we would expect labor costs to remain fixed. Unfortunately, memory structures typically do not seem to scale linearly. Individual memory controllers have a non-negligible cost, leading to the design for an internal on-chip interconnect of memories. The net result is that we can expect that chip development costs to continue to escalate roughly in line with the increases in tapeout costs, leading to an ongoing cost curve of about 1.5x every 2 years. Since each technology step roughly doubles memory, that implies that if demand grows faster than about (2x/1.5x) = 1.3x per year, then technology refresh will not be able to remain on a constant cost curve.

4.4. Heat and Power

Transistors consume power both when idle ("leakage current") and when switching. The smaller and hotter the transistors, the larger the leakage current. The overall power consumption is not linear with the density increase. Thus, as the need for more powerful routers increases, cooling technology grows more taxed. At present, the existing air cooling system is starting to be a limiting factor for scaling high-performance routers.

A key metric for system evaluation is now the unit of forwarding bandwidth per Watt-- [(Mb/s)/W]. About 60% of the power goes to the forwarding engine circuits, with the rest divided between the memories, route processors, and interconnect. Using parallelization to achieve higher bandwidths can aggravate the situation, due to increased power and cooling demands.

[Editor's note: Many in the community have commented that heat, power consumption, and the attendant heat dissipation, along with size limitations of fabrication processes for high speed parallel I/O interfaces, are the current limiting factors.]

4.5. Summary

Given the uncontrolled nature of its growth rate, there is some concern about the long-term prospects for the health and cost of the routing subsystem of the Internet. The ongoing growth will force periodic technology refreshes. However, the growth rate can possibly exceed the rate that can be supported at constant cost based on the development costs seen in the router industry. Since high-end routing is based on low-volume technology, the cost advantages that the bulk of the broader computing industry see, based on Moore's law, are not directly inherited. This leads to a sustainable growth rate of 1.3x/2yrs for the forwarding table and 1.2x/2yrs for the routing table. Given that the current baseline growth is at 1.3x/2yrs [CIDRRPT], with bursts that even exceed Moore's law, the trend is for the costs of technology refresh to continue to grow, indefinitely, even in constant dollars.

5. What Is on the Horizon

Routing and addressing are two fundamental pieces of the Internet architecture, thus any changes to them will likely impact almost all of the "IP stack", from applications to packet forwarding. In resolving the routing scalability problems, as agreed upon by the workshop attendees, we should aim at a long-term solution. This requires a clear understanding of various trends in the foreseeable future: the growth in Internet user population, the applications, and the technology.

5.1. Continual Growth

The backbone operators expect that the current Internet user population base will continue to expand, as measured by the traffic volume, the number of hosts connected to the Internet, the number of customer networks, and the number of regional providers.

5.2. Large Numbers of Mobile Networks

Boeing's Connexion service pioneered the deployment of commercial mobile networks that may change their attachment points to the Internet on a global scale. It is believed that such in-flight Internet connectivity would likely become commonplace in the not-too-distant future. When that happens, there can be multiple thousands of airplane networks in the air at any given time.

Given that today's DFZ RIB already handles over 200,000 prefixes [CIDRRPT], several thousands of mobile networks, each represented by a single prefix announcement, may not necessarily raise serious routing scalability or stability concerns. However, there is an open question regarding whether this number can become substantially larger if other types of mobile networks, such as networks on trains or ships, come into play. If such mobile networks become commonplace, then their impact on the global routing system needs to be assessed.

5.3. Orders of Magnitude Increase in Mobile Edge Devices

Today's technology trend indicates that billions of hand-held gadgets may come online in the next several years. There were different opinions regarding whether this would, or would not, have a significant impact on global routing scalability. The current solutions for mobile hosts, namely Mobile IP (e.g., [RFC3775]), handle the mobility by one level of indirection through home agents; mobile hosts do not appear any different, from a routing perspective, than stationary hosts. If we follow the same approach, new mobile devices should not present challenges beyond the increase in the size of the host population.

The workshop participants recognized that the increase in the number of mobile devices can be significant, and that if a scalable routing system supporting generic identity-locator separation were developed and introduced, billions of mobile gadgets could be supported without bringing undue impact on global routing scalability and stability.

Further investigation is needed to gain a complete understanding of the implications on the global routing system of connecting many new mobile hand-held devices (including mobile sensor networks) to the Internet.

6. What Approaches Have Been Investigated

Over the years, there have been many efforts designed to investigate scalable inter-domain routing for the Internet [IDR-REQS]. To benefit from the insights obtained from these past results, the workshop reviewed several major previous and ongoing IETF efforts:

1. The MULTI6 working group's exploration of the solution space and the lessons learned,

2. The solution to multihoming being developed by the SHIM6 Working Group, and its pros and cons,

3. The GSE proposal made by O'Dell in 1997, and its pros and cons, and

4. Map-and-Encap [RFC1955], a general indirection-based solution to scalable multihoming support.

6.1. Lessons from MULTI6

The MULTI6 working group was chartered to explore the solution space for scalable support of IPv6 multihoming. The numerous proposals collected by MULTI6 working group generally fell into one of two major categories: resolving the above-mentioned conflict by using provider-independent address assignments, or by assigning multiple address prefixes to multihomed sites, one for each of its providers, so that all the addresses can be topologically aggregatable.

The first category includes proposals of (1) simply allocating provider-independent address space, which is effectively the current practice, and (2) assigning IP addresses based on customers' geographical locations. The first approach does not scale; the second approach represents a fundamental change to the Internet routing system and its economic model, and imposes undue constraints on ISPs. These proposals were found to be incomplete, as they offered no solutions to the new problems they introduced.

The majority of the proposals fell into the second category--assigning multiple address blocks per site. Because IP addresses have been used as identifiers by higher-level protocols and applications, these proposals face a fundamental design decision regarding which layer should be responsible for mapping the multiple locators (i.e., the multiple addresses received from ISPs) to an identifier. A related question involves which nodes are responsible for handling multiple addresses. One can implement a multi-address scheme at either each individual host or at edge routers of a site, or even both. Handling multiple addresses by edge routers provides

the ability to control the traffic flow of the entire site. Conversely, handling multiple addresses by individual hosts offers each host the flexibility to choose different policies for selecting a provider; it also implies changes to all the hosts of a multihomed site.

During the process of evaluating all the proposals, two major lessons were learned:

- Changing anything in the current practice is hard: for example, inserting an additional header into the protocol would impact IP fragmentation processing, and the current congestion control assumes that each TCP connection follows a single routing path. In addition, operators ask for the ability to perform traffic engineering on a per-site basis, and specification of site policy is often interdependent with the IP address structure.

- The IP address has been used as an identifier and has been codified into many Internet applications that manipulate IP addresses directly or include IP addresses within the application layer data stream. IP addresses have also been used as identifiers in configuring network policies. Changing the semantics of an IP address, for example, using only the last 64-bit as identifiers as proposed by GSE, would require changes to all such applications and network devices.

6.2. SHIM6: Pros and Cons

The SHIM6 working group took the second approach from the MULTI6 working group's investigation, i.e., supporting multihoming through the use of multiple addresses. SHIM6 adopted a host-based approach, where the host IP stack includes a "shim" that presents a stable "upper layer identifier" (ULID) to the upper layer protocols, but may rewrite the IP packets sent and received so that a currently working IP address is used in the transmitted packets. When needed, a SHIM6 header is also included in the packet itself, to signal to the remote stack.

With SHIM6, protocols above the IP layer use the ULID to identify endpoints (e.g., for TCP connections). The current design suggests choosing one of the locators as the ULID (borrowing a locator to be used as an identifier). This approach makes the implementation compatible with existing IPv6 upper layer protocol implementations and applications. Many of these applications have inherited the long time practice of using IP addresses as identifiers.

SHIM6 is able to isolate upper layer protocols from multiple IP layer addresses. This enables a multihomed site to use provider-allocated

prefixes, one from each of its multiple providers, to facilitate provider-based prefix aggregation. However, this gain comes with several significant costs. First, SHIM6 requires modifications to all host stack implementations to support the shim processing. Second, the shim layer must maintain the mapping between the identifier and the multiple locators returned from IPv6 AAAA name resolution, and must take the responsibility to try multiple locators if failures ever occur during the end-to-end communication. At this time, the host has little information to determine the order of locators it should use in reaching a multihomed destination, however, there is ongoing effort in addressing this issue.

Furthermore, as a host-based approach, SHIM6 provides little control to the service provider for effective traffic engineering. At the same time, it also imposes additional state information on the host regarding the multiple locators of the remote communication end. Such state information may not be a significant issue for individual user hosts, but can lead to larger resource demands on large application servers that handle hundreds of thousands of simultaneous TCP connections.

Yet another major issue with the SHIM6 solution is the need for renumbering when a site changes providers. Although a multihomed site is assigned multiple address blocks, none of them can be treated as a persistent identifier for the site. When the site changes one of its providers, it must purge the address block of that provider from the entire site. The current practice of using the IP address as both an identifier and a locator has been strengthened by the use of IP addresses in access control lists present in various types of policy-enforcement devices (e.g., firewalls). If SHIM6's ULIDs are to be used for policy enforcement, a change of providers may necessitate the re-configuration of many such devices.

6.3. GSE/Indirection Solutions: Costs and Benefits

The use of indirection for scalable multihoming was discussed at the workshop, including the GSE [GSE] and indirection approaches, such as Map-and-Encap [RFC1955], in general. The GSE proposal changes the IPv6 address structure to bear the semantics of both an identifier and a locator. The first n bytes of the 16-byte IPv6 address are called the Routing Goop (RG), and are used by the routing system exclusively as a locator. The last 8 bytes of the IPv6 address specify an interface on an end-system. The middle (16 - n - 8) bytes are used to identify site local topology. The border routers of a site re-write the source RG of each outgoing packet to make the source address part of the source provider's address aggregation; they also re-write the destination RG of each incoming packet to hide the site's RG from all the internal routers and hosts. Although GSE

designates the lower 8 bytes of the IPv6 address as identifiers, the extent to which GSE could be made compatible with increasingly-popular cryptographically-generated addresses (CGA) remains to be determined [dGSE].

All identifier/locator split proposals require a mapping service that can return a set of locators corresponding to a given identifier. In addition, these proposals must also address the problem of detecting locator failures and redirecting data flows to remaining locators for a multihomed site. The Map-and-Encap proposal did not address these issues. GSE proposed to use DNS for providing the mapping service, but it did not offer an effective means for locator failure recovery. GSE also requires host stack modifications, as the upper layers and applications are only allowed to use the lower 8-bytes, rather than the entire, IPv6 address.

6.4. Future for Indirection

As the saying goes, "There is no problem in computer science that cannot be solved by an extra level of indirection". The GSE proposal can be considered a specific instantiation of a class of indirection-based solutions to scalable multihoming. Map-and-Encap [RFC1955] represents a more general form of this indirection solution, which uses tunneling, instead of locator rewriting, to cross the DFZ and support provider-based prefix aggregation. This class of solutions avoids the provider and customer conflicts regarding PA and PI prefixes by putting each in a separate name space, so that ISPs can use topologically aggregatable addresses while customers can have their globally unique and provider-independent identifiers. Thus, it supports scalable multihoming, and requires no changes to the end systems when the encapsulation is performed by the border routers of a site. It also requires no changes to the current practice of both applications as well as backbone operations.

However, all gains of an effective solution are accompanied with certain associated costs. As stated earlier in this section, a mapping service must be provided. This mapping service not only brings with it the associated complexity and cost, but it also adds another point of failure and could also be a potential target for malicious attacks. Any solution to routing scalability is necessarily a cost/benefit tradeoff. Given the high potential of its gains, this indirection approach deserves special attention in our search for scalable routing solutions.

7. Problem Statements

The fundamental goal of this workshop was to develop a prioritized problem statement regarding routing and addressing problems facing us today, and the workshop spent a considerable amount of time on reaching that goal. This section provides a description of the prioritized problem statement, together with elaborations on both the rationale and open issues.

The workshop participants noted that there exist different classes of stakeholders in the Internet community who view today's global routing system from different angles, and assign different priorities to different aspects of the problem set. The prioritized problem statement in this section is the consensus of the participants in this workshop, representing primarily large network operators and a few router vendors. It is likely that a different group of participants would produce a different list, or with different priorities. For example, freedom to change providers without renumbering might make the top of the priority list assembled by a workshop of end users and enterprise network operators.

7.1. Problem #1: Routing Scalability

The workshop participants believe that routing scalability is the most important problem facing the Internet today and must be solved, although the time frame in which these problems need solutions was not directly specified. The routing scalability problem includes the size of the DFZ RIB and FIB, the implications of the growth of the RIB and FIB on routing convergence times, and the cost, power (and hence, heat dissipation) and ASIC real estate requirements of core router hardware.

It is commonly believed that the IPv4 RIB growth has been constrained by the limited IPv4 address space. However, even under this constraint, the DFZ IPv4 RIB has been growing at what appears to be an accelerating rate [DFZ]. Given that the IPv6 routing architecture is the same as the IPv4 architecture (with substantially larger address space), if/when IPv6 becomes widely deployed, it is natural to predict that routing table growth for IPv6 will only exacerbate the situation.

The increasing deployment of Virtual Private Network/Virtual Routing and Forwarding (VPN/VRF) is considered another major factor driving the routing system growth. However, there are different views regarding whether this factor has, or does not have, a direct impact to the DFZ RIB. A common practice is to delegate specific routers to handle VPN connections, thus backbone routers do not necessarily hold

state for individual VPNs. Nevertheless, VPNs do represent scalability challenges in network operations.

7.2. Problem #2: The Overloading of IP Address Semantics

As we have reported in Section 3, multihoming, along with traffic engineering, appear to be the major factors driving the growth of the DFZ RIB. Below, we elaborate their impact on the DFZ RIB.

7.2.1. Definition of Locator and Identifier

Roughly speaking, the Internet comprises a large number of transit networks and a much larger number of customer networks containing hosts that are attached to the backbone. Viewing the Internet as a graph, transit networks have branches and customer networks with hosts hang at the edges as leaves.

As its name suggests, locators identify locations in the topology, and a network's or host's locator should be topologically constrained by its present position. Identifiers, in principle, should be network-topology independent. That is, even though a network or host may need to change its locator when it is moved to a different set of attachment points in the Internet, its identifier should remain constant.

From an ISP's viewpoint, identifiers identify customer networks and customer hosts. Note that the word "identifier" used here is defined in the context of the Internet routing system; the definition may well be different when the word "identifier" is used in other contexts. As an example, a non-routable, provider-independent IP prefix for an enterprise network could serve as an identifier for that enterprise. This block of IP addresses can be used to route packets inside the enterprise network. However, they are independent from the DFZ topology, which is why they are not globally routable on the Internet.

Note that in cases such as the last example, the definition of locators and identifiers can be context-dependent. Following the example further, a PI address may be routable in an enterprise but not the global network. If allowed to be visible in the global network, such addresses might act as identifiers from a backbone operator's point of view but locators from an enterprise operator's point of view.

7.2.2. Consequence of Locator and Identifier Overloading

In today's Internet architecture, IP addresses have been used as both locators and identifiers. Combined with the use of CIDR to perform route aggregation, a problem arises for either providers or customers (or both).

Consider, for example, a campus network C that received prefix x.y.z/24 from provider P1. When C multihomes with a second provider P2, both P1 and P2 must announce x.y.z/24 so that C can be reached through both providers. In this example, the prefix x.y.z/24 serves both as an identifier for C, as well as a (non-aggregatable) locator for C's two attachment points to the transit system.

As far as the DFZ RIB is concerned, the above example shows that customer multihoming blurs the distinction between PA and PI prefixes. Although C received a PA prefix x.y.z/24 from P1, C's multihoming forced this prefix to be announced globally (equivalent to a PI prefix), and forced the prefix's original owner, provider P1, to de-aggregate. As a result, today's multihoming practice leads to a growth of the routing table size in proportion to the number of multihomed customers. The only practical way to scale a routing system today is topological aggregation, which gets destroyed by customer multihoming.

Although multihoming may blur the PA/PI distinction, there exists a big difference between PA and PI prefixes when a customer changes its provider(s). If the customer has used a PA prefix from a former provider P1, the prefix is supposed to be returned to P1 upon completion of the change. The customer is supposed to get a new prefix from its new provider, i.e., renumbering its network. It is necessary for providers to reclaim their PA prefixes from former customers in order to keep the topological aggregatiblity of their prefixes. On the other hand, renumbering is considered very painful, if not impossible, by many Internet users, especially large enterprise customers. It is not uncommon for IP addresses in such enterprises to penetrate deeply into various parts of the networking infrastructure, ranging from applications to network management (e.g., policy databases, firewall configurations, etc.). This shows how fragile the system becomes due to the overloading of IP addresses as both locators and identifiers; significant enterprise operations could be disrupted due to the otherwise simple operation of switching IP address prefix assignment.

7.2.3. Traffic Engineering and IP Address Semantics Overload

In today's practice, traffic engineering (TE) is achieved by de-aggregating IP prefixes. One can effectively adjust the traffic volume along specific routing paths by adjusting the prefix lengths and the number of prefixes announced through those paths. Thus, the very means of TE practice directly conflicts with constraining the routing table growth.

On the surface, traffic engineering induced prefix de-aggregation seems orthogonal to the locator-identifier overloading problem. However, this may not necessarily be true. Had all the IP prefixes been topologically aggregatable to start with, it would make re-aggregation possible or easier, when the finer granularity prefix announcements propagate further away from their origins.

7.3. Additional Issues

7.3.1. Routing Convergence

There are two kinds of routing convergence issues, eBGP (global routing) convergence and IGP (enterprise or provider) routing convergence. Upon isolated topological events, eBGP convergence does not suffer from extensive path explorations in most cases [PathExp], and convergence delay is largely determined by the minimum route advertisement interval (MRAI) timer [RFC4098], except those cases when a route is withdrawn. Route withdrawals tend to suffer from path explorations and hence slow convergence; one participant's experience suggests that the withdrawal delays often last up to a couple of minutes. One may argue that, if the destination becomes unreachable, a long convergence delay would not bring further damage to applications. However, there are often cases where a more specific route (a longer prefix) has failed, yet the destination can still be reached through an aggregated route (a shorter prefix). In these cases, the long convergence delay does impact application performance.

While IGPs are designed to and do converge more quickly than BGP might, the workshop participants were concerned that, in addition to the various special purpose routes that IGPs must carry, the rapid growth of the DFZ RIB size can effectively slow down IGP convergence. The IGP convergence delay can be due to multiple factors, including

1. Delays in detecting physical failures,

2. The delay in loading updated information into the FIB, and

3. The large size of the internal RIB, often twice as big as the DFZ RIB, which can lead to both longer route computation time and longer FIB loading time.

The workshop participants hold different views regarding (1) the severity of the routing convergence problem; and (2) whether it is an architectural problem, or an implementation issue. However, people generally agree that if we solve the routing scalability problem, that will certainly help reduce the convergence delay or make the problem a much easier one to handle because of the reduced number of routes to process.

7.3.2. Misaligned Costs and Benefits

Today's rapid growth of the DFZ RIB is driven by a few major factors, including multihoming and traffic engineering, in addition to the organic growth of the Internet's user base. There is a powerful incentive to deploy each of the above features, as they bring direct benefits to the parties who make use of them. However, the beneficiaries may not bear the direct costs of the resulting routing table size increase, and there is no measurable or enforceable constraint to limit such increase.

For example, suppose that a service provider has two bandwidth-constrained transoceanic links and wants to split its prefix announcements in order to fully load each link. The origin AS benefits from performing the de-aggregation. However, if the de-aggregated announcements propagate globally, the cost is born by all other ASs. That is, the costs of this type of TE practice are not contained to the beneficiaries. Multihoming provides a similar example (in this case, the multihomed site achieves a benefit, but the global Internet incurs the cost of carrying the additional prefix(es)).

The misalignment of cost and benefit in the current routing system has been a driver for acceleration of the routing system size growth.

7.3.3. Other Concerns

Mobility was among the most frequently mentioned issues at the workshop. It is expected that billions of mobile gadgets may be connected to the Internet in the near future. There was also a discussion on network mobility as deployed in the Connexion service provided by Boeing over the last few years. However, at this time it seems unclear (1) whether the Boeing-like network mobility support would cause a scaling issue in the routing system, and (2) exactly what would be the impact of billions of mobile hosts on the global

routing system. These discussions were covered in Section 5 of this report.

Routing security is another issue that was brought up a number of times during the workshop. The consensus from the workshop participants was that, however important routing security may be, it was out of scope for this workshop, whose main goal was to produce a problem statement about addressing and routing scalability. It was duly considered that security must be one of the top design goals when we get to a solution development stage. It was also noted that, if we continue to allow the routing table to grow indefinitely, then it may be impossible to add security enhancements in the future.

7.4. Problem Recognition

The first step in solving a problem is recognizing its existence as well as its importance. However, recognizing the severity of the routing scaling issue can be a challenge by itself, because there does not exist a specific hard limit on routing system scalability that can be easily demonstrated, nor is there any specific answer to the question of how much time we may have in developing a solution. Nevertheless, a general consensus among the workshop participants is that we seem to be running out of time. The current RIB scaling leads to both accelerated hardware cost increases, as explained in Section 4, as well as pressure for shorter depreciation cycles, which in turn also translates to cost increases.

8. Criteria for Solution Development

Any common problem statement may admit multiple different solutions. This section provides a set of considerations, as identified from the workshop discussion, over the solution space. Given the heterogeneity among customers and providers of the global Internet, and the elasticity of the problem, none of these considerations should inherently preclude any specific solution. Consequently, although the following considerations were initially deemed as constraints on solutions, we have instead opted to adopt the term 'criteria' to be used in guiding solution evaluations.

8.1. Criteria on Scalability

Clearly, any proposed solution must solve the problem at hand, and our number one problem concerns the scalability of the Internet's routing and addressing system(s) as outlined in previous sections. Under the assumption of continued growth of the Internet user population, continued increases of multihoming and RFC 2547 VPN [RFC2547] deployment, the solution must enable the routing system to scale gracefully, as measured by the number of

- DFZ Internet routes, and

- Internal routes.

In addition, scalable support for traffic engineering (TE) must be considered as a business necessity, not an option. Capacity planning involves placing circuits based on traffic demand over a relatively long time scale, while TE must work more immediately to match the traffic load to the existing capacity and to match the routing policy requirements.

It was recognized that different parties in the Internet may have different specific TE requirements. For example,

- End site TE: based on locally determined performance or cost policies, end sites may wish to control the traffic volume exiting to, or entering from specific providers.

- Small ISP to transit ISP TE: operators may face tight resource constraints and wish to influence the volume of entering traffic from both customers and providers along specific routing paths to best utilize the limited resources.

- Large ISP TE: given the densely connected nature of the Internet topology, a given destination normally can be reached through different routing paths. An operator may wish to be able to adjust the traffic volume sent to each of its peers based on business relations with its neighbor ASs.

At this time, it remains an open issue whether a scalable TE solution would be necessarily inside the routing protocol, or can be accomplished through means that are external to the routing system.

8.2. Criteria on Incentives and Economics

The workshop attendees concluded that one important reason for uncontrolled routing growth was the misalignment of incentives. New entries are added to the routing system to provide benefit to specific parties, while the cost is born by everyone in the global routing system. The consensus of the workshop was that any proposed solutions should strive to provide incentives to reward practices that reduce the overall system cost, and punish the "bad" behavior that imposes undue burden on the global system.

Given the global scale and distributed nature of the Internet, there can no longer (ever) be a flag day on the Internet. To bootstrap the deployment of new solutions, the solutions should provide incentives to first movers. That is, even when a single party starts to deploy

the new solution, there should be measurable benefits to balance the costs.

Independent of what kind of solutions the IETF develops, if any, it is unlikely that the resulting routing system would stay constant in size. Instead, the workshop participants believed the routing system will continue to grow, and that ISPs will continue to go through system and hardware upgrade cycles. Many attendees expressed a desire that the scaling properties of the system can allow the hardware to keep up with the Internet growth at a rate that is comparable to the current costs, for example, allowing one to keep a 5-year hardware depreciation cycle, as opposed to a situation where scaling leads to accelerated cost increases.

8.3. Criteria on Timing

Although there does not exist a specific hard deadline, the unanimous consensus among the workshop participants is that the solution development must start now. If one assumes that the solution specification can get ready within a 1 - 2 year time frame, that will be followed by another 2-year certification cycle. As a result, even in the best case scenario, we are facing a 3 - 5 year time frame in getting the solutions deployed.

8.4. Consideration on Existing Systems

The routing scalability problem is a shared one between IPv4 and IPv6, as IPv6 simply inherited IPv4's CIDR-style "Provider-based Addressing". The proposed solutions should, and are also expected to, solve the problem for both IPv4 and IPv6.

Backwards compatibility with the existing IPv4 and IPv6 protocol stack is a necessity. Although a wide deployment of IPv6 is yet to happen, there has been substantial investment into IPv6 implementation and deployment by various parties. IPv6 is considered a legacy with shipped code. Thus, a highly desired feature of any proposed solution is to avoid imposing backwards-incompatible changes on end hosts (either IPv4 or IPv6).

In the routing system itself, the solutions must allow incremental changes from the current operational Internet. The solutions should be backward compatible with the routing protocols in use today, including BGP, OSPF, IS-IS, and others, possibly with incremental enhancements.

The above backward-compatibility considerations should not constrain the exploration of the solution space. We need to first find right solutions, and look into their backward-compatibility issues after

that. This way enables us to gain a full understanding of the tradeoffs, and what potential gains, if any, that we may achieve by relaxing the backward-compatibility concerns.

As a rule of thumb for successful deployment, for any new design, its chance of success is higher if it makes fewer changes to the existing system.

8.5. Consideration on Security

Security should be considered from day one of solution development. If nothing else, the solutions must not make securing the routing system any worse than the situation today. It is highly desirable to have a solution that makes it more difficult to inject false routing information, and makes it easier to filter out DoS traffic.

However, securing the routing system is not considered a requirement for the solution development. Security is important; having a working system in the first place is even more important.

8.6. Other Criteria

A number of other criteria were also raised that fall into various different categories. They are summarized below.

- Site renumbering forced by the routing system should be avoided.
- Site reconfiguration driven by the routing system should be minimized.
- The solutions should not force ISPs to reveal internal topology.
- Routing convergence delay must be under control.
- End-to-end data delivery paths should be stable enough for good Voice over IP (VoIP) performance.

8.7. Understanding the Tradeoff

As the old saying goes, every coin has two sides. If we let the routing table continue to grow at its present rate, rapid hardware and software upgrade and replacement cycles for deployed core routing equipment may become cost prohibitive. In the worst case, the routing table growth may exceed our ability to engineer the global routing system in a cost-effective way. On the other hand, solutions for stopping or substantially slowing down the growth in the Internet routing table will necessarily bring their own costs, perhaps showing up elsewhere and in different forms. Examples of such tradeoffs are

presented in Section 6, where we examined the gains and costs of a few different approaches to scalable multihoming support (SHIM6, GSE, and a general tunneling approach). A major task in the solution development is to understand who may have to give up what, and whether that makes a worthy tradeoff.

Before ending this discussion on the solution criteria, it is worth mentioning the shortest presentation at the workshop, which was made by Tony Li (the presentation slides can be found from Appendix D). He asked a fundamental question: what is at stake? It is the Internet itself. If the routing system does not scale with the continued growth of the Internet, eventually the costs might spiral out of control, the digital divide widen, and the Internet growth slow down, stop, or retreat. Compared to this problem, he considered that none of the criteria mentioned so far (except solving the problem) was important enough to block the development and deployment of an effective solution.

9. Workshop Recommendations

The workshop attendees would like to make the following recommendations:

First of all, the workshop participants would like to reiterate the importance of solving the routing scalability problem. They noted that the concern over the scalability and flexibility of the routing and addressing system has been with us for a very long time, and the current growth rate of the DFZ RIB is exceeding our ability to engineer the routing infrastructure in an economically feasible way. We need to start developing a long-term solution that can last for the foreseeable future.

Second, because the participants of this workshop consisted of mostly large service providers and major router vendors, the workshop participants recommend that IAB/IESG organize additional workshops or use other venues of communication to reach out to other stakeholders, such as content providers, retail providers, and enterprise operators, both to communicate to them the outcome of this workshop, and to solicit the routing/addressing problems they are facing today, and their requirements on the solution development.

Third, the workshop participants recommend conducting the solution development in an open, transparent way, with broad-ranging participation from the larger networking community. A majority of the participants indicated their willingness to commit resources toward developing a solution. We must also invite the participation from the research community in this process. The locator-identifier split represents a fundamental architectural issue, and the IAB

should lead the investigation into understanding of both how to make this architectural change and the overall impact of the change.

Fourth, given the goal of developing a long-term solution, and the fact that development and deployment cycles will necessarily take some time, it may be helpful (or even necessary) to buy some time through engineering feasible short- or intermediate-term solutions (e.g., FIB compression).

Fifth, the workshop participants believe the next step is to develop a roadmap from here to the solution deployment. The IAB and IESG are expected to take on the leadership role in this roadmap development, and to leverage on the momentum from this successful workshop to move forward quickly. The roadmap should provide clearly defined short-, medium-, and long-term objectives to guide the solution development process, so that the community as a whole can proceed in an orchestrated way, seeing exactly where we are going when engineering necessary short-term fixes.

Finally, the workshop participants also made a number of suggestions that the IETF might consider when examining the solution space. These suggestions are captured in Appendix A.

10. Security Considerations

While the security of the routing system is of great concern, this document introduces no new protocol or protocol usage and as such presents no new security issues.

11. Acknowledgments

Jari Arkko, Vince Fuller, Darrel Lewis, Tony Li, Eric Rescorla, and Ted Seely made many insightful comments on earlier versions of this document. Finally, many thanks to Wouter Wijngaards for the fine notes he took during the workshop.

12. Informative References

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

[RFC2547] Rosen, E. and Y. Rekhter, "BGP/MPLS VPNs", RFC 2547, March 1999.

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

[RFC4098] Berkowitz, H., Davies, E., Hares, S., Krishnaswamy, P., and M. Lepp, "Terminology for Benchmarking BGP Device Convergence in the Control Plane", RFC 4098, June 2005.

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

[RFC4192] Baker, F., Lear, E., and R. Droms, "Procedures for Renumbering an IPv6 Network without a Flag Day", RFC 4192, September 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.

[IDR-REQS] Doria, A. and E. Davies, "Analysis of IDR requirements and History", Work in Progress, February 2007.

[ARIN] "American Registry for Internet Numbers", http://www.arin.net/index.shtml.

[PIPA] Karrenberg, D., "IPv4 Address Allocation and Assignment Policies for the RIPE NCC Service Region", RIPE-387 http://www.ripe.net/docs/ipv4-policies.html, 2006.

[SHIM6] "Site Multihoming by IPv6 Intermediation (shim6)", http://www.ietf.org/html.charters/shim6-charter.html.

[EID] Chiappa, J., "Endpoints and Endpoint Names: A Proposed Enhancement to the Internet Architecture", http://www.chiappa.net/~jnc/tech/endpoints.txt, 1999.

[GSE] O'Dell, M., "GSE - An Alternate Addressing Architecture for IPv6", Work in Progress, 1997.

[dGSE] 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.

[PathExp] Oliveira, R. and et. al., "Quantifying Path Exploration in the Internet", Internet Measurement Conference (IMC) 2006, http://www.cs.ucla.edu/~rveloso/papers/imc175f-oliveira.pdf.

[DynPrefix] Oliveira, R. and et. al., "Measurement of Highly Active Prefixes in BGP", IEEE GLOBECOM 2005 http://www.cs.ucla.edu/~rveloso/papers/activity.pdf.

[BHB06] Boothe, P., Hielbert, J., and R. Bush, "Short-Lived Prefix Hijacking on the Internet", NANOG 36 http://www.nanog.org/mtg-0602/pdf/boothe.pdf, 2006.

[ROFL] Caesar, M. and et. al., "ROFL: Routing on Flat Labels", SIGCOMM 2006, http://www.sigcomm.org/sigcomm2006/discussion/showpaper.php?paper_id=34, 2006.

[CNIR] Abraham, I. and et. al., "Compact Name-Independent Routing with Minimum Stretch", ACM Symposium on Parallel Algorithms and Architectures, http://citeseer.ist.psu.edu/710757.html, 2004.

[BGT04] Bu, T., Gao, L., and D. Towsley, "On Characterizing BGP Routing Table Growth", J. Computer and Telecomm Networking V45N1, 2004.

[Fuller] Fuller, V., "Scaling issues with ipv6 routing+ multihoming", http://www.iab.org/about/workshops/routingandaddressing/vaf-iab-raws.pdf, 2006.

[H03] Huston, G., "Analyzing the Internet's BGP Routing Table", http://www.potaroo.net/papers/ipj/2001-v4-n1-bgp/bgp.pdf, 2003.

[BGP2005] Huston, G., "2005 -- A BGP Year in Review", http://www.apnic.net/meetings/21/docs/sigs/routing/routing-pres-huston-routing-update.pdf.

[DFZ] Huston, G., "Growth of the BGP Table - 1994 to Present", http://bgp.potaroo.net, 2006.

[GIH] Huston, G., "Wither Routing?", http://www.potaroo.net/ispcol/2006-11/raw.html, 2006.

[ATNAC2006] Huston, G. and G. Armitage, "Projecting Future IPv4 Router Requirements from Trends in Dynamic BGP Behaviour", http://www.potaroo.net/papers/phd/atnac-2006/bgp-atnac2006.pdf, 2006.

[CIDRRPT] "The CIDR Report", http://www.cidr-report.org.

[ML] "Moore's Law", Wikipedia http://en.wikipedia.org/wiki/Moore's_law, 2006.

[Molinero] Molinero-Fernandez, P., "Technology trends in routers and switches", PhD thesis, Stanford University http://klamath.stanford.edu/~molinero/thesis/html/pmf_thesis_node5.html, 2005.

[DRAM] Landler, P., "DRAM Productivity and Capacity/Demand Model", Global Economic Workshop http://www.sematech.org/meetings/archives/GES/19990514/docs/07_econ.pdf, 1999.

Appendix A. Suggestions for Specific Steps

At the end of the workshop there was a lively round-table discussion regarding specific steps that IETF may consider undertaking towards a quick solution development, as well as potential issues to avoid. Those steps included:

- Finding a home (mailing list) to continue the discussion started from the workshop with wider participation. [Editor's note: Done -- This action has been completed. The list is ram@iab.org.]

- Considering a special process to expedite solution development, avoiding the lengthy protocol standardization cycles. For example, IESG may charter special design teams for the solution investigation.

- If a working group is to be formed, care must be taken to ensure that the scope of the charter is narrow and specific enough to allow quick progress, and that the WG chair be forceful enough to keep the WG activity focused. There was also a discussion on which area this new WG should belong to; both routing area ADs and Internet area ADs are willing to host it.

- It is desirable that the solutions be developed in an open environment and free from any Intellectual Property Right claims.

Finally, given the perceived severity of the problem at hand, the workshop participants trust that IAB/IESG/IETF will take prompt actions. However, if that were not to happen, operators and vendors would be most likely to act on their own and get a solution deployed.

Appendix B. Workshop Participants

Loa Anderson (IAB)
Jari Arkko (IESG)
Ron Bonica
Ross Callon (IESG)
Brian Carpenter (IAB)
David Conrad (IANA)
Leslie Daigle (IAB Chair)
Elwyn Davies (IAB)
Terry Davis
Weisi Dong
Aaron Falk (IRTF Chair)
Kevin Fall (IAB)
Dino Farinacci
Vince Fuller
Vijay Gill

Russ Housley (IESG)
Geoff Huston
Daniel Karrenberg
Dorian Kim
Olaf Kolkman (IAB)
Darrel Lewis
Tony Li
Kurtis Lindqvist (IAB)
Peter Lothberg
David Meyer (IAB)
Christopher Morrow
Dave Oran (IAB)
Phil Roberts (IAB Executive Director)
Jason Schiller
Peter Schoenmaker
Ted Seely
Mark Townsley (IESG)
Iljitsch van Beijnum
Ruediger Volk
Magnus Westerlund (IESG)
Lixia Zhang (IAB)

Appendix C. Workshop Agenda

IAB Routing and Addressing Workshop Agenda
October 18-19
Amsterdam, Netherlands

DAY 1: the proposed goal is to collect, as complete as possible, a set of scalability problems in the routing and addressing area facing the Internet today.

0815-0900: Welcome, framing up for the 2 days
Moderator: Leslie Daigle

0900-1200: Morning session
Moderator: Elwyn Davies
Strawman topics for the morning session:
- Scalability
- Multihoming support
- Traffic Engineering
- Routing Table Size: Rate of growth, Dynamics
(this is not limited to DFZ, include iBGP)
- Causes of the growth
- Pains from the growth
(perhaps "Impact on routers" can come here?)
- How big a problem is BGP slow convergence?

1015-1030: Coffee Break

1200-1300: Lunch

1330-1730: Afternoon session: What are the top 3 routing problems
in your network?
Moderator: Kurt Erik Lindqvist

1500-1530: Coffee Break

Dinner at Indrapura (http://www.indrapura.nl), sponsored by Cisco

---
DAY 2: The proposed goal is to formulate a problem statement

0800-0830: Welcome

0830-1000: Morning session: What's on the table
Moderator: Elwyn Davies
- shim6
- GSE

1000-1030: Coffee Break

1030-1200: Problem Statement session #1: document the problems
Moderator: David Meyer

1200-1300: Lunch

1300-1500: Problem Statement session # 2, cont;
Moderator: Dino Farinacci
- Constraints on solutions

1500-1530: Coffee Break

1530-1730: Summary and Wrap-up
Moderator: Leslie Daigle

Appendix D. Presentations

The presentations from the workshop can be found on

http://www.iab.org/about/workshops/routingandaddressing

Authors' Addresses

David Meyer (editor)

EMail: dmm@1-4-5.net

Lixia Zhang (editor)

EMail: lixia@cs.ucla.edu

Kevin Fall (editor)

EMail: kfall@intel.com

Full Copyright Statement

Intellectual Property