# Insights on the Internet routing scalability issues.

Alberto Castro, Martín Germán, Marcelo Yannuzzi and Xavi Masip-Bruin
Advanced Network Architectures Lab (CRAAX), Universitat Politecnica de Catalunya (UPC)

***Abstract*—In recent years, the size and dynamics of the global routing table have increased rapidly along with an increase in the number of edge networks. The relation between edge network quantity and routing table size/dynamics reveals a major limitation in the current architecture. In this paper we introduce the two problematics target as the main cause for the Internet scalability issue. Subsequently, we describe the different proposals that address the scalability problem. We group them in three categories: Separation, Elimination and Geographic.**

## I. Introduction

A recent workshop report by the Internet Architecture Board (IAB) [1] revealed that Internet routing is facing a serious scalability problem. The current global routing table size in the default-free zone (DFZ) has been growing at an alarming rate over recent years [2] (see Fig. 1), despite the existence of various constraints such as a shortage of IPv4 addresses and strict address allocation and routing announcement policies. Though the deployment of IPv6 will remove the address shortage, there is an increasing concern that wide-scale IPv6 deployment could result in a dramatic increase of the routing table size, which may exceed our ability to engineer the operational routing system.

The workshop identified the following factors as the main driving forces behind the rapid growth of the DFZ RIB:

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

The major contributor to the growth of the routing table is site multihoming, where individual edge networks connect to multiple service providers for improved availability and performance. In the presence of network failures, a multihomed edge network remains reachable as long as any one of its providers remains functioning. In the absence of failures, the edge network can utilize multiple-provider connectivity to maximize some locally defined goals such as higher aggregate throughput, better performance, and less overall cost. However, for an edge network to be reachable through any of its providers, the edge networks address prefix(es) must be visible in the global routing table. In other words, no service provider can aggregate a multihomed edge networks prefix into its own address prefix, even if the edge network may be using a provider-assigned (PA) address block. In addition, more and more edge networks are getting provider-independent (PI) address allocations that come directly from the Regional Internet Registries to avoid renumbering when changing providers. In short, multihoming destroys topology-based prefix aggregation by providers and leads to fast global routing table growth.

Routing table size is not the only scalability concern. Equally important is the amount of updates the system must process. Under the current, flat inter-domain routing system [3], a connectivity flap to any destination network may trigger routing updates to propagate throughout the entire Internet, even when no one is communicating with the unstable destination network at the time. Several measurement studies have shown that the overwhelming majority of BGP updates are generated by a small number of edge networks (e.g., [4]). Unfortunately, a large-scale, decentralized system such as the Internet will surely contain a small number of poorly managed or even suspicious components.

The other problematic identified at the workshop is the overloading of IP address semantics. One of the fundamental assumptions underlying the scalability of routing systems was eloquently stated by Yakov Rekhter (and is sometimes referred to as ”Rekhter's Law”), namely:

*“Addressing can follow topology or topology can follow addressing. Choose one.”*

Following this idea some authors (e.g.: [5]) have tried to provide the architecture for a scalable routing system by making use of aggressive topological aggregation. Unfortunately there is some difficulty in creating and maintaining the envisioned congruence. This difficulty arises from the overloading of addressing with semantics of both end users identifiers and router locators: there is the need to identify both clients and routers and only one number space is available. Either way, the overloading has been felt and moreover deemed to have had profound implications for the scalability of the global routing system.

This paper presents the different strategies proposed to overcome the Internet routing scalability issues.

## II. Solution Strategies

In this section we first introduced the design goals for the scalable Internet routing as the Routing Research Group defined them [6]. Next, the different solutions strategies are presented. These strategies are analyzed as abstract architectural issues, not emphasizing in any particular architectural proposal. The different approaches are grouped following the categorization defined in [7] (i.e.: Separation

This work has been partially supported by the Spanish project ”Redes Multinivel: IP sobre redes de transporte” under contract TEC2008-02552-E and by the Catalan Government under contract 2009 SGR1508.

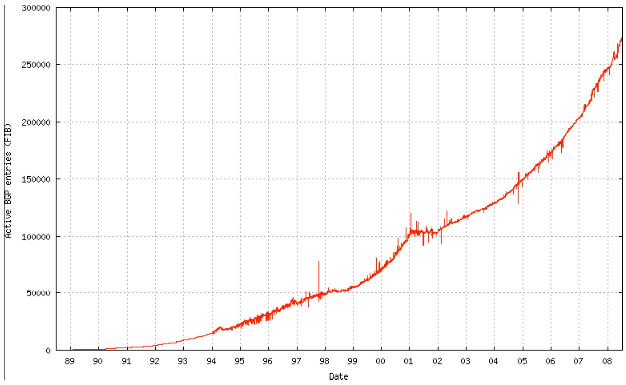

Fig. 1: Forwarding Information Base (FIB) entries per date.

and Elimination), adding the Geographic category.

### A. Design Goals

In order to overcome the challenges in scalability, mobility, multihoming and inter-domain traffic engineering the RRG defined ten design goals for that the future Internet architecture should accomplish:

1) Improved routing scalability
2) Scalable support for traffic engineering
3) Scalable support for multihoming
4) Scalable support for mobility
5) Simplified renumbering
6) Decoupling location and identification
7) First-class elements
8) Routing quality
9) Routing security
10) Deployability

### B. Separation

The first strategy proposes separating the routing space in two (e.g.: [8], [9]), the End-system Identifier (EID) space, which functions as the Globally Unique Identity (GUID), Session Identity (SID) component and local Location (LOC), and the Remote Locator (RLOC) space, which refers to a node attachment point in the Internet topology. Indeed, the local routing is done by the EID, but have each packet flow through an encoder which attaches a RLOC before the packet enters the internetwork core. The main idea behind this strategy is to not route by the EID in the core, instead route by RLOC. It is also necessary to limit the RLOC routing in the core so that only service providers (ISPs) with significant interconnection have their own RLOCs. Fewer than 10,000 such “core ISPs” exist today and the number is growing much more slowly than the routing table overall. Once the packet reaches the network identified by the RLOC, local routing by EID takes over for final delivery. In order to distribute RLOCs through the core a typical distance-vector or link-state routing protocol is needed. Additionally, as EIDs are not routable through the Internet, a mapping system is required to map an identifier onto a set of locators in order to reach this identifier.

Some variants of this approach include:

- Each core ISP has one RLOC. The RLOC’s existence and reachability is flood-propagated to the rest of the core.
- Each core ISP has a small number of RLOCs for traffic engineering (TE) proposes. The RLOCs existence and reachability is flood-propagated to the rest of the core.
- Each core ISP has an aggregated set of RLOCs which it may hierarchically assign to customers downstream and/or disaggregate for TE. The aggregated RLOCs existence and reachability is flood-propagated to the rest of the core.

Methods for mapping the EID to one or more RLOCs include:

- EIDs are statically mapped to each RLOC are periodically pushed towards a central or distributed registry. The full list is periodically downloaded to the encoders which add RLOCs to the packets.
- EIDs are dynamically mapped to each RLOC are pushed towards a central or distributed registry as they change. The registry pushes all incremental changes in near-real time to all encoders which add RLOCs to the packets.
- EIDs are dynamically mapped to each RLOC are pushed towards a central or distributed registry as they change. Encoders request and briefly cache individual mappings from the registry as needed (e.g.: [10], [11]).

Failure handling approaches include [1]:

- RLOC encoders detect when particular RLOCs are no longer reachable at all and fall back on secondary RLOCs for a particular EID. Encoders rely on active failure messages from some system in the RLOC-specified network to indicate that a host is no longer available via that RLOC, causing them to fall back on secondary RLOCs for that host (e.g.: [12], [13]).
- Link failures which prevent parts of the RLOC’s network from reaching a destination host or set of hosts it serves cause an external analysis element to make a dynamic change to the EID-to-RLOC map, depreferencing or removing the affected RLOC. The external analysis element may be under the control of the end-user destination network, the RLOC network or a third party under contract to one of them.

Compatibility approaches include:

- A new IP protocol. This would not be compatible with IPv4 and IPv6.

[1] Link failures in the Internet core cause the RLOCs to be rerouted with no change to the EID-to-RLOC map.

- A modified IP protocol. This would not be compatible with deployed IPv4 and IPv6.
- Standard IPv4 and IPv6 packets are encapsulated in a tunnel packet while they transit the Internet core (e.g.: simple IP over IP tunneling, or IP over UDP tunneling). Path-MTU issues are addressed by setting an Internet-wide maximum packet size enforced by the encoders and assuring that all core links support that size (e.g.: [8]).
- Standard IPv4 and IPv6 packets are encapsulated in a tunnel packet while they transit the Internet core. Path-MTU issues are addressed by returning packets which breach the MTU while in the core back to the encoder who must act as a proxy by returning a sensible packet-too-big message to the originating host.
- The IPv6 address space is partitioned into end-user address space (i.e.: EID space) and Internet core address space (i.e.: RLOC space). The EID-to-RLOC map is symmetric. Part of the IPv6 end-user address is swapped for the RLOC when the packet enters the Internet core and then restored when it leaves the Internet core.
- The IPv6 flow label or some other component(s) of the IPv6 header are used to contain the RLOC. The flow label is set before the packet enters the core. Non-local packets are routed based on the flow label.
- Steal bits from other functions in the IPv4 header (e.g.: checksum) to make space for an RLOC. Discard those components and set the RLOC when the packet enters the core. Restore the original bits when the packet leaves the core.

Possibles core routing methods:

- Distribute RLOCs through the Internet core via BGP (e.g.: [8]).
- Distribute RLOCs through the Internet core via a new distance-vector protocol.
- Distribute RLOCs through the Internet core via a link-state protocol.

Some disadvantages are:

- Handling path-MTU is a usually problem since the packets in the core are different than the origin host would recognize.
- Extra bandwidth is consumed by the Ingress Tunnel Router (ITR) figuring out whether the Egress Tunnel Router (ETR) is still available and functioning.
- Border filtering of source addresses (i.e: EID) becomes problematic.
- Deployment may require heavy weight "for the public good" relays in the non-upgraded part of the Internet to facilitate migration.

### C. Elimination

This proposal assigns hierarchically aggregatable RLOCs to every host (e.g.: [14], [15], [16]). It also assigns multiple RLOCs to each host such that in the network topography hosts appear as stubs in multiple locations instead of forming distant connections in the graph. Then, assigns one aggregated set of RLOCs to each core ISP, where a core ISP is one which has at least half a dozen major transit or peering links. Afterwards, it flood-propagates the aggregated RLOC's existence and reachability to the rest of the core.

Having reduced the network topology to something relatively close to a hierarchy, it performs plain old hierarchical aggregation on the RLOCs. In order to reflect changes in the nearby network hierarchies, it adds and removes RLOCs to each host dynamically during operation as needed.

Before the packet leaves the host, this approach attaches source and destination RLOCs. It routes the packets by first source then destination RLOC: move up the source network hierarchy until you can move laterally toward the destination RLOC in a permissioned manner. EID to RLOC maps are pushed from the host towards a distributed registry as they change (e.g.: DNS). Hosts request and temporarily cache individual mappings from the registry as needed.

Different RLOC variants include:

- A hierarchically aggregated numeric RLOC is dynamically assigned to each host from each upstream path. Each router receives a supernet from upstream and assigns a subnet downstream. Link state changes in the coreward path are satisfied by renumbering instead of rerouting: the host abandons the RLOC hierarchically associated with the old path. If a new path is available, the host acquires a RLOC hierarchically associated with the new path.
- A RLOC is an administratively-assigned loose source route instead of a single address. The first address in the loose source route is a universally-known waypoint router. The last address is the final destination. Link state changes in the coreward path are satisfied by reroutng in the appropriate routing domain when possible. If rerouting in the affected domain is not possible, the host abandons the impacted RLOC.
- Semi-hierarchical numeric RLOCs are administratively assigned. Local reconnection during link state changes is accomplished with rerouting instead of renumbering.

EID variants include:

- Each host has a single numeric EID to which the RLOCs are attached. This EID is used by the TCP layer and higher protocols to compose the SID.
- Each service provided by a host has a globally unique, hierarchic character-string EID to which the RLOCs are attached. Clients initiating communication with that service negotiate a numeric SID which is unique only within the scope of that service.

The major drawback of this approach is that it is not compatible with UDP or TCP, for example in [14] a shim layer between the TCP and IP layer is needed. This means that the protocol stack in every End-system must be upgraded.

### *D. Geographic*

Suppress distant routes by aggregating them into sets expected to be available in a given direction. Because RLOC reachability info is not flooded, the routing tables each router must deal with are relatively small. In the geographic aggregation all nodes within some geographic boundary are assigned the same RLOC. Routers move packets to any adjacent router deemed to be ”closer” to the RLOC in question.

The major criticism is that no one has been able to construct a protocol under this strategy without introducing constraints that are fundamentally incompatible with the Internet's economic model.

### *E. Solutions Proposals Summary*

A summary of the different proposals for the Internet routing scalability problem is depicted in Figure 2.

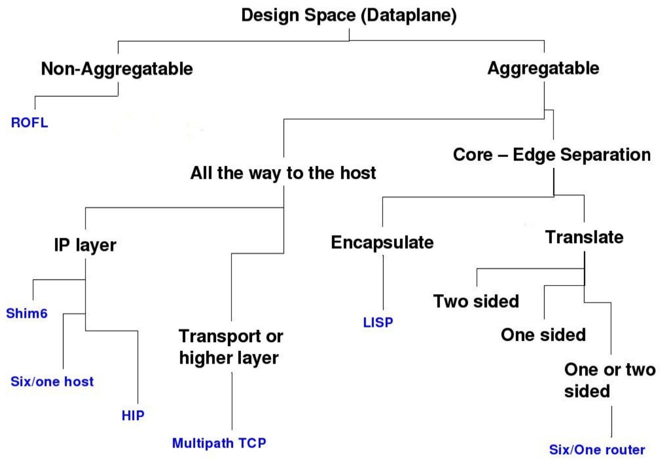

Fig. 2: Summary of the data pane proposals.

## III. Concluding Remarks

In this article we have presented the different proposed strategies to overcome the Internet routing scalability problem. We first introduce the two problematics identified as the main cause to the problem. Afterwards, the different approaches were described grouped in three categories: Separation, Elimination and Geographic. Each category defines a particular architectural approach.

## References

[1] D. Meyer, L. Zhang, and K. Fall, “Report from the IAB Workshop on Routing and Addressing,” IETF, RFC 4984, Sep. 2007. [Online]. Available: http://www.ietf.org/rfc/rfc4984.txt

[2] G. Huston, “The growth of the BGP table - 1994 to present.” http://bgp.potaroo.net.

[3] J. Saltzer, “On the naming and binding of network destinations.” RFC 1498, 1993.

[4] R. V. Oliveira, R. Izhak-Ratzin, B. Zhang, and L. Zhang, “Measurement of highly active prefixes in BGP,” in *Global Telecommunications Conference, 2005. GLOBECOM '05. IEEE*, vol. 2, Nov./Dec. 2005.

[5] J. N. Chiappa, “Endpoints and Endpoint Names: A Proposed Enhancement to the Internet Architecture,” 1999, endpoints.txt. [Online]. Available: http://mercury.lcs.mit.edu/~jnc/tech/endpoints.txt

[6] T. Li, “Design Goals for Scalable Internet Routing,” Jul. 2007, draft-irtf-rrg-design-goals-01.txt. [Online]. Available: http://tools.ietf.org/id/draft-irtf-rrg-design-goals-01.txt

[7] D. Jen, M. Meisel, H. Yan, D. Massey, L. Wang, B. Zhang, and L. Zhang, “Towards A New Internet Routing Architecture: Arguments for Separating Edges from Transit Core,” in *Hot Topics in Networks, 2008. HotNets 2008. 7th ACM Workshop on*, Calgary, Alberta, Canada, Oct. 6–7, 2008.

[8] D. Farinacci, V. Fuller, D. Meyer, and D. Lewis, “Locator/ID Separation Protocol (LISP),” Sep. 2009, draft-ietf-lisp-05.txt. [Online]. Available: http://tools.ietf.org/id/draft-ietf-lisp-05.txt

[9] C. Vogt, “Six/One Router: A Scalable and Backwards Compatible Solution for Provider-Independent Addressing,” in *MobiArch '08: Proceedings of the 3rd International Workshop on Mobility in the Evolving Internet Architecture*, Seattle, WA, USA, Aug. 22, 2008, pp. 13–18.

[10] V. Fuller, D. Farinacci, D. Meyer, and D. Lewis, “LISP Alternative Topology (LISP+ALT),” May 2009, draft-ietf-lisp-alt-01.txt. [Online]. Available: http://tools.ietf.org/id/draft-ietf-lisp-alt-01.txt

[11] V. Fuller and D. Farinacci, “LISP Map Server,” Oct. 2009, draft-ietf-lisp-ms-04.txt. [Online]. Available: http://tools.ietf.org/id/draft-ietf-lisp-ms-04.txt

[12] A. Castro, M. German, X. Masip-Bruin, M. Yannuzzi, R. Gagliano, and E. Grampin, “Advantages of a PCE-based control plane for LISP,” in *CONEXT '08: Proceedings of the 2008 ACM CoNEXT Conference*, Madrid, Spain, Dec. 9–12, 2008, pp. 1–2.

[13] M. Yannuzzi, X. Masip-Bruin, E. Grampin, R. Gagliano, A. Castro, and M. German, “Managing interdomain traffic in Latin America: a new perspective based on LISP [Topics in Network and Service Management],” *IEEE Commun. Mag.*, vol. 47, no. 7, pp. 40–48, Jul. 2009.

[14] E. Nordmark and M. Bagnulo, “Shim6: Level 3 Multihoming Shim Protocol for IPv6,” IETF, RFC 5533, Jun. 2009. [Online]. Available: http://tools.ietf.org/rfc/rfc5533.txt

[15] M. Handley, D. Wischik, and M. Bagnulo, “Multipath Transport, Resource Pooling, and implications for Routing,” Aug. 1, 2008, RRG. [Online]. Available: http://www.ietf.org/proceedings/08jul/slides/RRG-2.pdf

[16] X. Yang and X. sheng Ji, “Host Identity Protocolrealizing the Separation of the Location and Host Identity,” in *Information and Automation, 2008. ICIA 2008. International Conference on*, Changsha, Jun. 20–23, 2008, pp. 749–752.