Scaling The Internet Routing System: An Interim Report

The Special Issue Editor Team

Today's Internet routing infrastructure provides connectivity among billions of computers around the global. The Internet has gone through a phenomenal growth over the last three decades, and its routing system has encountered a multitude of challenges brought forth by the unprecedented scale of the system. In addition to the growth in the number of end hosts and customer networks, there has been an increasing trend toward customer networks becoming multihomed with multiple service providers to facilitate load balancing and fail-over between providers, and customers' desire for provider-independent IP address assignments over provider-allocated IP addresses to avoid internal renumbering when changing providers. Unfortunately, multihoming and provider-independent addressing have led to fast growth of the global routing system as measured by routing table sizes and update frequencies. At the same time, Internet service providers (ISPs) face economic constraints that may prevent them from promptly upgrading to the latest technologies to meet the demands.

More recently, the Internet routing architecture has also confronted two new challenges: the imminent exhaustion of IPv4 address space and hence the foreseeable wide deployment of IPv6, and the emerging mobile access of the Internet from billions of hand-held devices. The latter further drives the demand for IPv6 roll out, yet the sheer size of the IPv6 address space presents a great scaling concern to the routing system. It is imperative to solve the routing scalability problem in order to enable continued growth of the Internet while allowing ISPs to operate with acceptable upgrade intervals.

Routing scalability has long been recognized as an outstanding issue in the Internet. As early as 1991, RFC1287 predicted that "The total number of IP network numbers will grow to the point where reasonable routing algorithms will not be able to perform routing based upon network numbers." Since then several changes have been introduced to the global routing architecture, notably the introduction of Classless InterDomain Routing (CIDR) during mid 90's to scale the routing system via provider-based prefix aggregation. However as mentioned above, customer multihoming and provider-independent addressing dominate today's Internet connectivity and defeat CIDR's ability to control routing table growth through provider-based aggregation. In 2006 the Internet Architecture Board (IAB) held a workshop on Routing and Addressing to develop a shared understanding of the routing scalability problems facing the large backbone operators and to foster community efforts towards a solution. This call has sparked a plethora of research efforts over the last few years. Proposed solutions are diverse, ranging from backwards-compatible, evolutionary techniques, to revolutionary clean-slate designs.

The purpose of this special issue is to expose the readers to the latest research results on Internet routing scalability. This issue includes a collection of twelve research contributions, which cover a diversity of topics on the Internet routing architecture and protocols. We group all the papers into three sections: 1) assessments of today's global routing system; 2) techniques to improve the performance of the existing routing system; and 3) new solutions that are designed to address the routing scalability problem by changing the basic model of the Internet routing architecture.

Assessment of Today's Routing System

This section includes three papers that assess the scalability issue from different angles in today's global routing system. The first paper, "Evolution of Internet Address Space Deaggregation: Myths and Reality" by Cittadini *et al.*, examines the trends in IP address deaggregation and routing system dynamics. The routing table growth and update churn are two prominent scaling concerns. A common perception suggests that site-multihoming and traffic engineering are driving an increasing number of autonomous systems (ASes) that deaggregate IP prefixes, and that a small number of edge networks (edge ASes) generate a disproportionate amount of the Border Gateway Protocol (BGP) update messages. The measured results presented in this paper show that there is no trend towards more aggressive prefix deaggregation or traffic engineering over time. Furthermore, deaggregated prefixes generally do not generate a disproportionate number of routing updates with respect to their share of the routing table entries. These conclusions are not meant as an argument to deny the problem of IP prefix deaggregation in the routing system. Rather, the results show that the impacts of IP address space deaggregation and routing churn due to a small number of edge networks have not changed for the worse in recent years as measured by their proportion in the overall system; they have been present for a long time and increase at the same pace as the Internet itself. The findings indicate that the growth of the global routing table and routing dynamics is largely due to the growth of edge networks.

The second paper in this group, "On the Scalability of BGP: the Role of Topology Growth" by Elmokash *et al.*, examines the relation between the Internet's AS-level topological structure and the resulting routing update rates. The authors estimate the number of routing updates received by ASes at different levels in the AS hierarchy and characterize the churn increase experienced as the network grows. Because

one does not know with certainty how the Internet's AS-level topology may evolve over time, the authors use simulations to examine a number of "what-if" growth scenarios that can be either plausible directions in the evolution of the Internet or educational corner cases, and observe several interesting results. First, connectivity density in the core of the network is the most important topological factor that decides the total number of routing updates generated after a failure. In particular, because mid-tier transit providers multihome to tier-1 ISPs, they play a role of multiplying update messages, thus the number of mid-tier transit providers and their multihoming degrees directly affect the number of routing updates. However peering links between ASes play a very different role than transit links with respect to the routing update scalability; the peering degree in the Internet does not affect churn in the global scale. Second, the depth of the hierarchical structure in the Internet plays a significant role. If the Internet were moving towards having all customer networks and content providers connect to mid-tier ISPs, the number of routing updates at tier-1 ASes would be much higher than if they connected to tier-1 ASes directly. Furthermore, increased topological connectivity density through increased multihoming degree may have a different impact on routing update counts depending on the locations and the failure types. Increased multihoming of edge ASes can increase churn after a prefix failure, but reduces churn after a network link failure. On the other hand, a more densely connected core increases churn after either edge prefix or edge link failures.

The last paper in this group, "Routing Scalability: An Operators View", is authored by Zhao, Pacella, and Schiller who are all frontline network operators. Thus this paper presents a unique assessment on the routing scalability challenges that large network operators face today. The paper first gives a description on the exact locations of scalability bottlenecks on a router as well as in a large provider network. It then provides an estimate of the global routing table growth in next few years based on the observed IPv4 network growth rate and a projected IPv6 deployment. The paper further explains why network operators may not be able to deploy the latest router technology to meet the routing system scalability needs. Under todays business model, customers pay their providers for the physical connectivity and bandwidth they use. When customers request higher-speed circuits and pay for additional bandwidth, a provider can easily justify the cost to upgrade a legacy edge router. On the other hand, customers do not pay for the network prefixes they announce to the provider; they may announce as many network prefixes as they wish at no additional cost. As a result, there is no direct connection between revenue and the routing table growth, making it difficult to justify the cost of hardware upgrade to address routing scalability issues. The paper also provides a list of requirements that can help make a solution deployable in practice, and concludes with a brief assessment of several proposed solutions.

Techniques to Improve the Performance of the Existing Routing System

The next four papers propose and evaluate techniques that can improve the performance of the existing Internet routing system. Measurements in recent years have revealed that normal operations of BGP can amplify a simple failure event into extended event sequences across the entire network. A specific example of such behavior is *path exploration*, where withdrawing a single prefix can result in a superfluous series of path-lengthening announcements at intermediate routers before the prefix is finally withdrawn from the entire network. The first paper in this group, "A Technique for Reducing BGP Update Announcements through Path Exploration Damping" by Huston *et al.* , develops a router-level mechanism, Path Exploration Damping (PED) to reduce unnecessary routing updates within an AS as well as to decrease the time to restore reachability. PED achieves the above goals by identifying likely transient routing updates, then delaying and suppressing their propagation. The authors evaluate the effectiveness of PED through trace-driven simulations using the BGP update traffic captured at two autonomous systems. Their results show that, compared to the conventional use of the Minimum Route Advertisement Interval (MRAI) to damp BGP updates, a PED-enabled BGP router can reduce the total number of BGP updates by up to 32% and reduce path exploration by up to 77%. The authors also discuss the feasibility of incrementally deploying PED in the global Internet and potential areas for future work.

The second paper, "Rate Limiting in an Event-Driven BGP Speaker" by Harris and Griffin, also addresses the issue of how best to rate-limit BGP update streams. As a departure from the traditional timer-based way of implementing BGP rate-limiting, this paper presents a lazy event-driven BGP route processing pipeline to accommodate rate limiting. To evaluate the design the authors implemented it in the open-source router package XORP, and called it Lazy XORP. The experimental results show that Lazy XORP can offer significant performance benefits over the standard XORP, as well as a number of other advantages.

The third paper in this group, "BGP Add-Paths: The Scaling/Performance Tradeoffs" by Schrieck *et al.* , provides a qualitative analysis of how to select paths when advertising multiple paths for the same destination prefix. Although by definition BGP routers and route reflectors only select one best path for each destination prefix and announce to their internal BGP (iBGP) neighbors, multiple proposals have been made recently that allow the propagation of multiple iBGP paths for the same prefix in order to increase system resilience and speed up failure recovery. This paper analyzes the various selection options for the paths to be advertised. The results show that one should use different selections to fulfill different application needs such as fast recovery upon failure or MED oscillation avoidance, and that the cost and benefit bounds with these selections depend on the connectivity of the AS where they are deployed. The authors also develop a tool to measure the scaling of path selections in a given network, and illustrate the utilization of this tool on synthetic Internet topologies.

The last paper in this group, "Keychain-based Signatures for Securing BGP" by Yin *et al.* , proposes a keychain-based signature scheme called KC-x to secure BGP. Compared to the existing work, KC-x has lower CPU and memory overheads, and provides stronger incentive for incremental deployment on the Internet. KC-x also has the flexibility of using different signature algorithms and allowing them to co-exist in a hybrid deployment. The authors have provided two implementations of KC-x: KC-RSA based on RSA and KC- MT based on Merkle hash trees. The experimental results use real BGP workloads to show that KC-RSA is as efficient as SAS-V (the most efficient software solution for aggregated path authentication), and KC-MT can be three times faster than SPV with 40% smaller signatures. Through the hybrid deployment of KC-MT and KC-RSA, KC-x can achieve both small signature and high processing rate for BGP routers.

New Designs towards A Scalable Routing System

Different from the above papers, the remaining five contributions in this special issue propose new protocol designs to address the routing system scalability problem. These designs are in various different stages of their development. The first two are short design papers, representing new designs whose quantitative evaluations are yet to be carried out. Despite the limited evaluations, they are included in this special issue both to provide wider exposure for the concepts and to provide a historical archive. The next three papers describe designs that have gone through a preliminary evaluation stage, however, each represents a rather different design direction.

The first short paper, "Evolving the Internet Architecture Through Naming" by Atkinson *et al.* , describes a new network layer protocol named Identifier-Locator Network Protocol v6 (ILNPv6). The basic idea in ILNPv6 is to replace the IP address with a combination of an identifier and a locator. The identifier identifies a node, and the locator is topologically significant and is used to route data packets to the destination node. More specifically, ILNPv6 divides the current 16-byte IPv6 address into two parts, the lower 64-bit Identifier is an IEEE Extended Unique Identifier (EUI-64) and the higher 64-bit Locator (L64) identifies a (sub)network. ILNPv6 addresses the routing scalability problem through the elimination of provider-independent addresses from the global routing system. Instead of allowing multihomed networks to inject their own prefixes into the global routing system as in today's practice, ILNPv6 assigns multiple locators to each host in a multihomed network, one from each of its providers. This enables prefix aggregation by providers in the global routing system. The routing system uses the Locator only for data delivery, while transport layer protocols bind their session state to the Identifier only. The realization of ILNPv6 requires modification to the Domain Name System (DNS), so that when one performs a DNS look up, one can learn both the identifier and the locator of the destination host. By allowing multiple L64 values to be bound to the same identifier simultaneously, INLPv6 can provide support for both site multihoming and host multihoming.

The ILNPv6 approach above represents a departure from today's practice; today each multihomed edge network uses only one IP address block. A number of other proposed solutions take a different approach that do not require changes to the operations of multihomed edge networks. This class of solutions is generally referred to as *Map-and-Encap*. Similar to ILNPv6, Map-and-Encap solutions also aim to remove the edge prefixes from the global routing system and to enable provider-based prefix aggregation. Different from ILNPv6, they map the prefixes of an edge network to the IP addresses of all the routers this edge network is attached to, and deliver each data packet by encapsulating it with the IP address of one of the routers that the destination network is attached to; this router address is commonly named $RLOC$, short for *routing locators*. Thus this class of solutions requires adding a new *mapping system* into the routing architecture, and several mapping system designs have been proposed in the last few years.

The second short paper, "FIRMS: a Future InteRnet Mapping System" by Menth *et al.* , presents one of the mapping system designs. In FIRMS, each prefix owner provides a map-base that holds the mappings for all its IP addresses to the RLOCs, and registers necessary access information about its map-base (called map-base pointer or simply MBP) in a global MBP distribution network which collects all MBPs and constructs a global MBP table. Another type of entity, Map-Resolver (MR), can receive and use this MBP table to find out which map-base to contact for a given edge IP address. When a router needs to forward a packet, it queries the MR to get the RLOC for the destination address of the packet.

The next paper in the new design group, "LISP-TREE: A DNS Hierarchy to Support the LISP Mapping System" by Jakab *et al.* , describes another mapping system design. Instead of starting from scratch as in the previous paper, the design of LISP-TREE closely follows the design of the successful Domain Name System. As the name suggests, LISP-TREE builds a hierarchical mapping system that inherits a number of proven features from DNS, including robustness and scalability. The authors evaluated their design through measurement-driven simulations and the results show that LISP-TREE over-performs several alternative solutions.

Different from the last two papers which focus on the mapping system design, the next paper, "MILSA: A New Evolutionary Architecture for Scalability, Mobility, and Multihoming in the Future Internet" by Pan *et al.* , attempts an overall routing architecture design. Similar to ILNPv6, the MILSA design separates locators from identifiers; different from ILNPv6, this design defines MILSA Identifier (MID) which represents several different kinds of identifiers, including User-ID, Host-ID, and Routing-infrastructure-ID. In addition to a scalable routing function, the MILSA design also supports multicast and manycast (where a packet can be delivered to a user with multiple locators for different devices or services). The authors recognize that, as a whole new architecture, the MILSA design still requires significant research and experiments to be done in the future. The paper evaluates the design using BGP

measurement data and concludes with an extended discussion on how to transit today's Internet architecture to MILSA.

The last paper in this special issue, "Evolution Towards Global Routing Scalability" by Khare *et al.* , takes a fundamentally different design approach than MILSA. It proposes an evolutionary process towards controlling routing scalability that is incrementally deployable and provides immediate benefits to any adopting ASes. The basic premise of this design is that route aggregation removes from routing tables the unnecessary topological details about remote portions of the Internet. All the proposed solutions achieve routing scalability by means of aggregation, however they differ on the specific enabling techniques for route aggregation. This paper states a position that route aggregation can be implemented with increasing scopes, starting from a router and then within a network and then gradually expanding to include more and more networks, to address the FIB, RIB and churn scaling problems. The paper argues for an evolutionary path for moving the global routing system towards scalability with an incentive-driven adoption of aggregation techniques. The paper evaluates the gain and cost of each aggregation step, and shows that local aggregation techniques can offer attractive tradeoffs to adopting networks without the deployment barriers inherent in other proposed solutions. Between this proposed evolutionary process and an entirely new design such as ILNP or MILSA, time will tell which one will eventually become the course taken.

We hope that the papers included in this special issue present a sample snapshot on the current understanding and results in scaling the global routing system, and that the readers will find this special issue timely, informative, and stimulating.

ACKNOWLEDGMENT

We would like to thank all the authors who submitted their work to this special issue. We would also like to thank all the reviewers for their time and valuable comments. These comments greatly improved the quality of this special issue. Finally, we would like to thank Steven Low, J-SAC Board Representative in charge of this special issue, for his enthusiastic support and guidance throughout the process.