Understanding IPv4 Prefix De-aggregation: Challenges for Routing Scalability

Roque Gagliano*, Eduardo Grampín†, Javier Baliosian†, Xavier Masip-Bruin† and Marcelo Yannuzzi†

*Latin America and Caribbean Regional Registry (LACNIC)
roque@lacnic.net
†Computer Science Institute (INCO), University of the Republic (UdelaR)
{grampin,baliosian}@fing.edu.uy
‡Advanced Network Architectures Lab (CRAAX), Technical University of Catalonia (UPC)
{xmasip,yannuzzi}@ac.upc.edu

Abstract—The Internet has become a vast and complex infrastructure. One of the aspects of deeper concern in the community is routing scalability, which involves both the size and dynamics of the global Internet routing table. Many Future Internet initiatives stand for “clean slate” or disruptive approaches to overcome this issue; nevertheless, network operators need evolutionary solutions. In this paper we put forward a characterization of the IP prefix de-aggregation factor focussing on the Latin American region, where we found that a significant contribution to the growth of the routing table could be prevented by improving BGP configuration. In the context of our participation in the LISP+ALT testbed (an evolutionary Future Internet initiative), we identify initial management requirements for its deployment, according to our findings regarding the size and dynamics of the routing table.

I. INTRODUCTION

Future Internet initiatives such as FIRE in Europe and GENI in USA emphasize the Internet global success, and the fact that it has become a critical infrastructure both from a social and economic perspective, after almost 30 years of its birth in the 1980s as a research network. The fiery growth occurred in the Internet poses demanding challenges in different fields such as scalability, mobility, heterogeneity and security, with added growing management complexity.

One of the issues of deep concern for the Internet community is the matter of routing scalability. The Internet routing system is largely based in the Border Gateway Protocol (BGP), a long lived path-vector protocol which is used to exchange reachability information between Autonomous Systems (ASes). Being a policy-routing protocol, it gives operators the freedom to express their enterprise requirements and policies, allowing the attachment of several attributes for each route or network prefix.

The interdomain traffic engineering (TE) tools available for stub Internet Service Providers (ISPs) are coarse-grained and consist primarily of the combination of BGP-based techniques, like the utilization of variable prefix lengths together with BGP attributes such as AS-path prepending. The basic idea behind these techniques is to de-aggregate IP prefixes, increasing in this way the granularity of the advertisements, allowing thereby to better control the distribution of their traffic over heterogeneous transit links. Issues like multihoming, suboptimal address allocation, and Classless Inter-Domain Routing (CIDR) block de-aggregation for TE purposes, among others, are alarmingly impacting on the scalability of the global Internet routing table, fueling its size and dynamics, and approaching an estimate of three hundred thousand entries by February 2009. Given that the contribution of IPv6 prefixes is merely 1500 entries, only IPv4 characterization is relevant at the present moment. However, it is worth highlighting that, the growth foreseen for IPv6 is also raising concern within the industry.

Since 1990s, the core router memory size and forwarding speed have managed to keep up with the growth, and in fact, core ISPs carry about a million routes internally (public BGP routes plus customer VPN routes). However, routers that can handle a million routes and forward packets at many Gbit/s are costly, and we could reach a hardware limit shortly (i.e. in the next five years). This and other aspects of the routing scalability problem has been in the Internet agenda for a long time; the adoption of BGP4 and CIDR provided a (short term) fix, but several monitoring infrastructures such as the CIDR Report [1], the BGP Report [2] and the Global BGP Routing Table report [3] have shown an alarming growth trend. The problem is fully characterized in [4].

Part of the research community state that the limitations of the current Internet architecture calls for long-term, “clean slate” or “disruptive” approaches that should take into account the complexity across layers from network connectivity to service architectures. Some examples of the challenging issues under consideration are: the balance between intelligence in the core versus in the edges; network neutrality and the end-to-end principle; the integration of network, compute and services infrastructures.

On the other hand, network operators and the telecommunications industry need some evolutionary steps towards this brand new architectures, in order to cope with current stringent operational issues. The Internet Research Task Force (IRTF) Routing Research Group (RRG) is considering non-disruptive proposals in this direction.

In this paper we present the results of studying a per AS
behavior in the Global BGP Routing Table, seeking to better understand IP prefix de-aggregation, and its implications for scalability management for Latin American service providers. This was done in the context of our participation in the deployment of an Internet-wide testbed to demonstrate the LISP+ALT architecture [5], which assisted us to introduce some requirements for an evolved LISP control plane.

The rest of the paper is organized as follows: in next section we briefly review relevant related work; in the following section III, we present our study of IPv4 prefix de-aggregation, and afterwards, in section IV we will be describing the LISP+ALT initiative. We close with a discussion and some directions for future work.

II. RELATED WORK

Reachability information for a given prefix originates from the AS to which the prefix belongs and is selectively propagated from AS to AS by means of BGP messages known as route updates, which can be announcements or withdrawals. While a number of previous work focus on BGP table growth, there is also an important body of related work which discuss the relevance of both active and passive measurement of BGP updates, in order to characterize the dynamics of the BGP routing table.

In [6], the authors characterize BGP Routing Table growth considering various contributing factors such as address fragmentation, multihoming and load balancing. Data is extracted from RouteViews [7], and the results are confirmed using measurements from a number of public accessible BGP information repositories. They conclude that address fragmentation, caused by multi-homing and load balancing is the major reason of BGP table growth. In [8], a systematic study of highly active prefixes is presented, concluding that a small fraction of advertised prefixes are responsible for a relevant amount of churn in BGP; furthermore, they found that some generators of BGP beacons, used for active monitoring of BGP updates, appear as highly active.

Despite the big amount of related work, the dynamics of the BGP control plane information (i.e., the exchanging of updates messages due to the advertisement of new prefixes) remains unknown, but certain evidence exists of Long Range Dependence [9]. Note that BGP flexibility (and its frequent mis-use by network administrators) often impose difficulties to determine the underlying cause of routing behavior. As BGP propagates changes to the best path, a single router may send multiple updates based on one triggering event, and further cause induced updates at other locations; examples of such events are link failures, newly added networks, prefix de-aggregation and policy changes, among others. It is even possible for policy conflicts to occur that can potentially disrupt the entire Internet.

Moreover, it is important to notice that, since the routing information is subject to successive filtering by ASes policies, any route view of the network is always partial, determined by the local point of observation.

To the best of our knowledge, there is no previous study of the relationship between BGP Table characteristics and the connectivity of the highly active ASes. Since BGP is a complex and error prone protocol, our study focuses on this single aspect, and it seeks to identify configuration management practices in small ISPs, particularly in Latin America.

III. IPv4 PREFIX DE-AGGREGATION

One of the important aspects of the scalability problem, evidenced by the routing information reports mentioned earlier, is that the current number of prefixes in the routing table is twice as much as the number that could be in place if perfect aggregation had occurred. It is important to remark the likely causes of prefix de-aggregation, as stated in [4]: multihoming, traffic engineering, non-aggregatable prefixes (particularly due to legacy address allocations), business events (such as mergers and acquisitions), leakage of iBGP outside of local ASes, reduction of Denial of Service (DoS) attacks and improve routing security by avoiding more specific routes hijacking.

As any study of the Global BGP Routing Table, the results are relative to the point of observation; we have selected the different locations from the RIPE (Réseaux IP Européens) RIS (Routing Information System) project [10]. The measurements were taken on the period November 2008-January 2009, considering only IPv4 prefixes; a total of 29,269 ASes, originating 275,639 IPv4 prefixes were studied1.

A. De-aggregation Factor (DF) Definition

De-aggregation of IP prefixes is produced by advertising CIDR blocks with longer prefixes than those allocated by the Regional Internet Registries (RIRs), for example, advertise two /20 instead of one /19 prefix. In order to quantify this, the Internet community has defined the De-aggregation Factor (DF), which represents a measure of the current routing table size vs. its aggregated size, and it is formally defined as follows:

\[
DF = \frac{\text{Prefixes in the Global Routing Table}}{\text{Aggregatable Prefixes}}
\]  

The global DF has been increasing steadily since records began in 1999 [3]. Ideally, DF=1 means that a certain AS is only announcing the aggregated prefixes.

B. DF Classification

Seeking a better understanding of the de-aggregation configuration management practices, we chose a per AS instead of a per prefix analysis. The global results are presented in Table I, which columns represent the DF and the relation between %ASes and %Prefixes (labeled AS/Pfx(%) in the table), sorted by region.

The global results summarized in Table I(a) show that the LACNIC region de-aggregates twice as much as the global average (corresponding to 4% of the AS and 9% of the prefixes in the Global BGP Routing Table). This results are

---

1Partial results have already been presented by R. Gagliano at the IEPG meeting in Minneapolis on November 2008, at the LACNIC XI meeting in Salvador on May 2008 and at the GTER meeting in Sao Paulo in October 2008.
also shown as histograms of the DF distribution, where we found that globally almost 80% of ASes have a DF lower than 1.5 (Figure 1(a)) and this percentage decreases to 52% in the LACNIC region (Figure 1(b)).

ASes prefix de-aggregation is classified using different criteria: first, by type of advertised prefixes, either legacy (pre-RIR) or RIR space (see Table I(b)). Second, considering the origin, either allocated (PA, originated by connectivity providers) or assigned (PI, originated by end-users) (see Table I(c)). And third, considering the relationship between prefix de-aggregation and AS size (measured by the number of originated prefixes, see Table I(d)). We also present the relationship between prefix de-aggregation and the number of upstream providers for each AS in Figure 2(a), and the relationship between prefix de-aggregation and the number of originated prefixes, see Table I(d)). We also present the relationship between prefix de-aggregation and the number of upstream providers for each AS in Figure 2(a), and the relationship between prefix de-aggregation and the number of upstream providers for each AS in Figure 2(b).

In Table I(b) we can observe that the number of ASes that only originate legacy space is 10% and they originate solely 4% of the global prefixes. These ASes are frequently targeted as being out of the periodic controls of RIR as they normally do not have a contract relationship with them. However, we can see that globally they behave in the same manner as the other categories, which means that we can disregard this classification for the prefixes in the rest of the study.

In Table I(c) we categorize ASes as connectivity providers (typically ISPs) or end users (such as corporations, content providers or governmental institutions). RIR and legacy address space is normally allocated to connectivity provider and then assigned to end users. We used the information specified by the RIRs to classify ASes by the prefixes they originate as: allocated only (usually named “PA” or Provider Allocated), assigned only (usually named “PI” or Provider Independent) or both (“PA & PI”). Those ASes that originates prefixes that do not fit the classification where marked as “Other”. We observe that regarding the DF, except for the ARIN region, the ASes originating only assigned prefixes (PI) behave better than those originating only allocated prefixes (PA). The relationship between the PA and the PI percentage gave us an idea of the Internet market in each region, where a bigger percentage of PI ASes means a stronger presence of end users (typically Content Delivery Networks and big corporations), as is the case of the RIPE region, and a bigger percentage of PA means a more significant participation of connectivity providers (as is the case of the APNIC region). In this case we see that while there is a higher DF for PA space than the global average, the proportional weight of these ASes (75% of the total AS in the APNIC region) is 20% higher than in the global figures.

In Table I(d) we show average DF per AS, sorted by the amount of addresses being originated by each AS. We conclude that even though being a small percentage of ASes in all regions, the biggest IP consumers have a deep impact in the overall IPv4 prefix de-aggregation because of their high DF and high percentage of originated prefixes. These big IP consumers generally represent allocated space (ISPs), while small IP consuming networks (<=20) represent the majority of ASes in every region, being the tier with the lower DF.

It is often argued that IP prefix de-aggregation is needed in order to implement multihoming with fault resilience and load balancing among unequal capacity links. Following this reasoning, if we analyze the relationship between the number of upstream providers and the DF, we would probably perceive that the more upstream providers an AS has, the more it will de-aggregate. Nevertheless, our collected data contradicts this supposition, showing that 59% of the ASes have only 1
upstream provider, and, of those, 19% have a DF greater than 1, with a maximum DF=48 and an average DF=1.42. These ASes represent single-homed networks from our observation point of the Global BGP Routing Table; the relationship between DF and the number of upstream providers is shown in Figure 2(a) (considering only ASes with DF<=$20$).

The implication of this finding is that a considerable fraction of the BGP Routing Table growth is caused by ASes which has no evident reason to de-aggregate IP prefixes, since they are single-homed. To further explore the incidence of “noisy” ASes, we built a Normalized BGP Update Rank in the interval (0,1), where 1 means the worst, i.e., the ASes which introduce more churn in BGP. This information is plotted in Figure 2(b), which shows that the contribution of top de-aggregators to the Global BGP Routing table dynamics is also high.

**C. Summary of Results**

The data analysis hereby presented introduces firm evidence that relatively small regions such as Latin America may introduce noticeable churn in the Global BGP Routing Table, due to higher dynamics and DF than average. Apparently, an important amount of this anomalies may be easily prevented, since they come from single-homed ASes which have no technical reasons to de-aggregate prefixes, nor employ other techniques such as AS-Path prepending. The key factor is that BGP enable system administrators to freely express site policies. But being a very complex protocol, if mis-used it may also cause noticeable harm to the global routing system scalability. This situation calls for enhanced BGP configuration management, and/or to suppress its usage entirely when it is not needed. For example, an important amount of prefix de-aggregation could be prevented by proper usage of BGP communities and prefix filtering.

In the next chapter we present LISP, which can help to reduce the usage of BGP in single-homed ASes, among other benefits.

**IV. AN EVOLUTIONARY PROPOSAL FOR ROUTING SCALABILITY: THE LISP+ALT TESTBED**

The proposals that are currently being discussed at IRTF RRG concerning the scalability issues of the global routing table can be taxonomized into two categories, namely, those that are host-based, and those that are network-based. It is arguable that network-based solutions offer a more promising approach than host-based ones. The concern about host-based solutions is twofold: on one hand, they introduce substantial changes to end-hosts (e.g., the IP stack needs to be augmented and upgraded). On the other hand, a significant part of the scaling benefits are obtained by eliminating PI addresses, which may arise some undesirable effects, such as the need for renumbering for all edge networks that own their IP prefixes. In contrast, network-based solutions may enable incremental
A. LISP Overview

A recent study show that the Locator-Identifier Separation Protocol (LISP), one of the solutions under discussion at the RRG, can help to reduce the size of the current routing table by roughly two orders of magnitude [11]. This work also shows that LISP provides improved interdomain TE capabilities through a non-disruptive approach, given that it can be adopted today without introducing major changes to the Internet. LISP Data Plane deploy IP-over-IP tunnels among border routers located at different domains, named Ingress/Egress Tunnel Routers (ITRs/ETRs, or xTRs). LISP separate the IP address space in Routing Locators (RLOCs) and End System Identifiers (EIDs). In LISP, only RLOC addresses are globally routable through the Internet, while EID addresses are considered routable only within their local domain. In order to facilitate multi-homing and portability, LISP assign PA addresses to RLOCs, and PI addresses to EIDs. Since an AS usually has several border routers, EID addresses can be associated with multiple RLOC addresses, which offers the opportunity to dynamically choose among several network paths between two endpoints, in order to fulfill Traffic Engineering objectives.

The splitting of addressing space demand the existence of a mapping function between EIDs/RLOCs addresses. When an end-system \( H_b \) wants to send a packet to another end-system \( H_o \) in a foreign AS, it has first to determine the address \( EID_b \) using the DNS, and after that, the xTRs have to locate a proper \( RLOC_o = map(EID_b) \) before encapsulating and sending the first packet to the remote entity \( H_o \), to avoid dropping of the initial packets in a connection. That’s why LISP is considered a “map’n’encap” solution. RLOCs are assumed to be limited and stable, since they are owned and run by major ISPs, resulting in a compressed (and stable) global routing table. Nevertheless, the impact of the mapping database size and dynamics are yet to be well understood, and care should be taken to prevent the reproduction of the same problems presented by the legacy BGP routing system. In the next section we discuss LISP+ALT, one of the LISP Control Plane schemes.

B. LISP+ALT testbed

LISP+ALT, named after ”LISP Alternate Topology”, is a Control Plane overlay whose purpose is to incrementally run the LISP architecture over the legacy Internet. The overlay network is built using Generic Routing Encapsulation (GRE) tunnels, and run BGP among LISP+ALT routers to propagate EID prefix update information, which is learned either over eBGP connections from the authoritative xTR, or by configuration. The ALT topology mainly route LISP control messages such as Data Probe, Map-Requests and Map- Replies. The mapping database is distributed (as described in [12]) and is stored in the xTRs. The RRG has recently introduced the LISP Map-Server [13], a computing system which provides a simple LISP protocol interface as a front end to the mapping database. The purpose of the Map-Server is to simplify the implementation and operation of LISP xTRs.

The LISP+ALT is driven by Cisco, while an open source implementation is also under development [14]. A global testbed has been set with one location hosted at our premises at UdelaR [15], which enable us to undertake the building and testing of a traffic engineering infrastructure which can both take advantage and enhance LISP capabilities. We’re building our own LISP implementation, with particular interest in the Map-Server functionality, which enable to build arbitrary mapping functionality out of the xTRs.

The status of LISP is being discussed at the IETF. While some members of the community argue that it worth creating a Working Group to pursue standardisation, other stand for keeping it in the research area. Note that shim6, a host-based initiative which is competitive in some aspects with LISP, has been running as a IETF Working Group for a while.

C. Requirements for routing scalability management

From our study of prefix de-aggregation, we concluded that an important fraction of churn in the the Global BGP Routing Table may be prevented by improved BGP configuration. LISP offers the possibility of completely eliminate the usage of BGP in stub sites, even if they are multi-homed. The complexity is therefore translated to the mapping database, which introduce, among others, the following issues: query/response latency, scalability, pull vs. push strategies, centralized vs. distributed architectures (i.e P2P and/or DNS-based mapping overlay models) and security aspects.

An Internet-wide deployment of LISP present several management challenges. Regarding configuration, one of the lessons learned from BGP is the need for provisioning tools to overcome protocol complexity. Moreover, alternatives to BGP should be provided for end users (i.e. stub ASes with TE requirements). As mentioned in Section IV-A, EID addresses can be associated with multiple RLOC addresses; therefore, AS-wide policies are needed for address space and mapping provisioning. This implies the existence of either an “omniscient” (centralized) or distributed coordination functionality among xTRs in a given domain. A related fault management concern is to assure RLOC reachability. Mappings are cached in xTRs, therefore is important to determine the validity of any given RLOC/EID mapping. Besides the four methods provided by LISP (see [12] section 6.3), a generalized reachability communication between xTRs is needed in order to ensure the consistency of reachability information.

Security of the mapping information exchange must be guaranteed; authentication by the entity with the EIDs right of use and also cross-authentication by the entity with the mapped RLOC right of use shall be implemented. In the current LISP implementation an EID hijacking can be easily accomplished in the following way: an xTR may receive a map-request for a certain EID (ex. 240.0.0.1) and respond with a map-reply that includes that particular EID but also a much larger address space (such as 240.0.0.0/8), generating a “bogus” entry in the...
mapping cache. In order to detect and solve this and similar kind of failures, traceability of the mapping database activity is needed.

Regarding performance and accounting, LISP shall provide statistics to enable fine-grain provisioning with Traffic Engineering objectives. As previously mentioned, LISP mapping system can return multiple RLOC addresses for the same destination EID address. Each of the entries returned has a priority and a weight attribute. The priority determines the order in which the ETRs must be selected, while the weight tells how to distribute the traffic among ETRs with the same priority. A proper monitoring function is therefore decisive to define priorities and weights in line with TE objectives.

Note, again, that the Map-Server functionality, being tightly integrated in the LISP control plane, can help to fulfill some of the aforementioned requirements.

V. CONCLUSION AND FUTURE WORK

This paper present a per AS study of the de-aggregation factor, and concludes that a small percentage of “big” ASes, together with an important amount of “small” single-homed ASes make a considerable contribution to the size and dynamics of the global Internet routing table, particularly in Latin America. Single-homed ASes “noise” calls for improved BGP configuration management in order to “clean” unnecessary BGP prefix de-aggregation. The presented DF classification is a snapshot of the situation at the time of this writing; we plan to maintain a permanent processing of BGP information, with sources diversification.

We pointed out that LISP is an opportunity for routing table size reduction and stabilization; in particular it offers the possibility avoid using BGP in stub ASes (both single and multi-homed), and also provides means for evolved, fine-grained traffic engineering. Moreover, the reduction of BGP routing table size provides a growth path for IPv6 deployment.

Nevertheless, we state that the mapping functionality is a major concern of the LISP deployment; our participation in the LISP+ALT evolutionary Future Internet initiative enable us to identify some preliminary management requirements. In this regard, our basic idea is to benefit from the LISP Map-Server functionality and related protocols, which enable us to build mapping strategies to conceal intra and inter-domain traffic engineering objectives, using our previous experience in Intelligent Route Control (IRC) [16] and PCE-based TE [17]. Online computation and provisioning of mappings based on TE policies may be achieved using IRC-like techniques, Constraint-Based Routing algorithms and/or classical optimization techniques, such as the IDIPS prototype implementation [18]. We’re also exploring the applicability of LuPA [19] to this problem. This Autonomic Management infrastructure can be used to pursue general objectives, and has been designed to have a very small footprint; therefore, it may be applied in the RLOC/EID mapping context with very fast response times. Our preliminary idea is to build an overlay of agents with common objectives, for example, comprising ASes which exchange an important share of Internet traffic, and/or grouping of ASes as a consequence of merges and acquisition. Therefore, a fundamental line of future work is to build a working TE prototype over the global testbed, in order to validate the proposal and evaluate the impact in the routing table size.

ACKNOWLEDGMENT

The authors would like to thank the networking staff at UdelaR for the support building and running our branch of the LISP+ALT testbed.

REFERENCES