Routing Scalability: An Operator's View

Xiaoliang Zhao, *Member, IEEE,* Dante J. Pacella, and Jason Schiller

***Abstract*—The Internet and its associated global routing table continues to grow with time. Will the current routing infrastructure be able to scale itself to sustain such growth? Over the past several years, many efforts have been devoted to address this important question. This paper presents a unique view from a network operator's perspective. We first clarify the definition of the routing scaling problem in practical terms by providing the relevant background information. We point out several reality issues that, if overlooked, may impede the adoption and deployment of a solution. Based on our operational experiences, we identify several requirements for potential solutions, and provide brief comments on the existing solutions.**

***Index Terms*—routing scalability, BGP, network operation**

I. Introduction

WHEN an IP packet arrives at a router, the essential job the router needs to do is to look up its routing table to determine the outgoing link, then to forward the packet to that link. The basic challenge in carrying out these simple operations is that the router must do it *very* fast, say, up to hundreds of millions of packets per second per line card as of today, which is also known as "line-rate forwarding". To achieve such forwarding performance, a router uses the fastest hardware such as ASIC chips to maintain a local copy of the full routing table on each line card. The full routing table contains information on how to reach *all* the reachable Internet addresses around the world. Today this table holds more than 300,000 network prefixes, and the number keeps increasing [9]. Why is it necessary to have a local copy of such a big table? Because it is simply infeasible to query a remote data structure to find the outgoing link information, since such query-response process induces communication delays which compromises the performance goal of line-rate forwarding. The line-rate forwarding requirement dictates that the routing system cannot be built as hierarchical and/or distributed routing tables, even though structures can scale very well. Moreover, since IP forwarding is done on a hop-by-hop basis, it implies that every router along the forwarding path must store such a full routing table locally [1]. As the Internet continues to grow, the full routing table is getting bigger and bigger. The ignited routing scalability issue, from an operator's point of view, comes down to a very simple and basic question: *do the routers in my network have enough memory to hold the full routing table?*

Another related factor to the routing scalability is the routing dynamics which demands certain amount of computational resources from a router to process the routing updates. Network events such as individual networks connect to or disconnect from the Internet, links going up and down, or routing devices failures, constantly change the topological connectivity of the Internet. Such dynamic changes lead to routing updates being propagated across the whole Internet. We closely monitor the workload imposed by routing dynamics on our routers, in particular the CPU usage. Although we observed some CPU usage spikes during BGP session resets, in most time our records show that routers have plenty of computational power to handle routing dynamics. Moreover, [8] examined the Internet routing dynamics for the past two years and concluded that routing dynamics has not changed much while the Internet continues to grow. Both observations lead to the same conclusion that routing dynamics is not a major routing scalability concern, at least for the near future.

The routing scalability issue has been discussed for years [1]–[3], [5], [7], [8], [10]. Most of the discussions are taking place in academia community. In this paper, we provide a different view from a network operator's perspective. We emphasize that this paper is our attempt to provide relevant specifics regarding routing scalability based on our own operational experiences. We describe relevant background information in Section II, and offer a forecast on the global routing table growth in Section III. Section IV discusses a number of issues in the operational reality regarding new technology deployment. Section V outlines several requirements that we feel important for a solution to be adopted by industry, and section VI provides a brief review on the solution space.

II. Background

Generally speaking, a typical carrier-grade router can be viewed as a distributed system within a single chassis. By distributing different functionality to different components, router vendors have achieved better scalability, flexibility and reliability. However, such distribution also leads to multiple copies of routing tables being replicated and stored across multiple components. As illustrated in Fig. 1, a typical modern router is composed of three major components: 1) a *control engine* running various routing protocols with its neighbor routers to collect routing information and to select the best routes; 2) a number of parallel *forwarding engines*, also known as *line cards*; and 3) a *switch fabric* connecting control engine and line cards. Each line card hosts one or more high-speed forwarding ASIC chips to forward IP packets at line rate. Forwarding ASIC chips are specially designed and engineered

Manuscript received 9 November 2009; revised 14 June 2010. Disclaimer: the opinions expressed in this paper are authors' own personal opinions and do not represent their employer's view in any way.

X. Zhao is with Network Research Center, Tsinghua University, Beijing, China, 100084 (e-mail: xleonzhao@tsinghua.edu.cn). The majority of this work was done when he was with Verizon Business Inc., USA.

D. Pacella and J. Schiller are with Verizon Business Inc., Ashburn, VA 20147 USA (e-mail: {dante.j.pacella, jason.schiller}@verizonbusiness.com).

Digital Object Identifier 10.1109/JSAC.2010.1009xx.

[1]Under certain cases, some routers may only install a default route or a small subset of the full routing table. But such cases are limited to edge routers, not the routers in the Default Free Zone (DFZ).

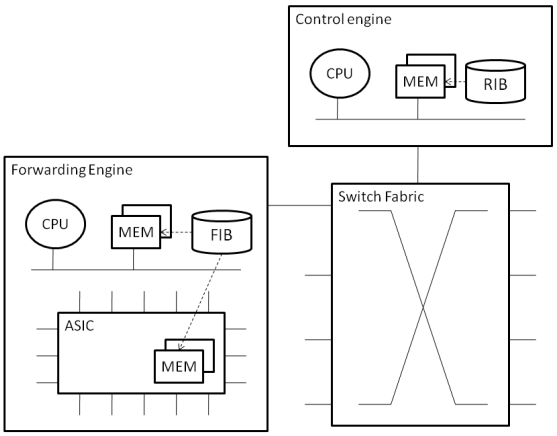


Fig. 1. An illustration of a typical modern router architecture

to meet the stringent performance requirements of high-speed line rate operations (currently at over 100Gbps per line card). Because of such special requirements, the line cards are the most expensive component in a router. Besides ASICs, a line card also has its own CPU and memory, usually running an embedded OS for control and management purposes.

Based on such a design, there are normally three copies of the routing table inside one router. First, the control engine maintains a large routing table called Routing Information Base (RIB) to store all network prefixes, including multiple possible paths for these destinations. The control engine selects the best routes from the RIB and installs them in a master forwarding table. Second, every line card maintains a local copy of the forwarding table in its own memory which mirrors the master forwarding table through real-time synchronization. Third, each forwarding ASIC chip maintains a highly compressed routing table in its built-in memory, usually implemented using a tree-based data structure to achieve fast IP routing lookup during packet forwarding process. This data structure inside ASIC chip is the ultimate source for a router to make the forwarding decision on which interface to forward the traffic out.

To achieve the optimal performance-cost ratio, different types of memory hardware are used to store different routing tables, depending on hardware capacity, speed, and cost. Because the routing information in the RIB is not directly used for packet forwarding, RIB memory normally uses larger, cheaper, and slower memory such as DRAM. On the other end, ASIC forwarding chip on a line card uses much faster but also more expensive and smaller memory such as SRAM. Line card memory hardware is somewhere in between. In most cases, the memory module is not a Field Replaceable Unit (FRU), thus it is not upgradable by users themselves; it will break the support contract if a user chose to do so. To expand memory capacity, vendors recommend to upgrade a routing engine or a line card as a whole. In this regard, it does not make a difference when either line card memory or ASIC built-in memory needs to be upgraded. With this in mind, we use a rather general but popular term, *Forwarding Information Base (FIB)* to refer to the routing table stored in either line card memory or ASIC memory.

Comparatively, a routing engine is much easier to upgrade than a line card in terms of cost and process. Unlike specially built line cards, a routing engine is often built using commercial off-the-shelf (COTS) hardware as long as it is capable to handle routing engine's functionality. In addition, a router is usually designed with redundant routing engines. A routing engine can be replaced during operation with almost no impact on traffic forwarding. On the other hand, replacing a line card has direct impacts on traffic, thus such activity must be planned and carried out carefully. Overall, RIB memory is a lot easier and cheaper to upgrade than FIB memory. Even beyond the physical difficulties of upgrading components, vendors are continually challenged to scale memory components without adversely impacting the cost. Continuing to add memory at a rate greater than Moore's Law has many challenges: power budget for the router, heat dissipation/cooling required and the diminishing return on air cooling, as well as the restricted footprint of the device due to the fact that components have a silicon landscape budget as well.

Different routers, depending on their capacity, functionality and cost, play different roles within an ISP. Typically, a router may play any of the following three main kinds of roles: 1) a *core router* is usually located at a major Point-of-Presence (POP) connecting to one or more core routers located at distant POPs to provide domestic or international long-haul transportation; 2) an *edge router* directly connects to customer networks to provide them the Internet connectivity; and 3) an *aggregation router* aggregates traffic from edge routers and then either sends the packets to core routers or loops them back to other edge routers within the same metropolitan area. Because a core router handles the aggregated traffic, it is often equipped with the high-speed and expensive line cards. Once a core router is unable to handle the ever-growing traffic or is about to be depreciated [2], it will be replaced by a newer model with the latest and greatest line cards. In most such cases, the old core router is still functional, so it makes perfect economical sense to keep it and turn it into either an aggregation router or an edge router. As time passes by, more and more legacy routers are gradually moved to the edge.

To a typical ISP, it is generally true that the number of its edge routers is proportional to the size of its customer base, while the number of the core routers is proportional to the number of POPs the ISP has a presence. The exact ratio of edge routers to core routers varies from network to network, but it is usually in the range of 4:1 to 10:1. Because of the large installation base of edge routers, upgrading all of them at one time is unlikely, if not impossible. Upgrades of edge devices are mainly driven by customers. When new customers require more ports to terminate them, or existing customers demand greater bandwidth, it is time to upgrade an edge router to a newer model with greater port density and faster interfaces.

[2]In terms of router replacement cycles, ISPs generally depreciate their equipment over a five year window. Add to that a six-month certification effort, and a one and a half year deployment time line, one may end up with a seven-year router replacement cycle.

III. Routing Table Growth

The Internet routing table growth has been mainly driven by the growth of Internet itself and deaggregation of the network address space. Regional Internet Registries (RIRs) such as ARIN, assign large blocks of IP addresses to service providers. The providers break up these large address blocks internally and assign individual pieces to different regions, POPs and various customers. Much of this deaggregation can be summarized externally, except in the cases of customer site multi-homing. Multi-homing benefits a customer with increased throughput and connectivity redundancy, but it also breaks the routing aggregation of the original CIDR design and adds more routing entries into the global routing table. From the protocol design's perspective, the current inter-domain routing protocol lacks of traffic engineering capability at AS level. Consequently BGP was often hacked with ad-hoc approaches, such as prefix de-aggregation or AS path prepending, to meet the real-world routing demands. Moreover, coupling the longest-match route selection with the absence of a built-in verification system in BGP, operators also have an incentive to intentionally announce more specific routes to prevent prefix hijacking.

The trend of the routing table growth is critical to understand the routing scalability issue. However, due to the uncertainty of the future, it is very challenging to make a fair prediction on such trend. Here we attempt to provide our empirical prediction based on our own experiences.

A. Assumptions

There will be two major events which affect future routing table growth: IPv4 address depletion and IPv6 adoption. It is predicted that all unallocated IPv4 addresses will be exhausted some time in 2011 or 2012 [20]. Past this date, IPv6 will be the only way to fulfill the demand for more IP addresses. IPv6 becomes an important aspect of business continuity for networks whose IP address needs continue to grow. Furthermore, new networks will be deployed as IPv6-only networks.

When old networks find a need to converse with these new IPv6-only networks, they will be driven to dual stack. It is possible that IPv4 to IPv6 Network Address Translation (NAT) may somewhat reduce the need for networks to go dual stack, but this approach has all of the same shortcomings as current IPv4 NAT. Furthermore deploying IPv4 to IPv6 NAT in the new IPv6-only networks requires new IPv4 addresses to interface with IPv4 networks, which may be difficult to obtain. Likewise deploying IPv4 to IPv6 NAT to legacy IPv4 networks requires the deployment of IPv6 which takes a similar amount of work as extending IPv6 to the Internet facing services of these legacy networks. This leads many to the conclusion that wide spread IPv6 adoption will likely occur around the time of IPv4 depletion for business continuity reasons.

ISPs generally upgrade their equipment over a seven year window. In terms of predicting the impact of routing table growth on routing hardware, one only needs to look within the seven year cycle. This means it becomes irrelevant to model the IPv6 adoption rate or carefully determine the IPv4 depletion rate if you believe there is a high likelihood that IPv4 depletion and wide spread IPv6 adoption will occur within the next seven year cycle. With this in mind, one approach to project the size of the IPv6 Internet table in seven years is to instead project the size of the IPv4 Internet table in seven years, and then interpolate from that table how big the IPv6 Internet table would be if all networks adopted dual stack and did IPv4 style multi-homing and traffic engineering.

The last assumption about routing table growth is that the routing table growth seen in the past is representative of future routing table growth. Without this assumption, one cannot make any reasonable routing table growth predictions. There are additional factors that might cause routing table growth in the future to be different such as fragmentation from an IPv4 market, IPv4 rationing to extend the life of IPv4, and a decrease in minimum allocation. The projection presented below did not take into account these considerations.

B. Projecting IPv4 Routing Table Growth

Default Free networks will need to carry all of the routes in the Default Free Zone (DFZ) [3]. In addition, each network will also carry additional routes about their internal topology, as well as more specific internal routes that may impact how their network delivers traffic, but may not impact how the greater Internet delivers traffic, such as static customers using the addresses assigned by their providers or BGP customers multi-homed to only a single AS. Such customers' addresses can be aggregated by their providers.

1) Impact of the Economic Downturn: The IPv4 Internet routing table growth data used in this document is derived from the CIDR report [21] up until October 2007. This is because in the early part of 2008, the routing table growth curve began to flatten out. It is possible that this slow-down of Internet routing growth is a result of the impact of the economic downturn.

The pre-2008 curve is steeper than Moore's Law. In early June 2008 through July 2009, the growth rate was flatter with the last three months being quite flat. Following this period of slow growth is a growth rate much steeper than the pre-2008 curve.

It is hard to judge the significance of the recent upturn in Internet routing table growth. It is possible that this is a blip that consists of simply noise. It is possible that this is a rebounding event that will result in returning the growth rate to the pre-2008 curve. It is also possible that this will be the new sustained growth rate. As such, this paper will use the pre-2008 curve as a basis for analysis.

2) Projecting IPv6 Routing Table Growth From The IPv4 Table: One approach to projecting the IPv6 Internet seven-year table size is to examine the seven-year IPv4 Internet table size and then correlate the IPv6 Internet table assuming all networks adopted dual stack and did IPv4-style multi-homing and traffic engineering. This is accomplished by dividing the IPv4 Internet table into three types of prefixes; aggregates, de-aggregates from growth into a non-contiguous space, and de-aggregates to perform traffic engineering. The assumption is that since all individual IPv6 address assignments are very

[3]One can trend the growth of the Default Free Zone (DFZ) from Geoff Huston's CIDR report found at [21].

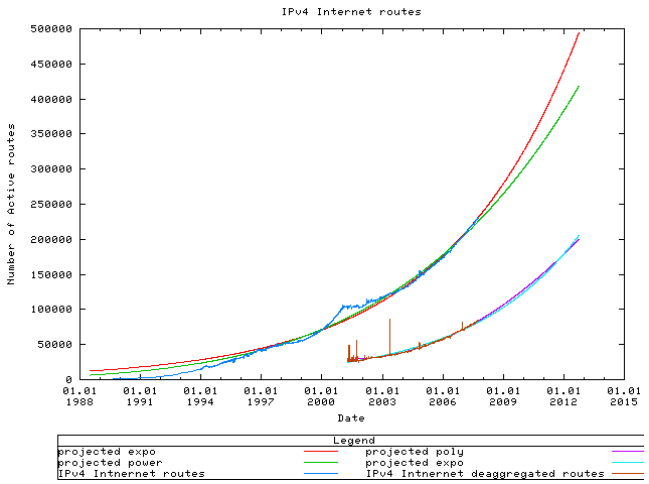


Fig. 2. IPv4 Internet Routes

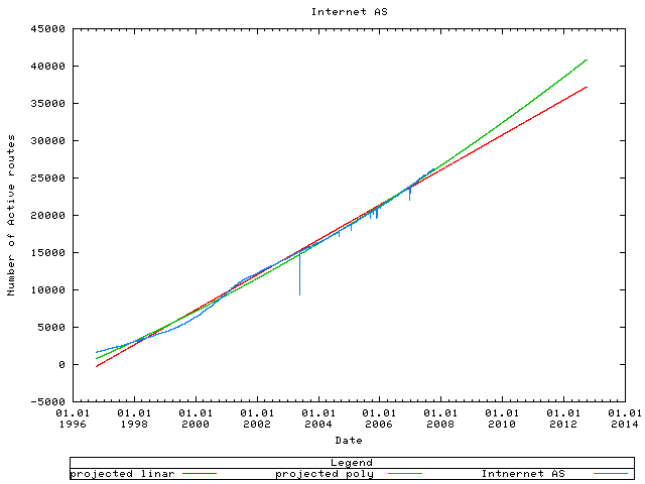


Fig. 3. IPv4 Active AS Growth.

large in size, the extra prefixes from growth into a non-contiguous space should not occur. Thus a network should only announce a single aggregate plus the number of more specifics needed to support multi-homing traffic engineering. As a result we could calculate the number of IPv6 routes in the Internet routing table as a function of the number of active ASes (each AS advertising one aggregate) and the additional more-specific prefixes needed for multi-homing traffic engineering.

a) Assumptions Not Accounted For: The above statement assumes that a single AS that currently announces multiple non contiguous addresses would only make one announcement if its non contiguous address blocks were replaced with a single contagious block. However one cannot determine with certainty whether the address announcements are separate only because they are not contiguous, or because they are both non-contiguous and used for separate traffic engineering. As a result this projection has erred on the conservative side, and assumed all non-contiguous addresses are not used for separate traffic engineering.

This projection does not account for all of the ASes that announce only a single /24 into the routing system because they are prohibited from announcing smaller prefixes. If instead they are given a large IPv6 assignment such as a /48, they might make multiple announcements for traffic engineering. Here this projection has erred on the conservative side, and assumed that these networks would only make one IPv6 announcement.

The projection does not account for the impact of Network Address Translation (NAT) in IPv6. Many people feel that with the abundance of IPv6 addresses, networks will no longer have a need to run Network Address Translation (NAT). Some of these networks may be using Network Address Translation (NAT) as their traffic engineering mechanism, by hiding behind multiple NAT addresses. With the removal of NAT, these networks will likely seek an alternate traffic engineering mechanism which might be advertising more specific routes. This is not accounted for.

This projection does not account for other possible impacting factors such as a grey market for IP space; changes in RIR IP addressing policy; more pervasive addressing needs, such as every handset, RFID tag, coffee maker, or other common appliance in every household in India and China. This projection also does not account for private VPN routes, which could be a sizable consideration in some networks.

C. Internet Table Growth

1) IPv4 Internet Table Growth: The graph in Fig. 2 depicts the growth of the IPv4 Internet table. The jagged blue line (the upper line) is the daily number of routes in the Internet table as reported by the CIDR report between 1988 and October of 2007. There are two corresponding curves projecting the growth rate of the IPv4 Internet table. The green curve is a best-fit power regression and the red curve is a best fit exponential.

As explained in Section III-B2, the jagged red line (the lower line) in Fig. 2 shows the daily number of intentional de-aggregates for traffic-engineering purpose, which is determined by taking the difference between the total number of IPv4 Internet routes and the IPv4 CIDR Aggregates. The blue curve is a best fit exponential projecting the growth of intentional de-aggregates.

2) IPv4 Active ASes: The graph in Fig. 3 depicts the growth of the number of active ASes of the IPv4 Internet table. The jagged blue line is the daily number of active ASes in the Internet table as reported by the CIDR report between 1988 and October of 2007. The red curve is a best-fit linear curve projecting the growth rate of the IPv4 Internet table.

3) Routing Table Growth of a Tier One ISP: One can interpolate the projected IPv6 Internet routing table by assuming each network will advertise one aggregate (the red curve in Fig. 3) as well as the number of intentional more specifics for multi-homing and TE (the blue curve in Fig.2). This can similarly be done for the internal routing table for a typical tier-1 ISP.

Fig. 4 depicts the predicted growth of the IPv4 and IPv6 routing table. These numbers reflect both the IPv4 and IPv6 Internet tables, as well as the IPv4 and IPv6 internal routing tables of a tier-1 ISP. The jagged green line (the upper jagged line) indicates the projected number of total IPv4 and IPv6

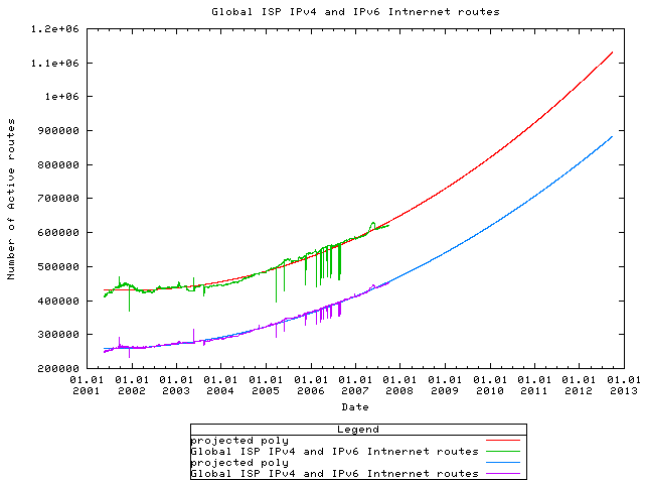


Fig. 4. IPv4 and IPv6 routing table growth for a Tier-1 ISP. The upper line shows the projected number of total IPv4 and IPv6 Internet and internal routes for the ISP assuming the high projection of 150,000 internal IPv4 routes, along with its best fit polynomial projection curve. The lower line shows the projected number of total IPv4 and IPv6 Internet and internal routes assuming the low projection of 50,000 internal IPv4 routes, along with its best fit polynomial projection curve.

Internet and internal routes for a tier-1 ISP assuming the high projection of 150,000 internal IPv4 routes. The red curve (the upper curve) is the best fit polynomial projection of the tier-1 ISP routing table growth. Similarly, the jagged purple line (the lower jagged line) indicates the projected number of total IPv4 and IPv6 Internet and internal routes for a tier-1 ISP assuming the low projection of 50,000 internal IPv4 routes. The blue curve (the lower curve) is the best fit polynomial projection of the tier one ISP routing table growth.

In short, if the routing table growth follows the predicted curve, the global routing table may have more than one million routes in next four or five years, and its size may exceed two millions routes in next ten years.

IV. Technology Trend vs. Deployment Reality

Similar to Moore's Law and Robert's Law [14], we found that the router's memory capacity also roughly follows a near-exponential growth. As shown in Table I[4], the router memory tends to double its size every couple of years, thus more or less matching an exponential growth curve. As of today, the latest router model has no problem to handle the full routing table. For example, one latest router is equipped with 4GB memory for control engine, 1GB memory for line card and 32MB ASIC memory. With about 300,000 network prefixes today [5], the typical memory utilization are well below 50% [6]. On the other hand, the recent routing table growth rate is more linear [9], and [8] predicted that such linear growth trend may remain for the next few years. Given the near-exponential memory capacity growth and a near-linear routing table growth, as questioned in [7], [8], do we really have a routing scalability problem? Well, what technology can do is

[4]Data is based on one specific vendor's product line.

[5]For most ISPs, they also need carry 110-140% more routes due to infrastructure address space and internally deaggregated prefixes.

[6]Actual percentage may vary from router to router, ISP to ISP.

TABLE I
Router Memory Capacity Growth

Year	RIB (MB)	Line card (MB)	ASIC (MB)
1998	256	128	8
2000	768	128	16
2002	2048	256	16
2004	2048	256	16
2006	2048	512	32
2007	3584	1024	32

one thing, what can be adopted and deployed in a production network is an entirely different thing. Let us take a closer look at the reality first.

It is often said that how much water a barrel can contain is determined by its shortest timber. If we apply the same rational here, we can roughly draw two corollaries.

- *Corollary 1*: The routing scalability problem for a router is determined by its smallest available memory pertaining to the routing table it stores.
- *Corollary 2*: The routing scalability problem for an ISP is determined by those routers which currently have the worst routing scalability problem.

Based on above understanding, as network operators, we do see an immediate routing scalability problem to address today. For example, a router purchased in 1998 only has 8MB ASIC memory, which has already approached its limit to handle today's full routing table. As we stated in Section II, there is a reason why such legacy routers are still running in a production network.

Today, such immediate concern is addressed on a case by case basis. For a router that cannot hold a full routing table, the easiest way is to install a default route when it makes sense, e.g., when an edge router only has one or two physical connections to its upstream router. But default route is a quite limiting option mainly because it gives away all routing knowledge and inevitably prevent us, as well as our customers, from engineering the traffic. As a tier-1 ISP, it is not uncommon that a downstream customer demands a full routing table for his/her own traffic engineering purpose. If the edge router connecting to the demanding customer happens to not have enough memory to hold the full routing table, a dedicated engineering effort will be needed to find a way to reduce the memory usage. One possible approach is to let an upstream router suppress those prefixes which are either aggregatable or fully internal. Then the "compressed" full routing table is sent to the incompetent edge router. Such compression can only improve the situation to a certain degree, and it is losing its effectiveness over time when the global routing table becomes more fragmented.

Undoubtedly, hardware upgrades is a readily available alternative to address routing scaling issue. However, hardware upgrades are not an economically sustainable solution. Per today's business model, an ISP's revenue mainly comes from providing the Internet connectivity or other Internet services to customers. Customers pay by physical circuits and bandwidth they rent. When customers request higher-speed circuits and pay extra for additional bandwidth, from business point of

view, it is easy to justify the cost to upgrade a legacy edge device. On the other hand, customers do not pay for adding network prefixes they sent to ISP's routing table. As a matter of fact, a customer literally can add as many network prefixes as he wishes at no additional cost. Because there is *no* direct connection between revenue generation and the routing table growth, again from business point of view, it is difficult to justify the cost of hardware upgrade associated to routing scalability problem. To resolve this dilemma, either the existing business model needs to be changed, or new technological solution needs to be developed to place an ultimate control over the routing table growth.

V. Requirements

In searching for a solution to the routing scalability problem, we summarize a few basic requirements from an ISP's perspective. By no means should this suggest a complete or mandatory list, rather the list intends to provide information in helping design a more deployable and practical solution. The requirements for a solution can be loosely grouped in three main categories of concern: Business Support, Compatibility, and OPEX Reduction.

A. Business Support

From past experiences, a new technology, when backed up by business support, is more likely to be adopted by industry. Business tends to welcome technologies which can either increase the revenue or decrease the cost. For example, VPN technology quickly becomes a popular service offered by many ISPs because it brings in new enterprise customers who requires the enhanced security and privacy than public Internet. Similarly, network based firewall and IDS technology were adopted to offer managed network security products to attract those customers who are seeking professional services to protect their networks. Ethernet now emerges as a promising long-haul transport technology to replace comparatively expensive SONET, mainly because significant cost savings are found by such transition. MPLS has been largely deployed in production networks for years. Compared to traditional intra-domain routing protocol like OSPF or ISIS, MPLS provides finer granularity traffic engineering and faster failure recovery capability. Such capability is critical to ISPs as it improves both the circuit utilization and customer satisfactions.

However, as discussed in Section IV, the routing table growth has yet found a direct connection to the revenue growth. It makes solving routing scalability problem especially challenging. If the situation does not change, an ISP will likely address the problem only on an as-needed basis. When that time comes, a solution with minimal cost and the least service impact is likely invoked. For a commercial ISP, the cost is the most determining factor, which not only includes the initial hardware and/or software purchase, but also the introduced changes to the existing back-end management systems, as well as the associated deployment and maintenance efforts.

Furthermore, the scaling stress impacts the largest networks first. Therefore, any solution should be able to be deployed gradually, thereby allowing smaller providers to forego what they may view as unnecessary migrations to reduce the total cost.

B. Backward Compatibility

Another important factor to be considered is to protect the existing heavy investments by ISPs on today's network infrastructure, including circuits and routing devices, data measurement infrastructure, back-end management systems, established business procedures, and so on. Backward compatibility is also critical to a solution. Radical changes, such as the one requires new hardware, flag-day event (network-wide change in same day), commissioning a new supporting infrastructure, or collaboration with other organizations, are not impossible, but more likely take longer time. IPv6 migration is a good example of such change. On the contrary, changes which are incremental, on the individual router basis, invisible to outside world are much easier to be tested and deployed.

In addition, any solution must at least support the current capabilities with regard to multi-homing, provider independence, and traffic engineering. The solution should not degrade the service or increase the risk of falling below Service Level Agreement (SLA) thresholds. Without addressing these capability and SLA probabilities, customers are likely to migrate away from ISPs that implement those methods.

C. OPEX Reduction

Operational costs with regards to time efficiency and network complexity are usually some of the most nebulous and difficult costs to quantify for network cost modeling. The most fundamental requirement resides in the deployment cost and Operational Expense (OPEX) associated with hardware upgrades. Large networks may be comprised of several hundred or even several thousand routers. Even component-level upgrade process may take a year or more from inception to completion in some large networks. With certification added to upgrade project timelines, the lifetime of the router's architecture may need to be up to seven years to fully depreciate the asset. Therefore, routers purchased in 2010 would need to have usability until at least 2017. Any practical solution must lower the initial and recurring expenses associated with operating and maintaining an ever-growing route table.

From network operation's point of view, it is desirable to have a new technology which is not only technologically sound, but also simple to understand and easy to operate. New products and other revenue drivers are easy to associate with training budgets, however, non-revenue related increases, such as those associated with external or ambient route growth, are more difficult to justify or have unanticipated increases as a Business As Usual (BAU) expense. Moreover, network operation staffs usually have full workload on a daily basis, which leaves less time for training and learning new concepts or new operational methodologies. Reducing the possible learning curve will facilitate a smooth deployment.

VI. Solution Space

In this section, we provide our view and comments on the solution space. To ease the discussion, we roughly categorized various proposed solutions into the following categories:

A. Business Model Change

There are some discussions about a new routing economy [19]. The concept is that those who operate routers in the Default Free Zone (DFZ) should be compensated to carry routes in the routing system. This approach is difficult to implement, because there is no easy mechanism to exchange money with all of the networks that make up the DFZ. Furthermore, there is a debate about whether particular networks in the DFZ have a real requirement to be default free. In addition, when distributing funds, one would have to consider the number of routers that carry the route, and the cost of upgrades. This could easily call into question particular routing designs, or upgrade practices and their associated costs. As each member of the DFZ would have to pay for routes they advertise to other members, the cost would have to be reflected in the services that each network provides. Ultimately, a routing economy would need to charge fees to their customers based on the number of prefixes each customer advertises.

An additional problem relates to smaller or regional networks. These smaller networks may carry the full external global routing table, but they do not have the same magnitude of internal or infrastructure prefixes and, therefore, would not need to upgrade their routers at the same rate. Creating such a routing economy would unnecessarily inflict burden upon these networks when they don't have such a premature need to upgrade their routing infrastructure to cope with the routing burden.

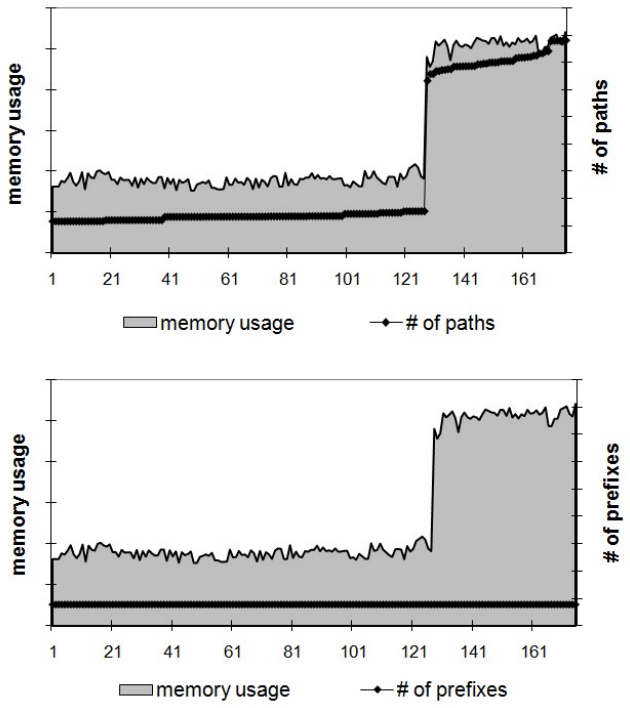


Fig. 5. RIB memory size vs. number of paths and number of prefixes. Each point at X-axis represents one router. The gray area shows the memory usage and the dotted line shows the number of paths or number of prefixes per router. To make it comparable, the same scale on right Y-axis is used in both figures. For the confidentiality purpose, the exact numbers have been removed. The figures indicates the memory usage increase is mainly determined by the increase of number of available AS paths.

B. Routing Architecture Re-design

The idea to separate location and identification address space has been discussed since 1990s and it remains as an attractive and promising direction today to resolve the routing scalability issue [12], [13]. The basic idea is to split the address space into two, one for the routing system (locator) and the other for end systems (identifier). The IP address prefixes for end systems will be moved out from today's DFZ routing infrastructure. Thus the size of the global routing table may be significantly smaller.

However, it seems to us that this type of new designs does not make the problem go away, rather it simply moves it into another problem domain. The mapping system in this new design faces the same challenges as today's routing architecture such as traffic engineering, scaling and synchronizing the database mapping identifiers to locators, routing security, and so on. As noted by Scudder [2], more efforts are needed to gain a solid understanding on those critical details. Our concern on new routing architecture design is that it has yet to demonstrate strong business supports, as first movers of the new designs do not see gains in their routing table reduction. In addition, since a typical development and deployment cycle takes about 3-5 years or even longer, re-architecture the routing infrastructure simply does not address the immediate scalability problem we have today.

C. Network Design Optimization

1) RIB size reduction: RIB memory size is mainly determined by two factors: one is the number of network prefixes, and the other is the number of available AS paths. By analyzing the routing table size and the memory usage data collected from more than 100 production routers, we found that the latter dominated the memory usage. In [16] McPherson *et al.* also made similar observations on the effect of multiple available paths. As shown in Fig. 5, when the number of available AS paths sharply increases (usually because of adding one or more BGP peers), the memory usage increases to a different level as well. The number of prefixes remains almost constant which indicates it is a least determining factor. Because the number of available paths is determined by the number of BGP peers, when engineering a routing infrastructure, one can limit the number of BGP peers for each router to control the RIB memory usage. For example, Route Reflector (RR) can greatly reduce the number of BGP peers for RR client routers.

2) BGP-free core vs. BGP-free edge argument: Some ISPs have built BGP-free core where a core router does not run BGP routing protocol, instead it forwards data via MPLS tunnels. However, a core router normally is more expensive as well as more resourceful than an edge router in terms of CPU power and memory size. Eliminating full routing table from core routers does not gain much with respect to routing scalability issue. Comparatively, edge routers have relatively limited resources, and in fact, many of them are legacy systems. Hence a BGP-free edge seems to be a better design goal. The most straightforward implementation is to install a default route on legacy edge routers, although this

approach may create sub-optimal routing and poor traffic engineering.

3) Virtual Aggregation: Virtual Aggregation (VA) [6] extended the concept of the default route by further dividing the whole IP address space into a small number of Virtual Prefixes (VP). Only VP routes are installed in the network, which greatly reduces the routing table size. Data traffic will be first forwarded to an Aggregation Point Router (APR), then redirected to the actual exit points through tunnels. Because of that, VA approach may introduce longer forwarding path and sub-optimal utilization of network resources. [6] proposed to resolve those problems by identifying and installing routes for a list of "popular prefixes" which carries the majority of data traffic. However, the operational cost to create and maintain such a dynamic popular prefix list is not trivial as it requires extensive traffic measurement and analysis. VA may also introduce new risks of possible network congestion, if the popular prefix list were misidentified or outdated. Moreover, certain software implementation attaches additional information, such as accounting, QoS, or firewall, to corresponding FIB entries. Removal FIB routes will also remove those extra bits which breaks the existing network operation. Overall, the VA approach looks promising to address immediate FIB memory emergency on individual routers, however, it is doubtful to be adopted as a network-wide solution due to its complexity and new issues introduced.

D. Software Optimization

As suggested by [17], for a near-term solution to address the immediate FIB memory shortage on legacy systems, one may consider possible software optimizations from vendors.

1) FIB size reduction: FIB Aggregation (FA), as detailed and evaluated in [15], is a local optimization of FIB table by removing more specific prefixes if they are already covered by an aggregated prefix with the same next-hop, or aggregating certain more specific prefixes which share the same next-hop. [15] reported that such optimization may reduce the current FIB table size by 30-50% if being "conservative", *i.e.,* keeping exact same routing and forwarding behavior as if no optimization. With more "aggressive" aggregation algorithms, one may reduce FIB size by 60-90% but it introduces extra routable space which may raise security concerns. Our internal tests revealed about 15-30% FIB reduction rate with the most conservative aggregation algorithm, while the actual number may vary network by network. FA is very promising in extending the lifetime of legacy systems for a couple of years, but at best it serves as a temporary patch to the overall routing scalability problem.

2) On-demand routing update: On-demand update of routing information may be another direction. The basic idea is FIB does not store a full routing table, instead it queries RIB for a particular route only when needed. The same idea can be realized at different levels, such as ASIC queries line card or master forwarding table, or an edge router queries a core router. The very concept has been implemented in early router models back to early 1990s, known as "route caching". However, the implementation encountered a big technical challenge on how to handle the initial traffic when the route was not available yet. The buffer inside a router was not large enough to hold the entire initial traffic. One way to handle this problem is to send the initial traffic to another routing device which has more resources, such as a core router. The idea was further examined in detail in [18].

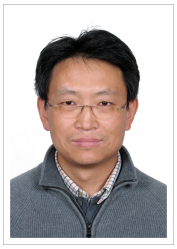
In summary, considering that there is no incentive yet for an ISP to resolve the routing scalability issue proactively, we feel FIB reduction may be the most promising solution for now. It only requires an software update internal to an individual router with no changes to existing hardware, operations or business procedures. Thus it meets almost all requirements outlined in Section V. FIB reduction offers a stopgap that can hopefully extend the lifetime of legacy devices for another couple of years while we are waiting for effective solutions being developed.

ACKNOWLEDGMENT

We would like to thank Lixia Zhang, Dan Massey, Gregory L. Stilwell and Sven Maduschke for their valuable comments and supports to this work.

REFERENCES

[1] D. Meyer, L. Zhang, and K. Fall (Ed.), *Report from the IAB Workshop on Routing and Addressing*, RFC 4984, Sep. 2007.
[2] John Scudder, *Routing/Addressing Problem: Solution Space*, ARIN XX Public Policy Meeting, Oct. 2007.
[3] T. Narten, *Routing and Addressing Problem Statement*, http://tools.ietf.org/html/draft-narten-radir-problem-statement-05, 2010.
[4] Y. Rekhter and T. Li, *An Architecture for IP Address Allocation with CIDR*, RFC1518, 1993.
[5] Jason Schiller, *Implications of Global IPv4/v6 Routing Table Growth*, ARIN XX Public Policy Meeting, Oct. 2007.
[6] P. Francis, X. Xu, H. Ballani, D. Jen, R. Raszuk, and L. Zhang, *FIB Suppression with Virtual Aggregation*, http://tools.ietf.org/html/draft-ietf-grow-va-02, 2010.
[7] K. Fall, G. Iannaccone, S. Ratnasamy, and P.B. Godfrey, *Routing Tables: Is Smaller Really Much Better?*, HotNets, 2009.
[8] Geoff Huston, *Scaling BGP - Or Not*, IEPG Meeting, http://www.potaroo.net/iepg/2010-03-ietf77/geoff.pdf, 2010.
[9] BGP Routing Table Analysis Reports, http://bgp.potaroo.net.
[10] The Routing Research Group (RRG), http://www.irtf.org/charter?gtype=rg&group=rrg.
[11] Y. Rekhter and T. Li, *An Architecture for IP Address Allocation with CIDR*, RFC1518, 1993.
[12] D. Farinacci, V. Fuller, D. Meyer, and D. Lewis, *Locator/ID Separation Protocol (LISP)*, Internet Draft, http://tools.ietf.org/html/draft-ietf-lisp-07, 2010.
[13] D. Jen, M. Meisel, D. Massey, L. Wang, B. Zhang, and L. Zhang, *APT: A Practical Tunneling Architecture for Routing Scalability*, UCLA Computer Science Department, Technical Report, 2008.
[14] L. Roberts, *Beyond Moore's law: Internet growth trends*, Computer, vol. 33, Jan. 2000, pp. 117-119.
[15] B. Zhang, L. Wang, X. Zhao, Y. Liu, and L. Zhang, *FIB Aggregation*, http://tools.ietf.org/html/draft-zhang-fibaggregation-02, 2009.
[16] D. McPherson, S. Amante, and L. Zhang, "The Intra-domain BGP Scaling Problem", NANOG 46, 2009.
[17] B. Zhang, L. Zhang, and L. Wang, *Evolution Towards Global Routing Scalability*, http://tools.ietf.org/html/draft-zhang-evolution-02, 2009.
[18] Changhoon Kim, Matthew Caesar, Alexandre Gerber, and Jennifer Rexford, *Revisiting Route Caching: The World Should be Flat*, in Proc. of PAM Conference, April 2009.
[19] Geoff Huston, *More ROAP Routing and Addressing at IETF68*, ISP Column, Feb. 2007.
[20] IPv4 Address Report, http://www.potaroo.net/tools/ipv4/index.html.
[21] CIDR Report, http://www.cidr-report.org/as2.0/.



Xiaoliang Zhao (M'01) received the B.S. and M. Eng. degree from Nankai University and Chinese Academy of Sciences, in 1995 and 1998, respectively, both in China. He earned his Ph.D. degree from Computer Science Department, North Carolina State University (NCSU) in 2002.

He is currently working at Network Research Center in Tsinghua University, Beijing, China. Previously, he worked at Verizon Business Inc., as a senior network engineer. His research interests include inter-domain routing and IPv6 deployment.

Dante J. Pacella graduated from Bryant & Stratton in 1994 before working on early Internet offerings from Avalon and LocalNet.

As Lead Network Architect, he has built large-scale IP backbones at Hyperion, Adelphia, and Verizon. At UUNET in the early part of this decade, he led efforts to address issues of scale in routing protocols. He has also led efforts to deploy one of the initial production instances of IP over both OC-192 and OC-768 as well as initial instances of multi-chassis routing. In 2003, he also developed Microburst Congestion Theory which has led to a shift in the way that QoS is implemented in IP backbones.

Mr. Pacella was also co-captain of the University of Buffalo Mad Turtles rugby team for two seasons 1989-90, leading them to a second-place finish in the Sevens Tournament at Delaware Park in 1990.

Jason Schiller graduated from American University in 1996 before working as a LAN Analyst at American University, a Network Support Specialist at Georgetown University, and a WAN Engineer at ManorCare where he maintained, supported, and augmented large enterprise networks. He first joined the UUNET High Speed Install Department before moving on to Global IP Network Engineering.

As a Senior Internet Network Engineer in the Global IP Network Engineering Department at Verizon, he is responsible for architecting, designing, evaluating, and qualifying networks for deployment in the UUNET / Verizon Business public IP network. He also completes field trials and acts as highest level of escalation. He is also responsible for defining and maintaining global standards for the UUNET / Verizon Business public IP networks.

Mr. Schiller continues to be active in the ARIN, NANOG, and IETF communities with particular interest in IPv4 exhaustion, IPv6 adoption, routing table growth, and the lack of a scalable solution for multi-homing and inter-AS traffic engineering. He is currently serving on the IETF Routing and Addressing Directorate (RADIR). His current term on the ICANN Address Supporting Organization Address Council (ICANN ASO AC) and the Number Resource Organization Number Council (NRO NC) expires December 31, 2010.