Multi-homing in IP networks based on geographical clusters

GÁBOR NÉMETH, GÁBOR MAKRAI, JÁNOS TAPOLCAI

Budapest University of Technology and Economics,
Department of Telecommunications and Media Informatics, High Speed Networks Laboratory
{nemethgab,tapolcai}@tmit.bme.hu, makraigabor@gmail.com

Keywords: aggregation, multi-homing, BGP, Rekhter's Law, identifier-locator split

In the past years it has became a widely accepted opinion that the current network layer protocol, IPv4, suffers from several serious problems. Besides the scalability issues of the BGP routing tables [1], several features (like mobility, security) need extensions (like MIP or IPSec, respectively). These features are available as extensions of the IPv4 protocol family. The 6th version of the Internet Protocol (IPv6) is assumed to become the network layer of the future Internet, which unfortunately inherits some weaknesses of IPv4. Such as, there are no easy solutions of the scalability problems caused by edge site multi-homing. In this paper we overview the available future internet trends, and propose an easy to deploy multi-homing strategy that decreases the number of entries in the routing tables by aggregating together the address space of several edge networks located close to each other. We sketch two possible solutions for such a scenario and investigate their performance.

1. Introduction

In the last decade, the Internet Protocol has become the leading technology for inter-machine communication. Meanwhile, a great number of different enhancements were introduced, deployed (e.g. Differentiated Services, Integrated Services), and the technological boundary of the protocol's capabilities have been reached. It turned out that the fundamental architecture assumptions that underlie the routing and addressing design of the Internet Protocol are no longer sustainable. This fact generated an increasing interest in changing or re-designing the routing and addressing architecture of the Internet [2,3], especially considering that a new version of IP, IPv6, is on its way to be globally deployed.

A fundamental design principle that has made it possible for IP to scale to the magnitude of today's Internet is that address prefixes can be aggregated at higher levels of the routing hierarchy. The classic IP aggregation scheme is based on provider-subscriber relationships, where the address space of the subscriber network is part of the larger address space of the provider network. In this case the provider does not have to announce every subscriber's address prefix separately, but it announces only its own address prefix, which trivially ensures global reachability of its subscribers.

Unfortunately, the assumption that prefixes can freely be aggregated at the provider networks is no longer valid. The two most important causes strengthening this effect are: *(i)* fast endpoint mobility and *(ii)* site multi-homing with provider independent (PI) addresses.

The growing difficulties in address aggregation results in the unprecedented increase of routing table sizes what we experience today *(Figure 1)*. The sheer volume of IP addresses to manage, and the frequent processing of update messages thereof, will inevitably lead to grave scalability problems in the long run.

A straightforward solution would be to completely remove the dependency of IP addresses from their location in the network. In such an unstructured (usually termed as "flat") routing architecture, neither endpoint mobility nor multi-homing would pose problems. Let us recall Rekhter's Law, which states *"Addressing can follow topology, or topology can follow addressing. Choose one"* [5]. Thus, according the Rekhter's Law there must be a congruency between the network topology and the addressing. Understanding the current future internet proposals we may derive a slightly different law, namely: *By removing the structure from the address space, we might need to reimplement a similar structure in the control plane.*

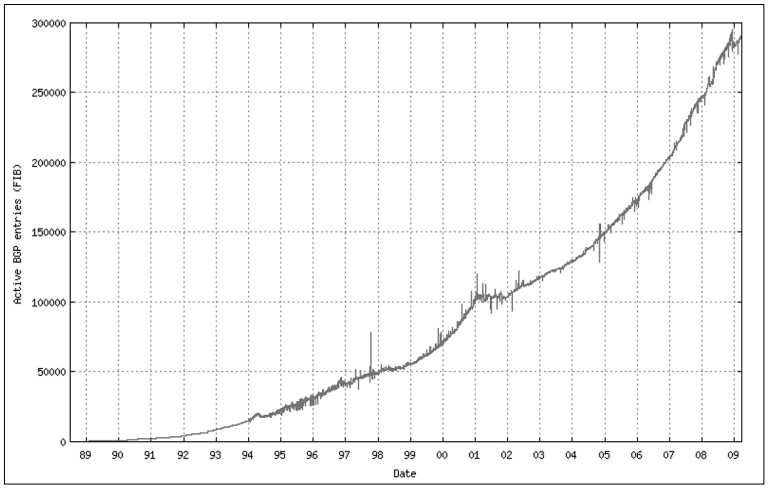


Figure 1. The BGP routing table growth [3]

In the next section we briefly describe the possible classification of the future internet proposals in order to emphasize the ideas behind the previously given law-interpretation. In Section 3 a scalable cluster based solution for multi-homed network aggregation is proposed. Later, in the same section, some simulation results are presented to characterize the proposed methods.

2. From Congruency to Structure

In this section we overview the popular trends of future internet architectures to solve the problems discussed in the previous section. An unique, usually unstructured, flat identifier (or name) is assigned for each host and/or network element, and somewhere in the network this identifier is resolved to a routeable locator (i.e., an address that can be used in a scalable routing system). The place, where this resolution is carried out can be either the host itself (like in HIP [7], shim6 [8], GSE/8+8 [9]) or somewhere deeper down in the network (like in LISP [10], eFIT [11]) (see also *Table 1*).

In the first case, users must invest in new devices, or at least update their software with a new protocol stack supporting the host based identifier-locator splitting mechanism. The majority of the home users are satisfied with the system used nowadays and may not be inclined to invest in a new one, because from their perspective the new equipment has the same capabilities. Thus, the deployment path for host-based identifier-locator splitting is somewhat obscure.

In the second case, the protocol stack of the hosts remains the same and it is the service provider that must install extra functions and (maybe) entities in its network. This approach looks more viable, however, it is still questionable whether the process of identifier-locator resolution can be made effective.

In both cases, the scalability problems are delegated to the control plane, that is, to the system coping with the identifier-to-locator and/or locator-to-identifier translation process. In the simplest case, this would amount to burdening the Domain Name System (or an appropriate successor thereof) with identifier-to-locator translation. It is yet to be seen whether the venerable DNS architecture can cope with that amount of load.

Methods differ, in addition, as to how the identifier-locator split is represented in the packets. It is possible to distinguish two different categories: map-and-encap and address rewriting (Table 1) [12].

The delivery of a packet in the map-and-encap schemes occurs as follows. A host, initially, puts the destination host's identifier into the packet. While the packet is in the transit through the core, it gets encapsulated and delivered using the locator into the destination's domain, where it is de-capsulated and delivered to its final destination using the identifier. Note that the map-and-encap schemes always append a new header to a packet instead of rewriting it, unlike the case with the rewriting methods discussed below.

	map-and-encap	*rewriting*
host-based	HIP	Shim6 Six/One
network-based	LSIP eFIT	GSE/8+8

Table 1.
The classification of the identifier-locator split proposals

The idea behind address rewriting is to divide the IP address space into two parts, and use the upper bits as the locator field and the lower bits as a unique endpoint identifier. The two parts of the address space are completely disjoint, which means that neither of the endpoints is aware of its locator. The locator part of the source address is filled in by the local egress router, while the locator part of the destination address is removed at the remote ingress router. From both sides of the communication the address of the remote host looks complete, as the locator and the identifier part is filled as it was returned by the DNS.

As a summary we may conclude, that all the introduced methods tackle the previously mentioned problems by introducing special structure in the addressing or by forcing special virtual structure of the network entities.

2.1 Push or Pull?

There are several ways to solve the problem of identifier-locator mapping: *(i)* empower the hosts to store their own bunch of locators and introduce a context establishment in the communication process (like Shim6 or Six/One) or *(ii)* delegate the mapping process to the (successor of the) DNS.

As per the enhancement of the DNS, there are two basic approaches. According to the push-based approach, the databases are pushed near the edge, i.e., the mappings are proactively fetched in order to minimize the lookup latency. In the pull-based approach, however, requests are delivered right to the authoritative server, i.e., the identifier-location mapping is performed upon the request, reactively.

Both of the push and the pull model have its own benefits:

– Pull systems have comparably low storage requirements, enabling finer and dynamically generated mappings (e.g., for mobility).
– Push systems have comparably lower latency, and hence less packet loss in the routers that buffer packets during the resolution.

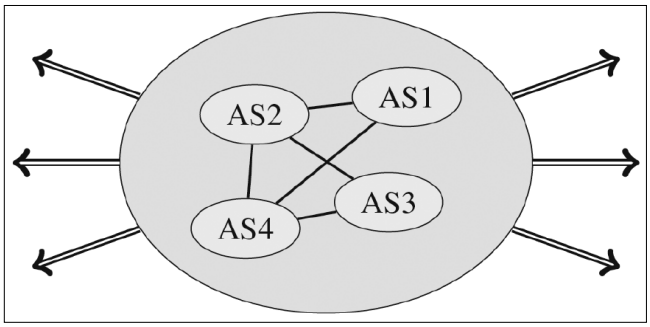
In the next section we propose a new routing architecture to deal with scalability isssues in multi-homing.

3. The proposed method: cluster based multi-homing

The scalability problems are mainly caused by small multi-homed edge networks, because the processing and storing of their small prefixes consume huge amount of resources. Thus, a new and scalable aggregation method is needed. At the same time the providers prefer not to invest much into hardware and software. Therefore, each viable method needs to be cheap and easy to deploy. Moreover, the method has to be *(i)* capable of aggregating the addresses allocated to different multi-homed edge networks and *(ii)* resistant against failure scenarios, as multi-homing was originally implemented.

These goals can be achieved by introducing special multi-homing (MH) addresses, which are completely independent from the address space of the providers' network and each multi-homed edge network gets its addresses from this special address space. Each multi-homed network has a few providers, called direct providers. These providers are able to forward packets directly to their multi-homed edge networks. However, they do not announce these addresses. These MH addresses are announced only after it is aggregated by a set of AS, called aggregation cluster (see also *Figure 2*). The edge networks aggregated together are strongly suggested to be "geographically close" to each other and to the cluster, to avoid large inter-continental hops in paths. Thus, the MH addresses have to be distributed geographically (probably by an international registry). Recall that this scheme faces only the problems caused by the small edge networks that are placed within geographically small regions.

Figure 2.
The autonomous systems called AS1, AS2, AS3 and AS4 are responsible together for a given MH address prefix. They announce the MH address together, and the incoming packets are distributed among them using IP tunneling.



Simply put, in such an architecture every packet heading to a multi-homed edge network *(i)* is first forwarded to the aggregating cluster, *(ii)* then, after detecting (one of) the real providers of the edge network, the packets are encapsulated and using IP tunnels forwarded to their real destination. After arriving to the network of the direct provider, (iii) the outer IP header is removed and the packets are forwarded using the original destination addresses.

3.1 Heuristic Models

Finding optimal aggregation clusters are almost impossible in a highly distributed and always varying system like the Internet. Instead, our focus is on simple and straightforward implementation guidelines. One can observe that the direct providers with edge networks of the same multi-homing prefix can be used as an aggregation cluster for that prefix. Also note that such a solution can minimize the AS-level path by putting the cluster as close as possible to the multi-homed edge networks. We wish to point out, that this scenario is also resistant against the same failure scenarios as the original (not scalable) architecture was. Note that, from the perspective of the edge networks, the previous cluster definition implies that they can connect only to provider networks that are part the aggregation cluster for their MH addresses. In the model presented so far the routing is performed in traditional IP forwarding until the packets reach the aggregation cluster. The routing and forwarding inside the cluster can be implemented in two different ways.

In the *Omni model* (Figure 2) each AS in a given cluster has enough information to send packet directly to its destination. That means, each AS knows every direct provider of each multi-homing edge network aggregated by the cluster. As a result, for each incoming packet only one decision is made, i.e., the selection of the direct provider's network. Thus, when a packet enters a cluster, the entry autonomous system simply matches a tunnel to it; a tunnel that ends at one of the direct providers of the destination edge network. Later in the direct provider's networks, the IP header which was used for tunneling is removed, and the packet is forwarded using the original IP address.

Special case of the Omni model can be reinterpreted as BGP Confederations [13]. In case of connected cluster, i.e., when the autonomous systems inside the same cluster are reachable through internal links, the packets can be forwarded inside the cluster without using any tunnel. This situation generally cannot be guaranteed; however from business perspective, providers may work together and deploy links among themselves. The major gain of this solution is in avoiding tunnels and so the encapsulation overhead; and the packets inside the cluster can use their own destination address, as these addresses are announced inside the cluster.

While in the Omni model each AS has sufficient information about the multi-homed edge networks in its cluster, in *MH^{2}T model* (Multi-Homing with Hash Tables) this information is reduced. The reduction is done by spreading this information among the provider networks inside the same cluster. A straightforward solution for this distribution is implementing distributed hash table (DHT).

3.2 Simulation Results

The goal for both multi-homing models was to reduce the size of BGP routing tables by improving the address aggregation procedure. First, let us theoretically estimate the reduction in the size of the routing.

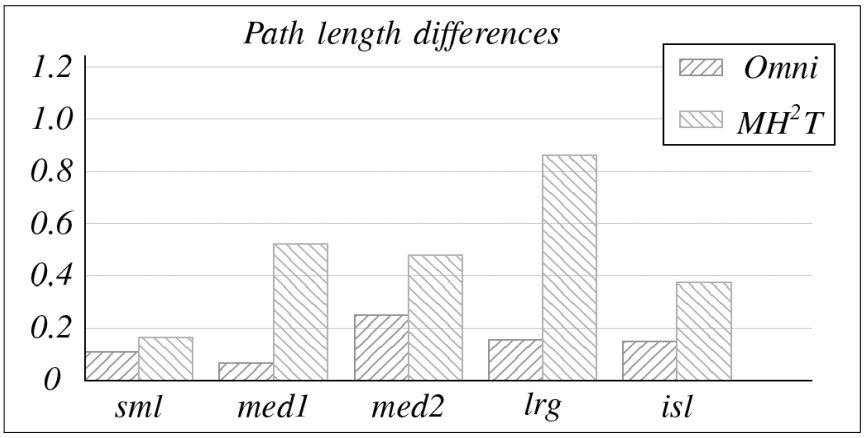


Figure 3.
Increase in the path lengths compared to the traditional BGP path lengths normalized with the size of clusters.

In the Omni model, unfortunately, the effect of the proposed aggregation model cannot be perceived in the local routing table sizes. As a result, an autonomous system participating in an aggregation cluster does not have direct benefits of the aggregation of its own edge networks, but all other autonomous systems outside the cluster do so. Thus, on the whole the size of the routing tables decreases; according to the actual configuration routing tables with 70-80-90% less entries can be reported. On the other hand, in the MH^{2}T model creating each cluster causes observable decrease in routing tables. Unfortunately, there is some data needed to maintain the DHT, which obviously increases the number of stored entries.

We have evaluated the performance of the proposed models – here we only want to show the results according to the average path length – using networks with different cluster sizes. We denoted the networks as *sml, med1, med2, lrg* and *isl*, where the cluster sizes were 3, 5, 5, 10 and 5, respectively (they were generated using reference networks in [13]). Simulation results were derived using bimodal traffic pattern.

The simulations showed that the average path length increases for both of the introduced cluster-based models. However, it never oversteps the diameter of the cluster *(Figure 3)*. Naturally, the paths observable when using the MH^{2}T model are longer than in case of the Omni model. This observation is also conforming with our expectations: the packets have to travel along the DHT to collect all the information necessary to forward them to their destination. It is important to reflect on large path length difference, mainly in case of the *lrg* network. In this case the cluster was rarely connected, i.e., packets had to enter and leave the same cluster several times. These extra hops outside the cluster have huge impact for path lengths.

We also tracked the effect of the cluster connectivity *(Figure 4)*. According to the figure, increasing the cluster connectivity the AS-level path stretch decreases rapidly. This is because inside the cluster the packets have to take shorter paths due to the dense inter-provider connectivity.

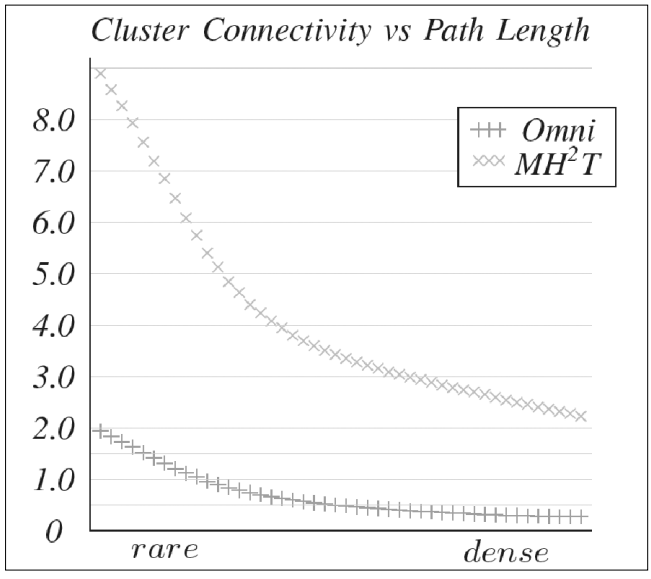
4. Conclusion

In this paper first the consequences of Rekhter's Law is discussed, as the key design principle of the future internet architectures.

In our understanding the law can be interpreted as follows: removing the structure from the address space should be performed with special care, otherwise a similar structure in the control plane must be reimplemented. It is followed by a novel proposal on scalable multi-homing solution, which is easy enough to configure to attract network administrators. The method is based on clusters that aggregate the multi-homing addresses of multiple edge network. Two different models are introduced describing the inner behavior of the cluster; in the Omni model where each AS knows every direct providers of each multi-homing edge network aggregated by the cluster, while in the MH^{2}T model this information is stored in a DHT.

While both models reduce the size of routing tables, they increase the path lengths. In the simulation the average path lengths for connected clusters vary about the original path lengths (there is an approximately 20% difference of the size of the cluster). Note that when the topology graph of cluster is not strongly connected, the packets reaching the cluster may leave the cluster and return at another AS, which cause longer paths and extra costs for the providers. In our simulations in this rarely connected case the path length increased approximately 80% relative to the size of the cluster. Unfortunately, not only the length of the paths grew, but also

Figure 4.
Increase in the path lengths compared to the traditional BGP path lengths, when increasing the cluster connectivity by adding links. The path lengths are hops between AS's, and the lengths are normalized by the size of the cluster (for lrg it is 10).



the load of the AS's. The load increase may be a drawback for the providers, thus they may avoid participating in any cluster, or even if they participate they advertise longer paths in BGP advertisements.

Our future work is to develop methods for load balancing to deny uncooperative announcements.

Authors

GÁBOR NÉMETH has received his M.Sc. degree in technical informatics from Budapest University of Technology and Economics (BME), Budapest, Hungary in 2007, where he is currently working towards his Ph.D. degree at the Department of Telecommunications and Media Informatics (TMIT). His research interests focus on multi-homing, routing and future Internet trends.

GÁBOR MAKRAI has received his B.Sc. degree in 2009 from Budapest University of Technology and Economics (BME), Budapest, Hungary, where he will continue his study and research towards his M.Sc. degree.

JÁNOS TAPOLCAI received his M.Sc. degree in Technical Informatics in 2000 and Ph.D. in Computer Science in 2005 from Budapest University of Technology and Economics (BME), Budapest, Hungary. Currently he is an associate professor at the High-Speed Networks Laboratory at the Department of Telecommunications and Media Informatics at BME. His research interests include applied mathematics, combinatorial optimization, linear programming, communication networks, routing and addressing, availability analysis and distributed computing. He has been involved in several related European and Canadian projects. He is an author of over 40 scientific publications, and is the recipient of the Best Paper Award in ICC'06.

References

[1] T. Bu, L. Gao, D. Towsley,
"On characterizing BGP routing table growth,"
In Proceedings of IEEE GLOBECOM'02,
Taipei, Taiwan, November 2002.

[2] FIND, NSF NeTS FIND Initiative Website,
http://www.nets-find.net/

[3] Internet Research Task Force Routing
Research Group,
http://www.irtf.org/

[4] BGP Routing Table Analysis Reports,
http://bgp.potaroo.net/

[5] D. Meyer, Ed., L. Zhang, Ed., K. Fall, Ed.,
"Report from the IAB Workshop on
Routing and Addressing,"
RFC 4984, September 2007.

[6] P. Savola, T. Chown,
"A Survey of IPv6 Site Multihoming Proposals,"
In Proceedings of the 8th International Conference of
Telecommunications (ConTEL) 2005,
Zagreb, Croatia, June 2005, pp.41–48.

[7] R. Moskowitz, P. Nikander,
"Host Identity Protocol (HIP) Architecture,"
RFC 4423, May 2006.

[8] P. Savola,
"IPv6 site multihoming using a host-based shim layer,"
In 5th International Conference on Networking and
the International Conf. on Systems (ICN/ICONS/MCL'06),
Mauritius, April 2006.

[9] Mike O'Dell,
"GSE – An Alternate Addressing Architecture for IPv6,"
IETF Draft, 1997.

[10] D. Farinacci, V. Fuller, D. Oran,
"Locator/ID Separation Protocol (LISP),"
IETF Draft, 2008.

[11] D. Massey, L. Wang, B. Zhang L. Zhang,
"A Proposal for Scalable Internet Routing & Addressing,"
IETF Draft, 2007.

[12] D. Meyer,
"Update on Routing and Addressing at IETF 69",
IETF Journal, Volume 3, Issue 2, October 2007.
http://www.isoc.org/tools/blogs/ietfjournal/

[13] D. Medhi, K. Ramasary,
"Network Routing Algorithms,
Protocols and Architecture",
Elsevier Inc., 2007.